Raisonnements non certains et changement de croyances

1. Introduction

Une des conclusions les plus largement admises de l'épistémologie contemporaine est négative : il n'est pas possible de trouver des fondements certains aux énoncés empiriques et aux raisonnements qui les affirment en évitant une régression à l'infini (trilemme de Fries). Le faillibilisme de Popper, bien qu'en accord avec cette thèse, considérait qu'une voie restait ouverte aux raisonnements purement déductifs - et donc certains - , celle de la réfutation. La thèse de Duhem-Quine va pourtant pourfendre ce dernier refuge de la certitude : d'un point de vue logique, tous nos énoncés, empiriques et même logiques, affrontent collectivement le tribunal de l'expérience. Pour rendre compte des raisonnements scientifiques, Neurath [NEU 33] considère dès lors que « *nous sommes des marins qui doivent reconstruire leur navire au grand large, sans jamais pouvoir le décomposer en cale sèche, afin de le reconstruire à partir des meilleurs composants* ».

La conception de la rationalité des croyances se trouve ainsi déplacée de la question de leur fondement à celle de leur changement. Les croyances scientifiques sont des données initiales modifiées au fur et à mesure de l'expérience pour conserver la cohérence de l'ensemble. Le principe fondamental de cette modification, baptisé « maxime de mutilation minimale » par Quine [QUI 72], consiste, face à des réfutations expérimentales qui affectent collectivement nos croyances, à préserver celles qui sont le plus profondément ancrées. Considérant que l'on peut traiter de la même façon les croyances scientifiques et ordinaires, l'Intelligence Artificielle a développé plus récemment des théories des changements de croyances fondées sur cette maxime.

Auteurs : Bernard WALLISER – Denis ZWIRN - Hervé ZWIRN

Les travaux d'Intelligence Artificielle proposent des modélisations de nombreux types de raisonnements ou d'énoncés ordinaires qui échappent tout autant à la certitude déductive que les raisonnements scientifiques :

- les raisonnements non-monotones dont les conclusions s'avèrent défaisables lorsqu'on acquiert de nouvelles informations ;
- les énoncés contrefactuels qui indiquent ce qui passerait si un antécédent non réalisé venait à s'actualiser ;
- les raisonnements abductifs qui consistent à dégager de faits symptomatiques des hypothèses explicatives ;
- les raisonnements probabilistes dont les conclusions sont affectées d'un degré de croyance de nature subjective.

Parmi les modélisations proposées, une approche apparaît de plus en plus capable d'offrir un cadre synthétique à toutes les autres: la théorie du changement des croyances. Ce corpus théorique, dû aux travaux fondateurs de Alchourron-Gärdenfors-Makinson [ALC 85], développe formellement l'idée fondamentale de Quine quant à la manière dont nos croyances évoluent. Les croyances y sont en effet dotées de « degrés d'enracinement épistémique » qui contraignent leurs méthodes de révision. L'objectif de cet article est d'indiquer, à travers plusieurs exemples, la fertilité de cette idée pour représenter les diverses formes de raisonnement non déductif.

Après avoir indiqué le cadre formel général adopté (§2), il décrit les principes des théories actuelles du changement des croyances en distinguant deux contextes de changement, la rectification (§3) et la mise à jour (§4). Il s'attache ensuite à indiquer la manière dont ces théories permettent d'une part d'appréhender la notion de raisonnement non-monotone (§5), d'énoncé contrefactuel (§6), de raisonnement abductif (§7), et d'autre part de fournir des principes pour choisir des règles de changement des croyances probabilisées (§8). La dernière section (§9) discute, en forme de conclusion, la possibilité d'utiliser ces règles de changement pour probabiliser les concepts évoqués dans les sections précédentes.

2. Le cadre formel du raisonnement

Considérons une syntaxe propositionnelle définie par un langage formel **L**, clos à partir d'un ensemble fini de propositions atomiques {a, b, …} par usage des connecteurs usuels : $\neg$ (négation), $\wedge$ (conjonction), $\vee$ (disjonction) et $\rightarrow$ (implication matérielle). Soit T et $\perp$ les deux constantes désignant respectivement le vrai et le faux. Soit $\vdash$ le symbole de l'opération de déduction au méta-niveau. Considérons par ailleurs une sémantique de nature ensembliste, définie à partir d'un ensemble fini de "mondes possibles" par les opérations - (complémentation),

$\cap$ (intersection), $\cup$ (union) et $\subseteq$ (inclusion). Soit W et $\varnothing$ respectivement l'ensemble de tous les mondes possibles et l'ensemble vide.

On peut dès lors établir une correspondance entre la syntaxe et la sémantique. Appelons A, B, C les sous-ensembles (ou "événements") caractérisés par le fait que les propositions a,b,c sont respectivement vraies dans chacun des mondes qui les composent et fausses dans les autres. Les deux cadres deviennent isomorphes, modulo des conditions standards, avec la correspondance suivante entre les connecteurs propositionnels et les opérations ensemblistes : aux connecteurs $\neg$, $\cap$, $\cup$, $\subseteq$ correspondent respectivement les opérations -, $\wedge$, $\vee$, $\vdash$. L'article utilisera le second cadre formel, plus commode, en introduisant pour chaque type de raisonnement des symboles correspondant au type d'inférence spécifique concerné.

3. La théorie de la rectification des croyances

La modélisation du changement - ou de la révision - des croyances est initialement due aux travaux de Alchourron, Gärdenfors et Makinson [ALC 85]. Une opération * de révision des croyances relie une croyance initiale K et un message A à une croyance finale K*A. Le message est une information considérée par l'agent comme prioritaire sur la croyance initiale, pour des raisons exogènes au modèle (il provient par exemple d'un expert de la question considérée). Croyance initiale et message portent sur un système réel statique, ce qui signifie que seul un des "mondes possibles" est le monde réel que le changement de croyances a pour objet de cerner. Ce contexte de changement a ultérieurement été baptisé contexte de "rectification" des croyances *(revising)*.

Ces auteurs [ALC 85] proposent un ensemble d'axiomes caractérisant la rationalité d'un agent dans ce type de contexte. On peut citer de manière non exhaustive :

- axiome de succès : $K*A \subseteq A$
- axiome de conservation : si $(K \subseteq A)$ alors $(K*A = K)$
- axiome d'inclusion : $K \cap A \subseteq K*A$
- axiome de préservation : si $(K \cap A \neq \varnothing)$ alors $(K*A \subseteq K \cap A)$

L'axiome de succès représente la priorité accordée au message ; les autres axiomes traduisent sous différentes formes un principe général d' « économie » consistant à minimiser autant que possible le changement des croyances initiales, conformément à la maxime de Quine.

