– « il *serait* bon que p soit vrai » est différent de « il *est* / *était* / *sera* bon que p soit vrai » ; c'est un constat effectué dans le présent mais dirigé vers un futur hypothétique ;

– « il serait bon que p soit vrai » est une notion relative à un monde de référence donné ;

– au sens où nous l'entendons, « il serait *bon* (respectivement *mauvais*) pour un agent que p soit vrai » signifie que p est actuellement faux, et que les états dans lesquels il est vrai sont nécessairement *meilleurs* (respectivement *pires*) que l'état actuel ;

– nous imposons qu'une chose ne soit pas à la fois bonne et mauvaise. Par exemple, si prendre un peu de poison serait bon, et en prendre beaucoup mauvais, nous ne pouvons rien en déduire sur la prise de poison en général. Nous n'affirmons donc pas que ce serait bon et mauvais, mais plutôt que ce ne serait ni bon ni mauvais. Cela revient à considérer que « il serait bon (resp. mauvais) que p soit vrai » implique « il ne serait pas mauvais (resp. bon) que p soit vrai » mais que la réciproque n'est pas valide. En effet p peut n'être ni bon ni mauvais (par exemple, les fatalités étant toujours soit vraies soit fausses, cela n'a pas de sens de dire qu'il serait bon (ou mauvais) qu'elles se réalisent puisqu'elles le sont déjà). Plus généralement, certaines choses peuvent être *indifférentes* au sens où elles peuvent être vraies dans certains mondes meilleurs et fausses dans les autres (mondes meilleurs), et pareillement pour les mondes pires.

– la propriété précédente impose que « Il serait *bon* (resp. *mauvais*) que p soit vrai » si et seulement si p est vrai dans au moins un monde *meilleur* (resp. *pire*) et est faux dans tous les mondes *pires* (resp. *meilleurs*).

Avant de définir les notions de bon et de mauvais, nous devons donc avant tout définir ce qu'est un monde meilleur et ce qu'est un monde pire. C'est l'objet de la section suivante.

3 Les préférences sur les mondes

Sémantiques. Soit W un ensemble quelconque de mondes classés les uns par rapport aux autres selon la relation $\preccurlyeq_i$: $w \preccurlyeq_i w'$ ssi w est un monde au mieux aussi bon que w' pour un agent i donné. $\preccurlyeq_i$ est un préordre total sur les mondes de W (relation réflexive et transitive).

Par définition, $w \succcurlyeq_i w'$ (préordre total) ssi $w' \preccurlyeq_i w$; $w \prec_i w'$ (relation transitive) ssi $w \preccurlyeq_i w'$ et non $w \succcurlyeq_i w'$; $w \succ_i w'$ (relation transitive) ssi $w' \prec_i w$; $w \sim_i w'$ (relation d'équivalence) ssi $w \preccurlyeq_i w'$ et $w \succcurlyeq_i w'$. (L'ordre étant total, il est impossible que non $w \preccurlyeq_i w'$ et non $w \succcurlyeq_i w'$ simultanément.)

Ainsi, les mondes possibles sont répartis en trois catégories par rapport au monde courant w : ceux qui sont meilleurs (*i.e.* $\{w' | w' \succ_i w\}$) ; ceux qui sont pires (*i.e.* $\{w' | w' \prec_i w\}$) ; et ceux qui sont équivalents (*i.e.* $\{w' | w' \sim_i w\}$). Remarque : par définition, $\preccurlyeq_i$ et $\succcurlyeq_i$ sont des relations inverses l'une par rapport à l'autre.

Pour chaque relation $R_i \in \{\preccurlyeq_i, \succcurlyeq_i\}$ nous définissons un opérateur modal comme suit : $w \Vdash \Box_{R_i} A$ ssi $\forall w'(w' \in W \ \& \ w' R_i w \Rightarrow w' \Vdash A)$ (et ce pour tout agent i). Ainsi, par exemple, $\Box_{\preccurlyeq_i} A$ se lit « pour l'agent i, A est vrai dans tous les mondes au moins aussi mauvais que le monde actuel ». On définit également les opérateurs duaux correspondants :

$$\Diamond_{R_i} A \stackrel{déf}{=} \neg \Box_{R_i} \neg A.$$

REMARQUE — En supposant que nous disposions d'opérateurs $\Box_{\prec_i}$ alors $A \wedge \Box_{\prec_i} A$ (signifiant « A est vrai dans le monde actuel et dans tous les mondes (strictement) pires que le monde actuel ») est différent de $\Box_{\preccurlyeq_i} A$ (qui signifie « A est vrai dans le monde actuel et dans tous les mondes équivalents ou (strictement) pires que le monde actuel »). De façon similaire, $A \wedge \Box_{\succ_i} A$ et $\Box_{\succcurlyeq_i} A$ sont également différents.

