MODELES D'INFORMATION ET THEORIE DE LA CONNAISSANCE

Philippe Mongin
Laboratoire d'économétrie, Ecole Polytechnique,
1 rue Descartes, F-75005 Paris
mongin@poly.polytechnique.fr

Février 2002

1. Le modèle des partitions d'information

La théorie de la décision, la théorie des jeux et la micro-économie de l'incertain, lorsqu'elle s'appuie sur ces deux disciplines, représentent le plus souvent l'information des agents à l'aide d'un *modèle de partitions*:

$$\langle \Omega, \Pi_i, i \in I \rangle$$

Comme dans la théorie des probabilités, qui a inspiré ce formalisme, *l'ensemble des états du monde* Ω fixe l'univers du discours. Les propriétés que le modèle permet de discuter s'identifient par hypothèse aux sous-ensembles (ou *événements)* $E \subseteq \Omega$ vérifiant ces propriétés. Par exemple, on traite "il fait beau aujourd'hui" ou "l'agent i sait qu'il fait beau aujourd'hui" comme des événements particuliers de Ω. Quant aux Π_i, ce sont des partitions de l'ensemble Ω; il y en a autant que d'agents $i \in I$. Elles représentent le pouvoir que chacun a de distinguer entre eux les états. L'appellation de *partitions d'information* s'explique ainsi: l'individu i perçoit l'information qui survient comme grossie et déformée par son Π_i. Plus précisément, l'information se présente objectivement, c'est-à-dire aux yeux du modélisateur, comme la réalisation d'un $\omega \in \Omega$, et elle

se présente aux yeux des agents comme la réalisation de l'élément de partition, ou *cellule*, $\Pi_i(\omega)$, auquel ω appartient.

Le modèle des partitions prépare l'introduction des concepts les plus caractéristiques de la théorie des jeux ou de la décision: les croyances des agents, qui sont généralement, mais non pas toujours, représentées par des mesures de probabilité μ_i ; les actions a_i dont les agents ont à décider; les fonctions d'évaluation grâce auxquelles ils comparent les actions. Lorsqu'elles entrent à leur tour dans l'analyse, ces notions doivent être cohérentes avec le modèle de partition choisi $\langle \Omega, \Pi_i, i \in I \rangle$. Techniquement parlant, on demandera que μ_i permette d'évaluer les ensembles engendrés par la partition Π_i et que les actions a_i, vues comme fonctions des ω, soient mesurables par rapport à Π_i. Suivant la modélisation probabiliste ordinaire, le flux de l'information se traduit objectivement par la réalisation d'états successifs ω et subjectivement par une révision bayésienne des probabilités $\mu_i(. \mid \Pi_i(\omega))$. En introduisant un modèle de partitions, le théoricien s'astreint donc à respecter des contraintes bien déterminées pour le reste de sa formalisation.

L'analyse d'inspiration logicienne qui s'est développée sous le nom de "théorie de la connaissance" permet d'approfondir le modèle de deux manières différentes: sous un certain angle, elle le fonde plus rigoureusement; sous un autre angle, elle en fait comprendre les limites et elle en propose des variantes peut-être moins contestables.

L'ASPECT "FONDATIONNEL". Si l'on veut faire une analyse ensembliste des propriétés, il faut que les états contiennent la description de *tous les faits du monde pertinents pour la discussion de ces propriétés.* Il n'y aurait pas de sens à identifier la propriété "il fait beau aujourd'hui" avec un sous-ensemble d'états $E \subseteq \Omega$ si sa présence ou son absence ne fait pas partie de la description de chaque état. Mais que veut-on dire lorsqu'on parle d'une description de "tous les faits du monde pertinents"? Une description du monde complète à un degré satisfaisant doit comporter plus que les faits objectifs, même si l'on convient de rattacher à ceux-ci les fonctions d'évaluation des agents. Elle doit inclure, au moins implicitement, *les faits de croyances* que le modèle permet de formuler. Par là on veut dire non seulement les croyances sur les faits objectifs, mais les croyances sur les croyances, à un niveau d'itération quelconque. De plus, on a débattu de l'appartenance des *actions* à une description complète du monde.