Dans une sémantique de mondes possibles, l'opération de révision des croyances peut par ailleurs être construite de la manière suivante. Considérons une relation de préférence $<_K$ (et une relation d'équivalence associée $=_K$) définie sur W et indexée sur K. Cette relation est supposée être un pré-ordre total respectant les deux propriétés suivantes :

(i) $w' \in K$ et $w'' \in K \Rightarrow w' =_K w''$

(ii) $w' \in K$ et $w'' \notin K \Rightarrow w' <_K w''$

Ce pré-ordre peut être représenté par un système de « couronnes » emboîtées autour d'un disque correspondant à l'ensemble des mondes de la croyance initiale K, les mondes étant d'autant plus éloignés de ce disque qu'ils se situent sur une couronne plus extérieure. On peut alors définir les "mondes minimaux" d'un événement quelconque X relativement à une croyance K par :

$\mathrm{Min}_K (X) = \{w \in X , \ \forall w' \in X \ w' <_K w \text{ est faux}\}$

En termes d'interprétation, ces mondes sont les mondes de X "les plus proches" de la croyance initiale K ; ce sont les mondes qui devraient être privilégiés par un agent lorsqu'il apprend X pour respecter les principes d'économie évoqués. On peut en effet établir un théorème de représentation qui énonce, conformément à cette attente :

$K*X = \mathrm{Min}_K (X)$

En reprenant la représentation en terme de couronnes, la croyance finale est définie par l'intersection du message avec la couronne la plus proche de la croyance initiale avec laquelle il possède une intersection non vide. Dans le cas où la croyance initiale et le message ont une intersection non vide (les deux sont compatibles), la croyance finale est précisément cette intersection.

Si l'on en revient à un cadre syntaxique, cette procédure revient à affecter un degré d'enracinement épistémique à chacune des propositions (le degré d'enracinement de l'événement associé est défini à partir du pré-ordre sur les mondes). L'opération de rectification revient alors à ajouter aux propositions de la croyance initiale la proposition représentée par le message (de degré d'enracinement maximal), puis à retirer de l'ensemble ainsi constitué les propositions les moins enracinées jusqu'à restaurer la cohérence de l'ensemble résiduel.

4. La théorie de la mise à jour des croyances

Une modélisation du changement de croyances, complémentaire de la précédente, a été proposée par Katsuno et Mendelzon [KAT 91]. Cette modélisation n'est pas conçue comme une alternative à celle de [ALC 85], mais plutôt comme indiquant des principes de changement différents dans un contexte différent. Une opération $\Box$ de "mise à jour" *(updating)* relie toujours une croyance initiale K et un message A à une croyance finale K $\Box$ A, mais dans un contexte dans lequel le système concerné par la croyance et le message est en évolution. A chaque étape du processus de raisonnement, le monde possible correspondant au monde réel peut lui-même se modifier. Il en résulte que le principe de la priorité accordée au message sur la croyance devient encore plus fort, au détriment si nécessaire du principe d'économie, car il traduit la priorité accordée à l'information la plus récente.

Ces auteurs [KAT 92] proposent un ensemble d'axiomes caractérisant la rationalité d'un agent dans un contexte de mise à jour. Plusieurs de ces axiomes sont communs avec ceux de [ALC 85], notamment les trois premiers énoncés à la section précédente ; certains axiomes d'économie sont relâchés parce qu'ils sont considérés comme trop forts, en particulier l'axiome de préservation qui cède la place à l'axiome nouveau suivant :

axiome de distributivité à gauche : $(K \cup K') \Box A = (K \Box A) \cup (K' \Box A)$

-

Le message révise la croyance initiale en considérant indépendamment chacune des hypothèses qu'elle contient.

D'un point de vue sémantique, l'opération de mise à jour des croyances est construite en remplaçant la relation de préférence globale $<_K$ par une famille de relations de préférence locales $<_w$ indexées sur les mondes possibles. Cette relation est à nouveau supposée être un pré-ordre total respectant la propriété suivante :

(j) $w' \neq w \Rightarrow w <_w w'$

On peut alors définir les "mondes minimaux" d'un événement quelconque X relativement à un monde possible initial w par :

$\text{Min}_w(X) = \{w' \in X , \ \forall w'' \in X \ w'' <_w w' \text{ est faux}\}$

Ces pré-ordres peuvent être à nouveau représentés par des systèmes de couronnes emboîtées autour cette fois de points correspondant à chacun des mondes possibles w, les mondes étant d'autant plus éloignés de ce point qu'ils se situent sur une couronne plus extérieure. La croyance finale est alors obtenue en considérant l'union de tous les mondes possibles minimaux du message

relativement à chacun des mondes de la croyance initiale. Un théorème de représentation relie à nouveau les axiomes à cette sémantique sous la forme :

$$K\Box A = \bigcap_{w \in K} Min_w(A) = \{w' \in A, \exists\, w \in K, \forall\, w'' \in A : w' \leq_w w''\}$$

Remarque 1. Les opérations de rectification et de mise à jour n'indiquent pas de processus de computation effectif de la croyance finale, mais seulement des contraintes générales de rationalité qui régissent la révision des croyances. La croyance finale résulte donc d'une règle particulière dans une classe de règles, sauf si on incorpore dans le modèle une spécification précise des relations de préférence sous-jacentes $<_K$ ou $<_w$.

Remarque 2. Les opérations de rectification et de mise à jour sont équivalentes lorsque la croyance initiale est réduite à un seul monde possible. Dans tous les autres cas, elles sont logiquement incompatibles (une même opération de changement ne peut pas satisfaire simultanément tous leurs axiomes). Elles se réfèrent à des contextes et à des processus épistémiques différents. Cette différence va se révéler dans les autres formes de raisonnements dont l'une ou l'autre de ces opérations de changement est capable de rendre compte.