Axiomatique. $\Box_{\preccurlyeq_i}$ et $\Box_{\succcurlyeq_i}$ sont définis dans des logiques modales normales. À ce titre, pour toute relation $R_i \in \{\preccurlyeq_i, \succcurlyeq_i\}$ nous avons [4], les règles de nécessitation de type

$$\frac{A}{\Box_{R_i} A} \qquad (\mathrm{RN}_{R_i})$$

et les schémas d'axiomes de type

$$\Box_{R_i} A \wedge \Box_{R_i}(A \rightarrow B) \rightarrow \Box_{R_i} B \qquad (\mathrm{K}_{R_i})$$

Nous ajoutons à ce système les schémas d'axiomes suivants :

$$\Box_{\preccurlyeq_i} A \rightarrow A \qquad (\mathrm{T}_{\Box_{\preccurlyeq_i}})$$
$$\Box_{\preccurlyeq_i} A \rightarrow \Box_{\preccurlyeq_i} \Box_{\preccurlyeq_i} A \qquad (4_{\Box_{\preccurlyeq_i}})$$

($\mathrm{T}_{\Box_{\preccurlyeq_i}}$) traduit la réflexivité et ($4_{\Box_{\preccurlyeq_i}}$) la transitivité des relations $\Box_{\preccurlyeq_i}$. Nous étendons également la logique par les deux axiomes d'interaction qui traduisent le fait que $\preccurlyeq_i$ et $\succcurlyeq_i$ sont des

relations inverses :

$$A \rightarrow \Box_{\succcurlyeq_i} \Diamond_{\preccurlyeq_i} A \qquad (\mathrm{I}_{\Box_{\succcurlyeq_i}, \Diamond_{\preccurlyeq_i}})$$
$$A \rightarrow \Box_{\preccurlyeq_i} \Diamond_{\succcurlyeq_i} A \qquad (\mathrm{I}_{\Box_{\preccurlyeq_i}, \Diamond_{\succcurlyeq_i}})$$

Théorèmes. $\mathrm{T}_{\Box_{\succcurlyeq_i}}$ et $4_{\Box_{\succcurlyeq_i}}$ sont démontrables à partir de la l'axiomatique précédente, ainsi que :

$$\Diamond_{\preccurlyeq_i} A \rightarrow \Box_{\succcurlyeq_i} \Diamond_{\preccurlyeq_i} A \qquad (1)$$
$$\Diamond_{\succcurlyeq_i} A \rightarrow \Box_{\preccurlyeq_i} \Diamond_{\succcurlyeq_i} A \qquad (2)$$

Complétude. Nos modèles sont de la forme $\langle W, R_{\succcurlyeq}, R_{\preccurlyeq}, V \rangle$ où W est un ensemble de mondes possibles, $R_{\preccurlyeq}$ et $R_{\succcurlyeq}$ associent des relations d'accessibilité $\preccurlyeq_i$ et $\succcurlyeq_i$ à chaque agent i, et V associe une valuation à chaque monde possible.

Il est aisé de vérifier que chaque axiome correspond à sa contrepartie sémantique, et que tous les axiomes font partie de la classe de Sahlqvist [14, 2] pour laquelle il existe un résultat de complétude général.

Exemple courant. L'agent i est dans le désert et souhaite boire l'eau d'un puits. Pour cela il a besoin d'un puits p, d'un seau s et d'une corde c pour plonger et récupérer le seau dans le puits. S'il ne dispose pas d'un objet, par exemple la corde c, nous notons $\neg c$.
On peut raisonnablement supposer que l'état w_0 où il n'a aucun objet est le pire de tous ; ceux où il dispose d'un seul objet sont meilleurs (w_1 à w_3) ; ceux où il dispose de deux objets sont encore meilleurs que les précédents (w_4 à w_6) ; le meilleur état de tous étant celui où il peut boire (w_7). Tous les états où il possède un seul objet sont équivalents ; de même ceux où il en possède deux. Ce préordre est représenté dans la figure 1.

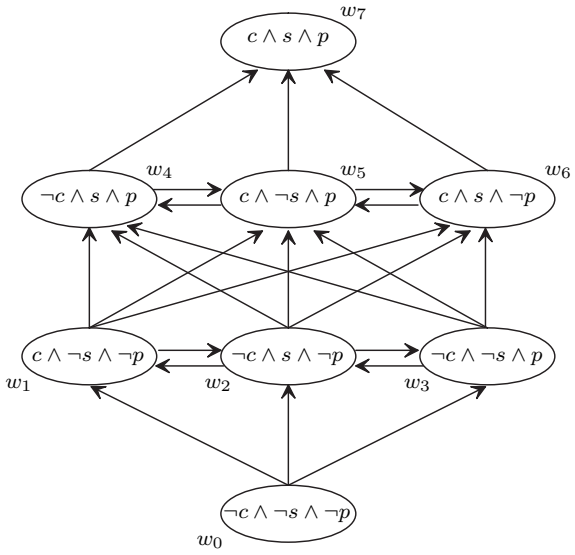

FIG. 1 – Exemple d'ordre total $\preccurlyeq_i$ pour le problème de l'eau

4 Les opérateurs $Good_i$ et Bad_i

4.1 De leurs relations avec les états

Définitions. Nous formalisons « Il serait bon pour l'agent i que A soit vrai » comme suit :

$$Good_i A \stackrel{déf}{=} \Box_{\preccurlyeq_i} \neg A \wedge \Diamond_{\succcurlyeq_i} A \qquad (\mathrm{Déf}_{Good_i A})$$

Autrement dit, « actuellement, il serait bon pour l'agent i que A soit vrai » signifie par définition que (1) A est faux dans le monde actuel et dans tous les mondes qui lui sont équivalents ou pires et (2) qu'il existe au moins un monde meilleur dans lequel A est vrai.

Selon cette définition, « il *serait* bon que A soit vrai » signifie que A ne peut être vrai que dans des mondes *strictement meilleurs*. On pourrait songer à étendre cette définition aux mondes *meilleurs ou équivalents*). Il suffirait pour cela de remplacer $\Box_{\preccurlyeq_i} A$ par $\neg A \wedge \Box_{\prec_i} \neg A$ dans ($\mathrm{Déf}_{Good_i A}$). Mais une telle définition entraînerait par exemple $w_1 \Vdash Good_i s \wedge Good_i p$ (autrement dit : il serait bon d'avoir l'un des objets que l'on n'a pas déjà, au risque de perdre celui qu'on a). Cette définition correspond plutôt à « il *pourrait être* bon que A soit vrai ». Par ailleurs, il ne serait plus le cas que toute chose qui est bonne dans un monde, l'est également dans tout monde équivalent (et inversement) ; pourtant, cette propriété paraît tout à fait intuitive.

Par ailleurs, imposer que A soit vrai dans tous les mondes meilleurs conduirait à une définition beaucoup trop restrictive. Dans notre exemple courant, $w_1 \Vdash Good_i(c \wedge s) \wedge Good_i(c \wedge p) \wedge Good_i(s \wedge p)$ (*i.e.*, à partir du moment où on ne dispose que d'une des choses nécessaires pour boire au puits, il serait bon d'en avoir deux). Si nous remplaçons $\Diamond_{\succcurlyeq_i}$ par $\Box_{\succ_i}$ [2] dans ($\mathrm{Déf}_{Good_i A}$), non seulement cette conjonction ne

[2] Dans ($\mathrm{Déf}_{Good_i A}$) nous utilisons $\Diamond_{\succcurlyeq_i}$ en place de $\Diamond_{\succ_i}$ (qui serait sémantiquement suffisant, puisque $\Box_{\preccurlyeq_i} \neg A$ assure que A est faux dans le monde actuel) car techniquement, l'irréflexivité n'est pas définissable en logique modale [2] (bien que l'on puisse obtenir une preuve de complétude en rajoutant éventuellement une règle d'inférence non standard). Autrement dit, on pourrait utiliser $\Diamond_{\succ_i}$ mais $\Diamond_{\succcurlyeq_i}$ est techniquement plus simple à utiliser sans pour autant modifier les propriétés sémantiques de l'opérateur $Good_i$.

serait pas satisfaite par w_1 (ce qui reviendrait à dire qu'il ne serait pas bon pour l'agent d'avoir deux objets plutôt qu'un seul), mais il ne serait également pas bon d'en avoir trois !

Enfin, comme la relation $\preccurlyeq_i$ sur laquelle elle est fondée, notre notion de bon est relative à un agent. Ainsi, dans notre exemple courant, si l'agent est non plus dans le désert mais à côté d'un robinet, on peut raisonnablement dire que les mondes w_0 à w_7 sont équivalents.

De façon similaire, nous définissons :

$$Bad_i A \stackrel{déf}{=} \Box_{\succcurlyeq_i} \neg A \wedge \Diamond_{\preccurlyeq_i} A \qquad (\text{Déf}_{Bad_i A})$$

auquel les mêmes remarques précédentes s'appliquent (en inversant les rôles de $\preccurlyeq_i$ et $\succcurlyeq_i$).

Théorèmes. Il découle de ($\text{Déf}_{Good_i A}$) et de ($\text{Déf}_{Bad_i A}$) les théorèmes suivants [3] :

– Les choses logiquement impossibles ou tautologiques sont des fatalités et en accord avec nos hypothèses (cf. Section 2) elles ne sont ni bonnes ni mauvaises :

$$\neg Good_i(\bot) \tag{3}$$
$$\neg Good_i(\top) \tag{4}$$

– Dans le cas où il serait bon que A entraîne B, alors dans le cas où il serait bon que A soit vrai, il serait bon que B soit vrai :

$$Good_i(A \rightarrow B) \rightarrow (Good_i A \rightarrow Good_i B) \tag{5}$$

– Une chose et son contraire ne peuvent pas être bonnes simultanément :

$$Good_i A \rightarrow \neg Good_i \neg A \tag{6}$$

– S'il était bon que $A \vee B$ soit vrai, alors il serait bon que l'un des deux au moins soit vrai (voire, les deux) :

$$Good_i(A \vee B) \rightarrow Good_i A \vee Good_i B \tag{7}$$

– Quand il est simultanément bon que A soit vrai et que B soit vrai, alors il est bon que $A \vee B$ soit vrai :

$$Good_i A \wedge Good_i B \rightarrow Good_i(A \vee B) \tag{8}$$

[3] Les mêmes propriétés sont vérifiées par l'opérateur Bad_i en inversant les symboles $\preccurlyeq_i$ et $\succcurlyeq_i$.