5. Le raisonnement non-monotone

Les conclusions des raisonnements empiriques sont rarement définitives. Ainsi en va-t-il par exemple des raisonnements courants, qui s'appuient sur les propriétés prototypales attribuées à un type d'objet (les oiseaux volent), l'héritage de ces propriétés à tous les objets du même type (à tous les oiseaux) pouvant souffrir d'exceptions (les autruches, les oiseaux venant de naître). Ainsi en va-t-il aussi, dans le domaine scientifique, de l'interprétation dans le langage observationnel des énoncés d'une théorie, toujours conditionnelle à une liste potentiellement infinie de "provisos" : on convient en général que l'attraction entre deux corps est régi par les lois de la gravitation, pourvu que ne s'exerce sur eux aucune autre force, dont la liste complète ne peut en fait pas être dressée [HEM 88]. L'idée commune à ces types de raisonnements est que la conclusion découle "normalement", mais non certainement, des prémisses. A l'encontre des raisonnements déductifs, une caractéristique importante de ces raisonnements est leur "non-monotonie" au sens suivant : une nouvelle prémisse peut défaire les conclusions initialement établies.

De nombreuses logiques ont été successivement proposées pour modéliser ce type de raisonnement (voir [GRE 90], [GAB 94]), par exemple la logique des défauts [REI 80] ou la logique auto-épistémique [MOO 88]. Le modèle d'inférence non-monotone proposé par Kraus, Lehmann et Magidor [KRA 90], à la suite des

travaux syntaxiques de Gabbay [GAB 85] et sémantiques de Shoham [SHO 87], offre un cadre synthétique qui résout de nombreuses difficultés des premières logiques non-monotones. Dans ce modèle, une inférence non-monotone notée A|~B est une relation d'inférence défaisable dont l'interprétation standard est précisément: "si A, normalement B".

[KRA 90] et [LEH 92] ont proposé un système d'axiomes régissant l'usage de cette relation d'inférence par un agent rationnel, contenant par exemple les axiomes suivants :

- axiome d'affaiblissement à droite : si A |~ B et B $\subseteq$ C alors A |~ C
- axiome de réflexivité : A |~ A
- axiome de conditionnalisation : si A |~ B alors T |~ A $\rightarrow$ B

Selon le système d'axiomes retenu, plusieurs logiques non-monotones (emboîtées) sont concevables. La logique non-monotone "rationnelle" se caractérise en particulier par le respect de l'axiome suivant :

- axiome de monotonie rationnelle :

si (non (A |~ -B) et A |~ C) alors (A $\cap$ B |~ C)

Cet axiome impose une version minimale de l'axiome classique de monotonie: pour que la conclusion d'une inférence soit conservée en ajoutant une prémisse, il suffit que cette nouvelle prémisse ne soit pas normalement exclue par les prémisses initiales.

D'un point de vue sémantique, considérons à nouveau un ordre partiel strict < sur W tel que les mondes minimaux ou "normaux" associés à un événement X soient définis par :

Min (X) = {w $\in$ X ,$\forall$ w'$\in$X, w' < w est faux}

Sous certaines conditions supplémentaires imposées à la relation d'ordre <, un théorème de représentation permet alors d'établir que "si A, normalement B" est vrai, si B est vrai dans tous les mondes normaux de A [LEH 92] :

A |~ B ssi Min (A) $\subseteq$ B

Cette logique peut être reliée directement à la théorie des changements de croyance dans un contexte de rectification, grâce au résultat suivant établi par [GAR 94] :

A $|\sim_K$ B ssi K*A $\subseteq$ B

Autrement dit, si la croyance initiale K sert de paramètre pour spécifier le contexte cognitif dans lequel se déroule l'inférence non-monotone, le raisonnement

“si A, normalement B” revient à réviser la croyance initiale K par le message A et à vérifier que B est déductible de la croyance finale.

A la correspondance sémantique entre les deux formes de raisonnement logique, on peut associer une correspondance directe entre certains de leurs axiomes. Par exemple, les axiomes de réflexivité et de conditionnalisation de la logique non monotone correspondent respectivement aux axiomes de succès et d’inclusion des logiques du changement de croyances. L’axiome de monotonie rationnelle correspond, quant à lui, à un axiome de “super-expansion”, qui généralise l’axiome de préservation, caractéristique de la rectification à l’encontre de la mise à jour, au cas de deux messages successifs.

Remarque : Il est également possible de relier la logique non-monotone à la logique de la confirmation des énoncés, et donc cette dernière à la théorie de la rectification des croyances ; l’inférence A |~ B, définie par un système d’axiomes conformes à des principes épistémologiques proposés notamment par Hempel [HEM 65], est alors interprétée par “A confirme absolument B” ou "en présence de A, j'accepte B" [ZWI 96].

6. Les conditionnels contrefactuels

De nombreux raisonnements scientifiques ou ordinaires reposent sur l’usage d’énoncés conditionnels du type : “si A alors B”, liant un antécédent A à un conséquent B. La première manière d’interpréter logiquement ce type d’énoncés est d’identifier leur valeur de vérité à celle de l’implication matérielle de la logique des propositions : $A \rightarrow B$, définie comme équivalente à (-A U B). Cela soulève toutefois de nombreuses difficultés et paradoxes (voir [JAC 91]), dans la mesure où on ne souhaite pas que tous les énoncés conditionnels puissent être vrais en vertu de la seule fausseté de leur antécédent ou de la seule vérité de leur conséquent. Une telle interprétation obligerait en particulier à considérer comme toujours vrais les “conditionnels contrefactuels”, c’est à dire ceux dont l’antécédent est faux, et dont le prototype est “si les kangourous n’avaient pas de queue, ils tomberaient”. Les énoncés contrefactuels sont utilisés couramment en science pour définir des propriétés physiques “dispositionnelles”, du type “soluble” ou “cassable” (voir [GOO 84]). Leur assimilation à des implications matérielles ne permet pas de rendre compte du fait que ces énoncés peuvent être vrais ou faux non en vertu de la seule logique, mais en fonction de la réalité empirique. Elle ne rend pas compte non plus de l’intuition attachée à la « relation causale » entre l’antécédent et le conséquent que ces énoncés tendent à exprimer.