– Quand deux choses A et B sont simultanément bonnes alors qu'isolément $\neg B$ est bonne, c'est que le caractère bon de A est indépendant de celui de B :

$$Good_i(A \wedge B) \wedge Good_i \neg B \rightarrow Good_i A \tag{9}$$

Par exemple, si pour une personne malade « il serait bon d'être malade et d'être milliardaire », et « il serait bon de n'être pas malade », alors « il serait bon pour cette personne d'être milliardaire en général, indépendamment de sa maladie ».

– Quand deux choses sont bonne séparément, et qu'il existe un monde meilleur dans lequel leur conjonction est vraie, c'est qu'il serait bon que ces deux choses soient simultanément vraies :

$$Good_i A \wedge Good_i B \wedge \Diamond_{\succcurlyeq_i}(A \wedge B) \rightarrow Good_i(A \wedge B) \tag{10}$$

– Si actuellement il était bon que A soit vrai, alors ce serait également le cas dans tout monde équivalent ou pire :

$$Good_i A \rightarrow \Box_{\preccurlyeq_i} Good_i A \tag{11}$$

– De façon similaire, si actuellement il n'est pas bon que A soit faux, ce ne sera jamais le cas dans un monde équivalent ou meilleur :

$$\neg Good_i A \rightarrow \neg \Diamond_{\succcurlyeq_i} Good_i A \tag{12}$$

– Si une chose est bonne alors elle n'est pas mauvaise :

$$Good_i A \rightarrow \neg Bad_i A \tag{13}$$

Propriétés non valides.

(i) $Good_i A \rightarrow Good_i(A \vee B)$

(ii) $Good_i(A \wedge B) \rightarrow Good_i A$

(iii) $Good_i A \wedge Good_i B \rightarrow Good_i(A \wedge B)$

(i) pourrait sembler souhaitable au premier abord (dans le sens de la logique classique : si A est vrai, alors $A \vee B$ l'est). Néanmoins il n'en est rien. Cela provient du fait que $Good_i A$ indique que A devrait être vrai (dans un hypothétique monde futur), mais dit également que A est actuellement faux. De la même manière, $Good_i(A \vee B) \rightarrow (\neg A \wedge \neg B)$. Cette propriété est invalidée dès lors que nous avons $Good_i A \wedge B$.

Le fait que (ii) ne soit pas valide signifie que l'opérateur $Good_i$ n'est pas monotone : s'il était le cas qu'il serait bon que deux choses soient simultanément vraies, il ne serait pas forcément

bon pour autant qu'elles le soient séparément. Dans notre exemple courant, dans w_1 il serait bon d'avoir une corde et un puits, mais il ne serait pas bon d'avoir une corde (on l'a déjà).

(iii) n'est pas valide (il suffit de se placer dans le cas où il n'existe pas de monde meilleur satisfaisant simultanément A et B mais où il en existe les satisfaisant séparément). Par exemple, s'il fait -10 degrés, il serait bon qu'il fasse 24 degrés et il serait bon qu'il fasse 25 degrés. Ainsi, s'il est impossible qu'il fasse 24 et 25 degrés simultanément, alors cela ne pourrait pas être bon. (À comparer au théorème (10).)

4.2 De leurs relations avec l'action

Des résultats en psychologie cognitive montrent que le bon est un critère de choix dans la prise de décision d'un agent. Il est donc essentiel de pouvoir caractériser les bonnes et les mauvaises actions. Dans ce qui suit, nous étudions dans quelle mesure les résultats précédents peuvent être appliqués aux actions. Ce n'est pas tant l'exécution de ces dernières qui nous intéresse que leur résultat (en termes d'effets sur les mondes atteints). C'est pourquoi nous cherchons à relier les notions de bon et de mauvais aux actions *via* leurs effets.

Sémantique de l'action. Une action est considérée comme une transition entre deux mondes [7]. Soit $ACT = \{\alpha, \beta ...\}$ l'ensemble des actions.

La formule $[\alpha]A$ exprime qu'après toute exécution de α la formule A est vraie. $[\alpha]\bot$ exprime que α est inexécutable.

$$\langle\alpha\rangle A \stackrel{déf}{=} \neg[\alpha]\neg A$$

signifie que l'action α est exécutable après quoi A sera vrai, et $\langle\alpha\rangle\top$ que α est exécutable.

La formule $[\alpha^{-1}]A$ exprime qu'avant toute exécution de α la formule A est vraie.

$$\langle\alpha^{-1}\rangle A \stackrel{déf}{=} \neg[\alpha^{-1}]\neg A$$

signifie que l'action α vient d'être exécutée avant quoi A était vrai, et $\langle\alpha^{-1}\rangle\top$ que α vient d'être exécutée.

Pour toute action $\alpha \in ACT$ il existe une relation $R_\alpha : ACT \longrightarrow (W \longrightarrow 2^W)$ associant des ensembles de mondes $R_\alpha(w)$ à w ; on note R_α^{-1} la relation inverse. Les conditions de vérité sont : $w \Vdash [\alpha]A$ si et seulement si $w' \Vdash A$ pour tout monde $w' \in R_\alpha(w)$; $w \Vdash [\alpha^{-1}]A$ si et seulement si $w' \Vdash A$ pour tout monde $w' \in R_\alpha^{-1}(w)$.