Un conditionnel est en fait défini comme une proposition logique (ou un événement) et non comme un mode d’inférence. Plusieurs modélisations ont

néanmoins été proposées pour rendre compte des raisonnements sur les conditionnels contrefactuels, en particulier par Stalnaker [STA 68] et par Lewis [LEW 73]. En traduisant la logique "officielle" VC de Lewis dans un cadre ensembliste, l'événement $A > B$ (ou de manière équivalente la proposition $a > b$) s'interprète comme "si A était le cas, alors B serait le cas" ; il obéit à un système d'axiomes (et de règles), dont les noms ci-dessous sont dus à [NUT 80], en particulier :

- axiome ID (Identité) : $A > A$
- axiome MP (Modus Ponens) : $(A > B) \subseteq (A \rightarrow B)$
- axiome CS : $(A \cap B) \subseteq (A > B)$

La logique C2 de Stalnaker se caractérise par l'ajout au système VC de l'axiome caractéristique suivant :

- axiome CEM (Conditional Excluded Middle) : $A > A$ ou $A > -A$

Plusieurs sémantiques permettent de définir les conditions de vérité de l'événement $A > B$ en chaque monde w. Lewis [LEW 73] utilise en particulier un système de couronnes emboîtées autour de w obéissant à certaines conditions . On peut montrer qu'il est équivalent à la famille des relations de pré-ordre local $<_w$ utilisée pour définir l'opération de mise à jour. Moyennant cette équivalence, on peut établir directement le résultat suivant [GRA 91] :

$w \subseteq (A > B)$ ssi $w \,\square\, A \subseteq B$

Autrement dit, le conditionnel contrefactuel "si A était le cas, alors B serait le cas" est vrai dans un monde w si, en apprenant que ce monde a évolué de telle manière qu'il valide le message A, la croyance finale qui en résulte valide B.

A nouveau, la correspondance sémantique entre les deux logiques peut être associée à une correspondance directe entre certains de leurs axiomes. Par exemple, l'axiome ID de la logique VC de Lewis correspond à l'axiome de succès, l'axiome MP à l'axiome d'inclusion et l'axiome CS à l'axiome de conservation.

Ce résultat doit être comparé à une conjecture plus ancienne, due à Ramsay [RAM 50], selon laquelle la bonne méthode pour évaluer un conditionnel consiste à ajouter l'antécédent au stock de croyances initiales, puis à effectuer tous les ajustements nécessaires pour restaurer la cohérence sans contredire l'antécédent, et enfin à vérifier si le conséquent s'en déduit. Cette méthode, baptisée "test de Ramsay", peut être formalisée dans les termes de la théorie du changement de croyances de la manière suivante, en notant Δ l'opération de changement inconnue utilisée :

[test de RAMSAY] $(A > B)_K$ ssi $K \,\Delta\, A \subseteq B$

Gärdenfors [GAR88] a démontré que le test de Ramsay n'est pas satisfait si on interprète Δ par l'opération * de rectification des croyances. Il est par contre possible de démontrer qu'il est satisfait si Δ est interprété par l'opération $\Box$ de mise à jour des croyances et si l'événement $(A > B)_K$ est défini de la manière suivante :

$(A > B)_K = \{w \in K,\ w \subseteq (A > B)\}$

Ce résultat permet de généraliser l'équivalence entre la logique des conditionnels contrefactuels et la théorie de la mise à jour des croyances, en l'étendant au cas où la croyance initiale est incertaine et contient plusieurs mondes.

7. Le raisonnement abductif

Le concept d'abduction a été introduit par le philosophe Charles Peirce (1931,1958) pour décrire la forme de raisonnement incertain conduisant de manière très générale à formuler des hypothèses explicatives. Dans un contexte épistémologique, le raisonnement abductif peut être une inférence de faits vers des faits, comme en physique lorsque Leverrier formula l'hypothèse de l'existence d'une nouvelle planète, Neptune, ou une inférence de faits vers des théories ou des lois, comme en économie lorsqu'une fonction d'utilité est construite pour un consommateur à partir de l'observation de ses consommations dans un contexte économique donné. Dans un contexte de raisonnement ordinaire, le raisonnement abductif a été mis en avant, en Intelligence Artificielle, en relation avec le diagnostic réalisé par un expert. Il en est ainsi du diagnostic d'un médecin qui conjecture une maladie en présence de symptômes ou de celui d'un policier qui devine un coupable en présence d'indices.

Selon une première définition proposée par Peirce**,** l'**abduction$_1$** intervient dans le contexte syllogistique du "schéma déductif-nomologique", ultérieurement popularisé par Popper [POP 59] ou Hempel [HEM 65]. Considérons un syllogisme reliant :

- une règle ou une loi L (par exemple : tous les avions volent)
- un cas ou un fait E (par exemple : ceci est un avion)
- une conséquence ou un phénomène P (ceci vole)

Alors que la prédiction associe déductivement E et L à P, l'induction associe non déductivement E et P à L et l'abduction$_1$ associe non déductivement P à E et L. En effet, l'observation que ceci vole peut s'expliquer par le fait que ceci est un avion et que les avions volent (bien qu'il ne s'agisse pas de la seule explication possible du phénomène).

Cette théorie de l'abduction présente l'avantage de la simplicité et de la précision formelle, mais elle se heurte à plusieurs limites :

- elle admet des raisonnements abductifs « anormaux », effectués en utilisant des lois très peu plausibles (dans le cas précédent, l'inférence que l'objet est un vaisseau intergalactique car tous les vaisseaux intergalactiques volent) ;
- elle exclut a contrario des raisonnements abductifs courants, qui s'appuient sur des lois seulement probables ou prototypales (dans le cas précédent, l'inférence que ~~ceci~~ l'objet est un oiseau car tous les oiseaux volent).

Selon une seconde définition proposée par Peirce, l'**abduction$_2$** est un mode de raisonnement plus général défini dans un contexte scientifique dynamique. Lorsqu'un savant découvre un fait nouveau, éventuellement surprenant, son "calme cognitif" est troublé. La méthode scientifique se déploie alors en trois étapes:

- l'abduction$_2$ correspond à une première étape au cours de laquelle le scientifique imagine des hypothèses explicatives de ces faits ;
- la déduction correspond à une seconde étape dans laquelle il en déduit des conséquences testables ;
- l'induction correspond à une troisième étape qui consiste à confirmer les hypothèses en testant expérimentalement leurs conséquences.

L'abduction$_2$ semble appartenir au contexte (non purement logique) de la découverte, au contraire de la deuxième et de la troisième étapes qui relèvent plus du contexte de la preuve. Toutefois, dès cette première étape, toutes les hypothèses explicatives ne sont pas de bonnes candidates pour la suite : même si l'abduction ne conduit à aucune croyance ferme dans une seule hypothèse, un scientifique ne considèrera sérieusement que certaines hypothèses abductives possibles. Il faut donc formuler un critère logique de sélection de ces « bonnes explications », qui seules seront considérées comme des candidates à une croyance valide.

La théorie logique la plus simple d'abduction comme inférence explicative est celle de la déduction inverse, selon laquelle l'abduction est l'inverse de la déduction, elle-même conçue comme la forme la plus directe d'explication. L' **abduction classique** d'un fait A vers une hypothèse B est ainsi définie par :

$A \Vdash B$ ssi $B \subseteq A$

Cette définition n'est toutefois pas satisfaisante car elle constitue une version dégénérée du syllogisme et se heurte ainsi aux deux limites déjà rencontrées par l'abduction$_1$. Une théorie satisfaisante du raisonnement abductif se doit d'éviter ces deux écueils, en considérant que les bonnes explications d'un fait doivent respecter les deux conditions suivantes :

a. elles doivent être des explications acceptables dans un certain contexte de croyance.

b. elles doivent pouvoir être des explications non déductives, par exemple non-monotones

On peut alors proposer les trois définitions suivantes de l'inférence abductive en mobilisant d'emblée le formalisme de la théorie du changement de croyances dans un contexte de rectification (le contexte de mis à jour n'est a priori pas pertinent dans la mesure où il s'agit ici de formuler des hypothèses à propos d'un monde statique) :

1. **l'abduction non transitive** : $A \ \|\sim\ B$ ssi $\varnothing \neq K^*B \subseteq A$
2. **l'abduction non réflexive** : $A \ \|\pi\ B$ ssi $\varnothing \neq B \subseteq K^* A$
3. **l'abduction ordonnée** : $A \ \|\approx\ B$ ssi $\varnothing \neq K^*B \subseteq K^*A$

L'abduction non transitive, proposée par Boutilier et Becher [BOU 95] sous le nom d' « explication prédictive », consiste à inverser l'opération de rectification des croyances, une hypothèse étant abduite d'un message si la rectification de la croyance initiale par cette hypothèse valide le message ; elle affaiblit donc l'abduction classique en considérant que l'explication du message envisagée peut être non-monotone (compte tenu de la relation entre rectification des croyances et inférence non-monotone), ce qui satisfait la seconde condition. L'abduction non réflexive, qui traduit dans le langage de la théorie de la rectification un critère proposé par Mayer et Pirri [MAY 96], consiste à n'admettre comme hypothèses abduites que celles qui valident les cas "normaux" du message, c'est-à-dire les cas les plus plausibles compte tenu de la croyance initiale, ce qui valide la première condition. L'abduction ordonnée, proposée par les auteurs de ce papier [WAL 01b], représente une synthèse des deux critères précédents, apte à satisfaire simultanément les deux conditions. Elle admet comme hypothèses abduites toutes les explications non-monotones qui valident les cas normaux du message à expliquer; plus précisément, une hypothèse abduite est telle que si on rectifie la croyance initiale par cette hypothèse, on valide la rectification de la croyance initiale par le message.

Les auteurs [WAL 01b] ont proposé un système axiomatique pour chacune de ces trois formes d'abduction, et ont établi un théorème de représentation pour l'abduction non réflexive et pour l'abduction ordonnée. Les noms donnés à ces trois formes d'abduction correspondent à des axiomes qu'elles sont les seules (parmi les trois) à respecter. L'abduction ordonnée est en fait caractérisée par un pré-ordre sur les événements : elle est à la fois réflexive et transitive ; elle satisfait d'autres axiomes qui semblent intuitivement adéquats pour caractériser le raisonnement abductif, par exemple :

- axiome de disjonction à droite : si $[(A \|\approx B) \wedge (A \|\approx C)]$ alors $(A \|\approx B \cup C)$

- axiome de monotonie faible : si $[(A \,\|\!\approx B) \wedge (B \subseteq C)]$ alors $(A \cap C \,\|\!\approx B)$
- axiome de conjonction à gauche : si $[(A \,\|\!\approx B) \wedge (C \,\|\!\approx B)]$ alors $[(A \cap C) \,\|\!\approx B]$

Aucun des axiomes précédents ne correspond directement (compte tenu de la définition de l'abduction ordonnée) à un axiome unique de la théorie de la rectification des croyances, mais ils se déduisent tous d'un ensemble d'axiomes de cette théorie. Un axiome simple tel que l'axiome de conservation peut cependant être associé directement à l'axiome suivant, valide dans la logique de l'abduction ordonnée :

- axiome de conservation abductive : si $T \,\|\!\approx B$ et $(B \,\|\!\approx A)$ alors $T \,\|\!\approx A)$

Cette approche de l'abduction par la théorie de la rectification des croyances pourrait encore être améliorée selon trois voies : en cherchant à rendre compte par la notion de « pouvoir explicatif » de la sélection des meilleures explications d'un fait, en l'étendant au calcul des prédicats pour retrouver l'intuition attachée à l'abduction$_1$, et bien sûr en étudiant ses liens avec le raisonnement probabiliste. D'autres travaux récents en Intelligence Artificielle tentent par ailleurs de préciser les frontières et relations entre les concepts d'abduction et d'induction [FLA 00].

8. La théorie du changement de croyances probabiliste

Une autre manière de traiter les raisonnements incertains est d'affecter des degrés de croyance à leurs prémisses et à leurs conclusions. Ces degrés peuvent être représentés par des probabilités, bien que d'autres types de mesures soient également envisageables (par exemple les fonctions de croyance [SHA 76] ou les familles de probabilités [CHA 89]). Les probabilités utilisées dans ce cas sont nécessairement des probabilités subjectives, qui peuvent toutefois être reliées à des probabilités objectives attribuées au monde réel par l'intermédiaire de principes tels que le « principe de Miller » [MIL 66]. Dans la théorie du changement de croyances, le raisonnement probabiliste consiste, à partir d'une distribution de probabilités initiale (a priori) P sur les états d'un système réel, et d'un message certain A reçu sur ce système, à calculer la distribution de probabilités finale (a posteriori) P^*_A . La méthode la plus connue pour effectuer ce calcul est la règle de Bayes, qui s'écrit : $P^*_A(B) = P(B,A) = P(B \cap A) / P(A)$. Le bayésianisme épistémique n'est autre que la doctrine qui consiste à affirmer que les degrés de croyance doivent être mesurés par des probabilités et que celles-ci doivent être révisées par la règle de Bayes. La règle de Bayes implique une réallocation homothétique des probabilités initiales des états vers les états conservés dans le changement consécutif au message. Elle est naturelle dans le contexte de probabilités objectives issues par exemple de données statistiques dans la mesure où elle traduit une conservation de proportions. Mais elle requiert une justification

supplémentaire dans le contexte de probabilités subjectives où l'homothétie ne repose pas sur un fondement cognitif aussi clair.