Axiomatique. Pour toute action α, nous disposons des règles de nécessitation RN_α et $\text{RN}_{\alpha^{-1}}$ ainsi que des axiomes K_α et $\text{K}_{\alpha^{-1}}$ (voir (K_{R_i}) et (RN_{R_i}), avec $[\alpha]$ et $[\alpha^{-1}]$ se substituant respectivement à $\Box_{R_i}$).

Afin de rendre compte du fait que R_α et R_α^{-1} sont des relations inverses l'une par rapport à l'autre, nous ajoutons également les deux schémas d'axiomes suivants :

$$A \rightarrow [\alpha]\langle\alpha^{-1}\rangle A \qquad (\text{I}_{[\alpha],\langle\alpha^{-1}\rangle})$$
$$A \rightarrow [\alpha^{-1}]\langle\alpha\rangle A \qquad (\text{I}_{[\alpha^{-1}],\langle\alpha\rangle})$$

Complétude. On ajoute R à notre modèle qui associe à chaque action de ACT les relations d'accessibilité R_α et R_α^{-1}. Les nouveaux axiomes font également partie de la classe de Sahlqvist, et le résultat de complétude établi précédemment est préservé.

Lois d'action. Une action α est caractérisée par ses conditions d'exécutabilité et ses effets. Les premières sont formalisées par des *lois d'exécutabilité* de la forme $A \leftrightarrow \langle\alpha\rangle\top$ signifiant que « l'action α est exécutable si et seulement si la condition d'exécutabilité A est vraie ».[4] Les seconds le sont par des *lois d'effet* de la forme $A \rightarrow [\alpha]A'$ et signifiant « si on se trouve dans un contexte d'exécution où A est vrai, alors après toute exécution de l'action α la postcondition A' est vraie ». Autrement dit, le contexte d'exécution d'une loi d'effet influence l'effet d'une action sans préjuger du fait qu'elle a été exécutée ou non ; la précondition d'une loi d'exécutabilité indique à quelle condition cette action pourra être exécutée. L'ensemble des lois d'exécutabilité et des lois d'effet constitue l'ensemble des *lois d'action* noté LAW. Techniquement, les lois d'action sont des axiomes globaux.

EXEMPLE (LÀ OÙ UN GÉNIE INTERVIENT) — Dans notre exemple courant, supposons qu'un bon génie propose à l'agent de lui fournir les objets manquants, mais uniquement si celui-ci possède au moins la corde. Si α est l'action de fournir les objets manquants, alors on a les lois d'action

[4] Une forme plus faible consiste à ne donner qu'une condition suffisante (mais non nécessaire) : l'action peut alors être exécutable dans d'autres conditions non spécifiées par la loi d'exécutabilité. Nous ne traitons pas ce cas ici. Par ailleurs, l'équivalence permet d'obtenir une solution élégante au problème du décor (voir [13, 15]).

suivantes :

$$c \rightarrow \langle\alpha\rangle\top$$
$$\neg s \rightarrow [\alpha]s$$
$$\neg p \rightarrow [\alpha]a$$

(Pour que le génie puisse agir il suffit que l'agent ait une corde, et en fonction des objets que celui-ci possède déjà, l'effet de l'action du génie sera différent : soit il donnera le seau, soit le puits, soit les deux.) Dans notre exemple courant, α est exécutable dans w_1, w_5, w_6 et w_7. Dans tous les cas, après que α ait été exécutée, l'agent se retrouvera dans un monde où $c \wedge s \wedge p$ est vrai.

Première tentative de définition de ce qu'il serait bon de faire. La première idée qui s'impose est de définir « Il serait bon de faire α » (que nous notons simplement $Good_i\alpha$ pour l'instant) à partir de $Good_iA$ où A serait une formule rendant compte de l'accomplissement (futur ou passé) de α. Quatre cas sont possibles :

1. $Good_i\alpha \stackrel{déf}{=} Good_i\langle\alpha\rangle\top$ signifie « il serait bon que α soit exécutable », ce qui implique en particulier, par définition, que α est inexécutable dans tout monde pire que le monde actuel ce qui semble pour le moins contre-intuitif. Qui plus est, on ne cherche pas à traduire l'idée selon laquelle il serait bon qu'on *puisse faire* α, mais qu'on *fasse* α ;

2. $Good_i\alpha \stackrel{déf}{=} Good_i\langle\alpha^{-1}\rangle\top$ signifie « il serait bon que α vienne d'être exécutée », ce qui implique que dans tous les mondes pires que le monde actuel, α n'a pas été exécutée. Là encore, cela va à l'encontre de notre intuition car cette contrainte est trop forte ; c'est l'exécution d'une action dans le monde actuel qui nous intéresse, et non pas celle dans un monde meilleur ou pire ;

3. $Good_i\alpha \stackrel{déf}{=} Good_i[\alpha]\bot$ signifie « il serait bon que α soit inexécutable », ce qui va à l'encontre du sens que l'on cherche à capturer. De plus, cela implique qu'actuellement α est exécutable, ce qui est une contrainte trop forte (il suffit que α le soit dans le monde dans lequel elle doit être exécutée).

4. $Good_i\alpha \stackrel{déf}{=} Good_i[\alpha^{-1}]\bot$ signifie « il serait bon que α ne vienne pas d'être exécutée », ce qui là encore ne correspond intuitivement pas au sens que l'on cherche à capturer.

Aucune de ces définitions ne peut donc capturer, par la simple substitution d'une modalité de la logique dynamique à A, une définition formelle de $Good_i\alpha$ à partir de $Good_iA$. Il faut donc chercher une définition plus complexe.

Seconde tentative. La notion que nous cherchons à capturer est « il serait bon de faire α ». Dans le cadre de $Good_iA$, le fait que A soit bon était établi comparativement au monde actuel. Dans le cas d'une action, on est face à une alternative : soit nous comparons les mondes atteints à la suite de l'exécution de α à des mondes futurs alternatifs (ceux atteints si α n'avait pas été exécutée), soit nous les comparons au monde actuel.

Pour les raisons évoquées précédemment, la notion de bon doit conduire l'agent vers des mondes strictement meilleurs. Le premier choix impose donc que les mondes atteints suite à l'exécution de α doivent être meilleurs que ceux atteints en faisant autre chose que α (ou en ne rien faisant). Il n'est pas réaliste d'envisager de comparer le résultat de chacune des actions par rapport à toutes les autres. Et se limiter par exemple aux actions exécutables ne fait pas de sens car la notion que nous cherchons à capturer ne stipule pas, intuitivement, que l'action qu'il serait bon de faire doive être exécutable dans le monde actuel. De plus, cette définition interdirait que deux actions soient simultanément bonnes à exécuter. Enfin, cette définition semble plutôt caractériser la meilleure chose à faire (et non celles qu'il serait simplement bon de faire).

Le second choix impose que les mondes atteints suite à l'exécution de α doivent être meilleurs que le monde actuel. Ceci est cohérent avec la précédente notion de bon (portant sur une formule). Par exemple, un agent propose un jeu à un autre agent : « si tu ne fais rien, je te prends 100 €. En revanche si tu fais quelque chose, je ne t'en prends que 50 ». Il ne fait aucun doute qu'il serait meilleur de faire quelque chose que de ne rien faire. Mais nous n'en déduisons pas pour autant qu'il serait bon pour lui de faire quelque chose, puisqu'il perd tout de même de l'argent. Nous disons plutôt qu'il serait mauvais pour lui de ne rien faire, et qu'il serait moins mauvais pour lui de faire quelque chose. À l'inverse, dans « si tu ne fais rien, je te donne 100 €. Si tu fais quelque chose, je ne t'en donne que 50 », il serait bon de faire quelque chose, bien qu'il soit le cas qu'il serait encore meilleur de ne rien faire. Ce choix correspond donc bien à notre intuition.

REMARQUE — En raisonnement sur les actions, il est commun de faire l'hypothèse que toute action peut être découpée en une action purement ontique (à effet exclusivement sur le monde) et une action purement épistémique (à effet exclusivement sur les états mentaux). Par exemple, l'action de jouer à pile ou face peut se décomposer en l'action de jeter la pièce (action ontique) et l'action de regarder le résultat (action épistémique).
Nous ne considérons pour l'instant que les actions purement ontiques, non déterministes (dont l'effet est tiré aléatoirement dans un ensemble d'effets possibles) et conditionnelles (dont les effets changent selon le contexte). Une action a donc en général plusieurs mondes conséquents possibles.
Dans le cas des actions non déterministes la notion de bon prend toute son importance. Par exemple, un jeu fait gagner dans 99 % des cas un lot de 1000 €, et dans seulement 1 % perdre 1 €. Un simple calcul d'espérance de gain indique qu'il serait bon de jouer. Seulement, il existe au moins un monde pire où le joueur perd 1 €. Et même en recommençant n fois, il existe toujours une probabilité, aussi infime soit-elle, qu'il perde les n fois consécutives. Il ne serait donc pas bon (ni mauvais) dans l'absolu de jouer à ce jeu. Ce résultat, s'il est fort, n'est pas contre-intuitif.

Nous retenons donc le second choix et définissons qu'il serait bon de faire une certaine action si et seulement si tous les mondes atteints suite à l'exécution de cette action sont meilleurs que le monde actuel. Autrement dit, il serait bon d'exécuter une action si et seulement si il serait bon ques ses effets soient réalisés.

Dans notre exemple courant, l'action du génie conduit à ce que l'agent dispose de tous les objets dont il a besoin. Selon ce qui précède, il serait bon qu'il exauce ce voeu dans tous les mondes possibles, à l'exception de w_7 (où son effet est déjà réalisé).

Définition. Pour une action α donnée, soient

$$A_1 \rightarrow [\alpha]B_1,$$
$$A_2 \rightarrow [\alpha]B_2,$$
$$\dots$$

ses lois d'effet. Alors :

$$Good_i\alpha \stackrel{déf}{=} (A_1 \wedge Good_iB_1) \vee (A_2 \wedge Good_iB_2) \vee \dots \quad (\text{Déf}_{Good_i\alpha})$$

Autrement dit, ($\text{Déf}_{Good_i\alpha}$) signifie qu'il serait bon de faire une certaine action α si et seulement si il y a au moins (1) un contexte d'exécution A_k qui est actuellement vrai et (2) qu'il serait bon que l'effet B_k associé à ce contexte soit vrai.