Plusieurs types de justification du raisonnement bayésien ont été proposés, de nature purement algébrique (par exemple [HECK 88]) ou de nature décisionnelle (arguments de type « Dutch Book » [SAV 54]). Ces justifications ne reposent toutefois sur aucun principe général de rationalité épistémique. La théorie du changement de croyances offrant un ensemble de tels principes, il est intéressant de la confronter à la règle de Bayes pour voir si celle-ci pourrait bénéficier d'une telle justification de nature épistémique. Le principe général de cette confrontation est le suivant :

1. On considère qu'une méthode de calcul de la fonction de probabilité révisée P^*_A (par exemple la règle de Bayes) constitue une sémantique pour le raisonnement probabiliste dans un contexte de changement ;

2. On considère qu'il existe des axiomes du raisonnement probabiliste qui doivent être respectés dans un contexte de changement ;

3. On définit des types de changement de croyances probabiliste en établissant des théorèmes de représentation de certaines axiomatiques par certaines méthodes de calcul ;

4. On compare les axiomatiques des différents types de changement probabiliste à celles des types de changement ensembliste pour voir dans quelle mesure ces dernières soutiennent les premières. Pour réaliser cette dernière opération, on définit des conventions de transcription des axiomes ensemblistes en axiomes probabilistes. Ces conventions peuvent être des « transcriptions faibles », qui transposent directement les axiomes ensemblistes en supposant que le « support » d'une fonction de probabilité, c'est-à-dire l'ensemble des mondes de probabilité non nulle, représente la croyance ensembliste associée à cette fonction. Elles peuvent être des « transcriptions fortes », interprétant de manière plus ou moins directe dans un cadre probabiliste l'intuition associée aux axiomes ensemblistes.

A titre d'exemple, la transcription faible des trois axiomes communs à la logique de la rectification et à la logique de la mise à jour énoncés au §3 est la suivante :

- axiome de succès probabiliste : $P^*_A(A) = 1$

- axiome de conservation probabiliste : si $P(A) = 1$ alors $(P^*_A(X) > 0 \Leftrightarrow P(X) > 0)$

- axiome d'inclusion probabiliste : $P(X \cap A) > 0 \Rightarrow P^*_A(X) = 1$

Une transcription forte de l'intuition associée à l'axiome de conservation ensembliste est la suivante :

- axiome de conservation probabiliste fort : si P(A) = 1 alors $P^*_A(X) = P(X)$

On peut alors établir les deux résultats fondamentaux suivants [WAL 01a] :

1. Dans les cas non triviaux où la croyance initiale est incertaine (elle comporte plus d'un monde possible), la règle de Bayes n'est compatible qu'avec un contexte de rectification des croyances. Dans un contexte de mise à jour, seule la règle dite d'« imaging » proposée par Lewis [LEW 76] est adéquate. Cette règle consiste à réaffecter la probabilité initiale de chaque monde w incompatible avec le message A au(x) monde(s) le(s) plus proche(s) de w et compatible(s) avec A. Elle viole un axiome caractéristique de la règle de Bayes, valable dans le seul contexte de rectification :

- axiome de préservation probabiliste : si P(A) > 0, alors ($P^*_A(X) > 0 \Rightarrow P(X) > 0$)

2. Dans un contexte de rectification, la règle de Bayes n'est pas la seule règle possible si on s'en tient à une transcription « faible » des principes ensemblistes. Une infinité d'autres règles de « conditionnalisation » sont admissibles, dont le point commun est qu'elles respectent le principe de préservation probabiliste précédent. Ainsi, la « règle égalitaire », qui répartit à parts égales la probabilité initiale de tous les mondes incompatibles avec le message entre les mondes de probabilité initiale positive compatibles avec le message, est une autre règle de conditionnalisation possible, valide dans un contexte de rectification des croyances. Pour obtenir un théorème de représentation de la règle de Bayes seule, il faut adopter une transcription « très forte » de l'axiome ensembliste suivant :

- axiome de distributivité à droite : $K^*(A \cup B) = (K^*A) \cup (K^*B)$

sous la forme linéarisée suivante, qui force la réallocation homothétique, mais n'a pas de réelle justification cognitive :

- axiome de Gärdenfors [GAR 88] :

si $A \cap A' = \varnothing$ alors $\exists\, a \in [0,1]$ tel que $\forall x$, $P^*_{AUA'}(X) = a\, P^*_A(X) + (1-a)\, P^*_{A'}(X)$

En résumé, les théories du changement de croyances permettent de mettre en évidence deux familles alternatives de raisonnement probabiliste en situation de changement, la conditionnalisation et l'imaging. Elles conduisent également à relativiser, à ce stade, la justification épistémique du privilège accordé à la règle bayésienne dans un contexte de rectification, sans toutefois qu'aucune autre méthode de conditionnalisation n'apparaisse comme plus naturelle.

9. La probabilisation des raisonnements non certains

Les sections 5 à 7 ont permis d'indiquer le rôle central des théories du changement de croyances pour appréhender différentes formes de raisonnement non classiques (le raisonnement non-monotone, les énoncés contrefactuels, l'abduction). La section 8 a montré que les deux contextes de changement des croyances pouvaient être associés à deux types de méthodes de révision probabiliste. Il est alors intéressant d'envisager la possibilité de probabiliser les raisonnements non classiques évoqués, dont le point commun est d'associer des prémisses (ou des antécédents) certaines à des conclusions (ou des conséquents) non certaines ou défaisables. Cette tentative de probabilisation peut relever de deux approches différentes, correspondant aux niveaux logiques différents de ces raisonnements.

La première approche a pour point de départ la relation classique de déduction logique $A \vdash B$, traduite par $A \subseteq B$. Il s'agit d'une relation d'inférence reliant une prémisse à une conclusion vraie dans tous les mondes où la prémisse est vraie. L'implication probabiliste de cette définition est triviale : si $A \subseteq B$ alors $P^*_A(B) = 1$. Une première méthode de probabilisation consiste alors à attribuer aux conclusions des raisonnements non-classiques une probabilité inférieure à 1.

Le raisonnement non-monotone représente un affaiblissement de l'inférence déductive tel que la conclusion B n'est pas nécessairement vraie dans tous les mondes de la prémisse A. La définition sémantique de l'inférence $A \mid\sim B$ exposée au §5 signifie qu'on s'attend seulement à ce que B soit vraie dans les mondes les plus « normaux » de A. On peut alors chercher à traduire cette idée en termes probabilistes par le fait que la probabilité conditionnelle, définie par la règle de Bayes, soit supérieure à un certaine seuil α (supérieur à 0.5) :

$A \mid\sim B$ ssi $P(B,A) \geq \alpha$

Elle exprime que les inférences non-monotones ont des conclusions « fortement probables », cette forte probabilité expliquant notre disposition à les accepter. Toutefois, cette interprétation intuitive se heurte à des difficultés, dont la plus notable est le « paradoxe de la loterie » [KYB 61] : elle conduit à violer un axiome naturel des logiques non-monotones :

-axiome de conjonction : si $A \mid\sim B$ et $A \mid\sim C$ alors $A \mid\sim B \cap C$

La violation de cet axiome conduit à rejeter l'interprétation probabiliste proposée ; en effet, on peut démontrer qu'aucun critère probabiliste utilisant une méthode de conditionnalisation représentable en termes de relations entre les seules probabilités des deux événements considérés ne peut échapper au paradoxe de la

loterie [ZWI 96]. D'autres outils de représentation de l'incertitude - qui sortent du cadre du présent papier - sont toutefois envisageables : les ε-probabilités d' Adams [ADA 75] ou les probabilités non additives (voir par exemple [BEN 97]) .

Le raisonnement abductif conduit à des conclusions plus faibles encore que celles du raisonnement non-monotone, dans la mesure où les hypothèses abduites ne sont pas forcément acceptées par l'agent et ne sont donc pas nécessairement « fortement probables ». Comme on l'a noté au §7, il est par ailleurs assez naturel de considérer que l'explication sous-jacente puisse être de nature probabiliste, comme dans la relation symptôme/maladie utilisée dans un diagnostic. Une interprétation simple de l'abduction en termes probabilistes qui tienne compte de ces deux remarques serait alors que :

$A \| \approx B$ ssi $P(A, B) > P(A)$

Elle exprime simplement que l'hypothèse abduite accroît la probabilité des faits qu'elle explique. Toutefois, cette interprétation n'est manifestement pas plus satisfaisante dans la mesure où elle est sujette à la version inversée du paradoxe de la loterie : elle ne respecte par l'axiome de conjonction à gauche (cf.§7), requis par toutes les formes de logique abductive évoquées. Là encore, d'autres voies de probabilisation sont à rechercher peut-être du côté des probabilités non-classiques ou du second-ordre (voir à ce sujet [GAR 88], chap.8).

La seconde approche a pour point de départ l'énoncé conditionnel de la logique classique, $A \rightarrow B$. Il s'agit d'une proposition reliée à l'inférence déductive par la version sémantique du théorème de la déduction : $A \subseteq B$ ssi $T \subseteq A \rightarrow B$. L'implication probabiliste de cette équivalence est encore triviale : $P^*_A(B)=1$ ssi $P(A \rightarrow B) = 1$. Une seconde méthode de probabilisation consiste alors à chercher à identifier un énoncé conditionnel $A \rightsquigarrow B$ qui permette de « gradualiser » ce résultat sous la forme générale suivante :

[test de STALNAKER] $P(A \rightsquigarrow B) = P^*_A(B)$

Ce test, proposé par Stalnaker [STA 70], est une probabilisation du test de Ramsay évoqué au §6. Il est facile de vérifier que ce test n'est pas satisfait si on interprète $A \rightsquigarrow B$ par l'implication logique $A \rightarrow B$ et si on calcule la probabilité a posteriori $P^*_A(B)$ par la règle de Bayes. Un résultat plus général, dû à Lewis [LEW 76], montre que le test de Stalnaker peut par contre être satisfait si on calcule la probabilité a posteriori $P^*_A(B)$ par la méthode de l'imaging et si on interprète $A \rightsquigarrow B$ par le contrefactuel $A > B$, défini par les axiomes de la logique C2 de Stalnaker (c'est à dire qu'il respecte l'axiome de tiers-exclu pour les contrefactuels). Un autre résultat, dû à Lepage [LEP 91], montre qu'on ne peut échapper à cette condition restrictive qu'en se plaçant dans le cadre d'une logique conditionnelle à

plusieurs valeurs, ce qui revient en fait à dissoudre la logique qualitative des conditionnels contrefactuels dans un cadre gradualiste.

En conclusion, si les théories du changement de croyances permettent de fournir un cadre synthétique pour représenter un certain nombre de raisonnements non classiques aux conclusions incertaines, la théorie classique des probabilités n'offre pas directement un tel cadre synthétique pour les mêmes raisonnements. Les possibilités déjà explorées ou à explorer pour relier ces concepts logiques à des mesures quantitatives conduisent soit à abandonner les logiques à deux valeurs, soit à adopter des mesures non-probabilistes. De fait, il est possible de montrer que ces dernières peuvent être des sémantiques adéquates de la logique non-monotone ou de la théorie du changement de croyances (voir par exemple [PEA 88], [MON 94], [BEN 97]).

Bibliographie

[ADA 75] ADAMS, E., *The logic of conditionals*, Dordrecht, Reidel, 1975.

[ALC 85] ALCHOURRON C. E., GÄRDENFORS P., MAKINSON, D., « On the logic of theory change : partial meet contraction and revision functions » , *Journal of Symbolic Logic*, 50, 510-530, 1985.

[BEN 97] BENFERHAT, S., DUBOIS, D., PRADE, H., « Nonmonotonic reasoning, conditional objects and possibility theory », *Artificial Intelligence*, 92, 259-276, 1997.

[BOU 95] BOUTILIER, C., BECHER, V., « Abduction as Belief Revision » , *Artificial Intelligence*, 77(1), 43-94, 1995.