Nous n'imposons pas qu'une bonne action soit nécessairement exécutable, tout comme nous n'imposions pas qu'une bonne chose soit réalisable. Cette logique accepte donc des phrases du genre : « il serait bon pour l'agent i qu'il nage » tout comme elle accepte : « il serait bon pour i qu'il soit milliardaire », étant donné que i ne sait pas nager et ne peut pas devenir milliardaire.

Cette définition traduit entièrement la définition informelle précédente grâce au fait que la définition du $Good_i$ tient compte des mondes incomparables. En effet, si ce n'était pas le cas, on tiendrait pour bonnes des actions qui mènent vers des mondes incomparables, ce qui est contre-intuitif puisque par définition, ça ne fait pas de sens de juger s'il serait bon d'accomplir une action si on ne peut pas comparer les effets de cette action par rapport au monde où elle est exécutée.

Le jugement d'une suite d'actions ne porte que sur la comparaison entre le monde actuel et l'ensemble des mondes conséquents de la suite d'actions, sans regard sur les mondes intermédiaires. Cette définition est intuitive et permet de traduire la notion de bon sur le long terme.

4.3 De leurs relations avec les croyances

Jusqu'à présent, les notions de bon et de mauvais ont été définies pour un agent, mais indépendamment de ses croyances. En quelque sorte, on décrivait de façon objective ce qui était bon (et mauvais) pour un agent donné. Cependant, l'intérêt de ces notions est qu'un agent autonome puisse par lui-même juger de ce qui est bon ou mauvais, pour lui ou pour d'autres agents. Nous voulons donc rendre les notions de bon et de mauvais subjectives, ce qui nécessite l'introduction d'un ensemble d'opérateurs doxastiques (*i.e.* relatifs à la croyance).

Nous entendons par *croyance* le savoir subjectif : si un agent croit A, alors c'est que pour lui A est vrai dans tous les mondes possibles constituant une alternative au monde réel (ses mondes *épistémiques*).

Le jugement des actions ne traduit pas le prin-

cipe que la fin justifie les moyens, bien qu'une bonne fin puisse être atteinte en passant par des étapes intermédiaires mauvaises. En effet, nous avons justifié qu'une fin qui présente à la fois des bénéfices et des maux ne peut être jugée en termes de bon et de mauvais. Mais l'agent ne peut se référer qu'à ses croyances (et donc, se tromper). Par exemple, un agent peut croire qu'il serait bon pour lui de manger de la viande crue car il ignore qu'il est possible qu'elle soit contaminée (cette dernière possibilité ayant pour conséquence, selon notre définition, que manger de la viande crue n'est pas bon dans l'absolu).

Un agent est amené à modifier ses croyances au cours de ses expériences, soit dans le sens d'un affinement de ses connaissances à mesure de son apprentissage, soit par des bouleversements du monde ou de lui-même. Ainsi, un agent peut découvrir que le feu brûle et classer les mondes où il est brûlé comme pires que ceux où il ne l'est pas. Ou encore, avec le temps, son organisme peut devenir résistant à des maladies, et ce qu'il croyait mauvais peut devenir indifférent. Ce problème se rapporte au problème plus général de la révision et de la mise à jour des croyances [1, 5, 11] pour lesquels nous disposons de quelques solutions [8, 9].

Sémantique de la logique épistémique. Soit $\mathcal{B}_i$ la relation d'accessibilité de la logique épistémique [10, 6, 12]. $\mathcal{B}_i(w)$ est l'ensemble des mondes que l'agent i ne sait distinguer du monde actuel w. En d'autres termes, c'est l'ensemble des mondes dans lesquels l'agent i croit être.

La condition sémantique est la suivante : $w \Vdash Bel_i A$ ssi $\forall w'(w' \in \mathcal{B}_i(w) \Rightarrow w' \Vdash A)$. « L'agent i croit A dans le monde w si et seulement si tous les mondes dans lesquels il croit être satisfont A. » On impose que $\mathcal{B}_i$ soit une relation sérielle, transitive, et euclidienne.

Axiomatique et complétude. En plus de la règle de nécessitation pour la croyance, nous avons les schémas d'axiomes suivants :

$$Bel_i(A \rightarrow B) \rightarrow (Bel_i A \rightarrow Bel_i B) \quad (\mathrm{K}_{Bel_i})$$
$$Bel_i A \rightarrow \neg Bel_i \neg A \quad (\mathrm{D}_{Bel_i})$$
$$Bel_i A \rightarrow Bel_i Bel_i A \quad (4_{Bel_i})$$
$$\neg Bel_i A \rightarrow Bel_i \neg Bel_i A \quad (5_{Bel_i})$$

On rajoute B à nos modèle qui associe à chaque agent une relation d'accessibilité $\mathcal{B}_i$. Là encore, on peut montrer aisément que la complétude est préservée.