[CHA 89] CHATEAUNEUF,A., JAFFRAY,J.Y., « Some characterization of lower probabilities and other monotone capacities through the use of Möbius Inversion », *Mathematical Social Sciences*, 17, 263-283.

[FLA 00] FLACH, P., KAKAS, A., *Abduction and induction*, Kluwer.

[GAB 85] GABBAY, D., « Theoretical Foundations for non-monotonic reasoning in expert systems », in K.R. Apt ed., *Proceedings NATO Advanced Study Institute on Logics and Models of Concurrent Systems*, Springer-Verlag, Berlin, 439-457, 1985.

[GAB 94] GABBAY, D., HOGGER, C.J., ROBINSON, J.A., *Handbook of Logic in Artificial Intelligence and Logic Programming*, Oxford Science Publication, coll. ed. by Gabbay, Hogger, Robinson, 1994.

[GAR 88] GÄRDENFORS P., *Knowledge in Flux*, MIT Press, 1988.

[GAR 94] GÄRDENFORS P., MAKINSON, D., « Nonmonotonic inference based on expectations », *Artificial Intelligence*, 65, 197-245, 1994.

[GOO 84] GOODMAN, N., *Fact, Fiction and Forecast*, Cambridge, Mass., 1984.

[GRA 91] GRAHNE, G., « Updates and counterfactuals », in Allen, J.A., Fikes, R., Sandwell, E. (eds.), *Principles of Knowledge Representation and Reasoning : Proceedings of the Second International Conference*, San Mateo, California, Morgan Kaufmann, 269-276, 1991.

[GRE 90] GREGOIRE, E., *Logiques non monotones et intelligence artificielle*, Hermes, 1990.

[HEC 88] HECKERMAN D.E., « An axiomatic framework for belief updates », in J.F.Lemmer, L.N. Kanal (eds), *Uncertainty in Artificial Intelligence 2*, , North Holland, Amsterdam, 11-22, 1988.

[HEM 65] HEMPEL, C.G., *Aspects of scientific explanation and other essays in the philosophy of science*, The Free Press, 1965.

[HEM 88] HEMPEL, C.G., « A Problem Concerning the Inferential Function of Scientific Theories », in A.Grünbaum and W.C.Salmon eds. *The Limitations of Deductivism*, University of California Press, 1988.

[JAC 91] JACKSON, F., *Conditionals,* Oxford University Press, 1991.

[KAT 92] KATSUNO A., MENDELZON A., « On the difference between updating a knowledge base and revising it », in P. Gärdenfors ed., *Belief Revision*, Cambridge University Press, 183-203, 1992.

[KRA 90] KRAUS, S., LEHMANN D., MAGIDOR, M., « Non monotonic reasoning, preferential models and cumulative logics », *Artificial Intelligence* , 44,167-208, 1990.

[KYB 61] KYBURG, H.E., *Probability and the Logic of Rational Belief*, Middletown, Conn., Wesleyan University Press, 1961.

[LEH 92] LEHMANN D., MAGIDOR, M. (1992), « What does a conditional base entail ? » *Artificial Intelligence*, 55, 1-60, 1992.

[LEW 73] LEWIS, D.K., *Counterfactuals*, Harvard University Press, 1973.

[LEW 76] LEWIS, D.K., « Probabilities of conditionals and conditional probabilities », *Philosophical Review*, 85, 297-315, 1976.

[MAY 96] MAYER, M.C., PIRRI, F. (1996), « Abduction is not Deduction-in-Reverse », *Journal of the IGPI*, 4(1), 1-14, 1996.

[MIL 66] MILLER D., « A paradox of information », *British Journal for the Philosophy of Science*, 17, 59-61, 1966.

[MON 94] MONGIN, P., « The logic of belief change and nonadditive probability », in Prawitz, D., Westertåhl (eds.), *Logic and Philosophy of Science in Uppsala*, 49-68, Kluwer Academic Publishers, 1994.

[MOO 88] MOORE, R.C., « Autoepistemic logic », in Smets P. et al. (eds), *Non Standard Logics for Automated Reasoning*, Acadamic Press, London, 105-136, 1988.

[NEU 33] NEURATH, O., «Protokollsätze, Erkentniss 3», 1933; trad.fr. « Enoncés protocolaires », in *Manifeste du Cercle de Vienne et autres écrits*, sous la direction de Antonia Soulez, P.U.F., 1985.

[NUT 80] NUTE, D., *Topics in Conditional Logic*, Reidel, Dordrecht, 1980.

[PEA 88] PEARL, J., *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo, CA, 1988.

[PEI 31] PEIRCE, C. S., *Collected papers of Charles Sanders Peirce*, ed. by C. Hartshorne & P. Weiss, Harvard University Press, 1931-1958.

[POP 59] POPPER, K.R., *The Logic of Scientific Discovery*, Hutchinson & Co., 1959.

[RAM 50] RAMSAY, F.P., « General Propositions and Causality », in Ramsay ed., *Foundations of Mathematics and Other Logical Essays*, 237-257, New-York, 1950.

[REI 80] REITER, R., « A logic for default reasoning », *Artificial Intelligence*, 13, 81-131, 1980.

[SAV 54] SAVAGE, L.J., *Foundations of Statistics*, Macmillan, New York, 1954.

[QUI 72] QUINE, W.V.O., *Méthodes de logique*, Armand Colin, 1972

[SHA 76] SHAFER G., *A Mathematical Theory of Evidence,* Princeton University Press, 1976.

[SHO 87] SHOHAM, Y., « A semantical approach to nonmonotonic logics », in *Proceedings Logics in Computer Science*, Ithaca, NY, 275-279, 1987.

[STA 68] STALNAKER, R., « A theory of conditionals », in N.Rescher ed., *Philosophical Quarterly Monograph Series*, 2, Basil Blackwell, 1968.

[STA 70] STALNAKER, R., « Probability and conditionals », *Philosophy of Science*, 37, 64-80, 1970.

[WAL 01a] WALLISER B., ZWIRN D., Can Bayes rule be justified by cognitive rationality principles ?, mimeo ENPC, 2001.

[WAL 01b] WALLISER B., ZWIRN, D., ZWIRN,H., Abductive logics in a belief revision framework, mimeo ENPC, 2001.

[ZWI 96] ZWIRN D., ZWIRN H., « Metaconfirmation », *Theory and Decision*, 3, 195-228, 1996.