Formalisation. Notre approche consiste à simplement préfixer toute modalité par l'opérateur de croyance Bel_i. Du fait des propriétés de la logique KD45, nous avons que $Bel_i Good_i A \leftrightarrow Bel_i \Box_{\preccurlyeq_i} \neg A \wedge Bel_i \Diamond_{\succcurlyeq_i} A$. En d'autres termes, l'agent i croit qu'il serait bon que A soit vrai si et seulement si il croit que A est faux dans le monde actuel et dans tous les mondes pires, et qu'il croit qu'il existe un monde meilleur dans lequel A est vrai.

4.4 Conclusions et perspectives

Globalement, le formalisme développé dans cet exposé permet une meilleure interprétation des actes de langage indirects en proposant d'une part une définition des notions de bon et de mauvais, et d'autre part une articulation logique entre ces notions et les notions déjà formalisées que sont l'action et la croyance.

Après avoir montré qu'une logique normale ne pouvait pas capturer les notions de bon et de mauvais, nous avons pu développer une logique modale non normale, basée sur une sémantique rigoureuse qui traduit fidèlement l'intuition qu'une chose est bonne si elle va dans le sens d'une amélioration.

La notion de bon ainsi caractérisée a été adaptée de façon non triviale aux actions puis aux croyances. Une étude de la logique dynamique a permis de mettre en évidence quelques subtilités lors de sa mise en rapport à l'intuition de ce qu'est une bonne action. De même, la logique épistémique apporte une solution satisfaisante à la représentation subjective de la notion de bon.

Du point de vue des agents (qui sont susceptibles de changer de croyances) nous avons montré que la révision et la mise à jour de ce qui est bon et mauvais sont traitées *via* celles des croyances (pour lesquelles nous disposons déjà de solutions formelles connues).

Ce travail apporte donc aux travaux en cours un nouvel outil logique pour doter un agent rationnel autonome des capacités d'interprétation du langage naturel.

Bien que non présentés ici, quelques résultats nous permettent d'aller plus loin, notamment en donnant une première caractéristique de la propriété selon laquelle *il serait meilleur d'être*

dans l'état E que dans l'état F. À terme, nous espérons que ces résultats seront les prémisses d'un système de prise de décision.

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), June 1985.

[2] Patrick Blackburn, Maarteen de Rijke, and Yde Venema. *Modal Logic*. Number 53 in Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, Cambridge, 2001.

[3] Maud Champagne, Andreas Herzig, Dominique Longin, Jean-Luc Nespoulous, and Jacques Virbel. Formalisation pluridisciplinaire de l'inférence d'actes de langage non littéraux. *Information, Interaction, Intelligence*, Hors série :197–225, 2002.

[4] B. F. Chellas. *Modal Logic : an introduction*. Cambridge University Press, 1980.

[5] Peter Gärdenfors. *Knowledge in Flux : Modeling the Dynamics of Epistemic States*. MIT Press, 1988.

[6] Joseph Y. Halpern and Yoram Moses. A guide to the modal logics of knowledge and belief : preliminary draft. In *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI'85)*. Morgan Kaufmann Publishers, 1985.

[7] D. Harel. Dynamic logic. *Handbook of Philosophical Logic*, 2 :497–604, 1984.

[8] Andreas Herzig and Dominique Longin. Sensing and revision in a modal logic of belief and action. In F. van Harmelen, editor, *Proc. of 15th European Conf. on Artificial Intelligence (ECAI 2002)*, pages 307–311, Amsterdam, 2002. IOS Press.

[9] Andreas Herzig and Dominique Longin. C&L intention revisited. In Didier Dubois, Chris Welty, and Mary-Anne Williams, editors, *Proc. 9th Int. Conf. on Principles on Principles of Knowledge Representation and Reasoning (KR2004) , Whistler, Canada*, pages 527–535. AAAI Press, 2-5 juin 2004.

[10] J. Hintikka. *Knowledge and Belief : An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca, 1962.

[11] Hirofumi Katsuno and Alberto O. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence Journal*, 52, 1991.

[12] Robert C. Moore. A formal theory of knowledge and action. In J.R. Hobbs and R.C. Moore, editors, *Formal Theories of the Commonsense World*, pages 319–358. Ablex, Norwood, NJ, 1985.

[13] Ray Reiter. The frame problem in the Situation Calculus : A simple solution (sometimes) and a completeness result for goal regression. In Vladimir Lifschitz, editor, *Artificial Intelligence and Mathematical Theory of Computation : Papers in Honor of John McCarthy*, pages 359–380. Academic Press, San Diego, CA, 1991.

[14] H. Sahlqvist. Completeness and correspondence in the first and second order semantics for modal logics. In S. Kanger, editor, *Proc. 3rd Scandinavian Logic Symposium*, volume 82 of *Studies in Logic*, 1975.

[15] Richard Scherl and Hector J. Levesque. The frame problem and knowledge producing actions. In *Proc. Nat. Conf. on AI (AAAI'93)*, pages 689–695. AAAI Press, 1993.

[16] J.R. Searle. *Sens et Expression*. Expression and Meaning. Cambridge University Press, 1979.