Même si la théorie des jeux pratique ce type de questionnement plus souvent que la théorie de la décision, il s'impose déjà lorsqu'on envisage un agent isolé.

En même temps qu'un problème général de délimitation, la notion d'état du monde semble entraîner un risque de *circularité*. En effet, les croyances portent normalement *sur* les états; mais on vient de voir qu'elles étaient aussi dissimulées *dans* les états. Comment envisager la relation de ces deux manières d'envisager la croyance, l'explicite et l'implicite?

La théorie de la connaissance permet d'approfondir la question générale à quoi se ramènent celles que nous venons de soulever: qu'est-ce qu'un état du monde dans le modèle des partitions et, plus généralement, dans un modèle d'information?

L'ASPECT CRITIQUE. Pourquoi devrait-on supposer que l'information parvienne aux agents filtrée par une partition? N'y a-t-il pas, logiquement parlant, des modèles plus généraux pour représenter l'information et préparer l'introduction des croyances, probabilistes ou non? S'il s'avérait que ces modèles sont plus conformes à la nature du raisonnement naturel, celui des partitions ne se justifierait finalement que par sa commodité mathématique et une longue tradition d'emploi dans les sciences sociales formalisées.

De telles préoccupations semblent plus concrètes et plus faciles à prendre en compte que les recherches "fondationnelles". Il se trouve qu'elles y ramènent une fois qu'on les approfondit suffisamment. Pour l'exposé qui suit, nécessairement sommaire et simplificateur, nous avons retenu comme fil directeur la contribution critique de la théorie de la connaissance, mais nous n'avons pas renoncé entièrement à évoquer les bases logiques de la théorie des jeux et de la théorie de la décision.

2. Au-delà des partitions: les correspondances de possibilité

Si Π_i est une partition de Ω, nous pouvons la considérer comme une fonction qui associerait à chaque ω la cellule $\Pi_i(\omega)$ à laquelle il appartient. La première généralisation à laquelle on puisse penser consiste à remplacer les cellules par des sous-ensembles a priori quelconques. On définira donc une *correspondance de possibilité pour* i comme une application:

$$P_i : \Omega \to 2^{\Omega}$$

On suppose en général que $P_i(\omega) \neq \emptyset$, mais cette hypothèse n'est pas nécessaire à la définition. Au lieu de dire que l'individu est incapable de distinguer ω d'autres états, on dira désormais que, quand ω se produit, *il juge possible* - ou *il n'exclut pas* - que certains états ω' se produisent. On ne présuppose pas de relation particulière entre ω et ω', par exemple la symétrie qui est évidemment sous-entendue lorsqu'on dit: "l'agent ne distingue pas entre ω et ω'". L'idée des correspondances de possibilité préserve la notion d'information tout en autorisant une souplesse qui manque au modèle des partitions.

On retrouve le cas particulier des partitions Π_i à partir des trois conditions :
(P1) $\omega \in P_i(\omega)$
(P2) Si $\omega' \in P_i(\omega)$, alors $P_i(\omega') \subseteq P_i(\omega)$
(P3) Si $\omega' \in P_i(\omega)$, alors $P_i(\omega) \subseteq P_i(\omega')$.
Elles sont évidemment nécessaires. Pour vérifier la suffisance, il est commode d'introduire la relation binaire:

$$(*)\ \omega R_i \omega' \text{ ssi } \omega' \in P_i(\omega)$$

On voit aussitôt que R_i est réflexive, transitive et symétrique, donc qu'elle est une relation d'équivalence, de sorte que P_i définit une partition.

EXEMPLES. "A noir, E blanc, I rouge, U vert, O bleu: voyelles,
Je dirai quelque jour vos naissances latentes".
Au contraire de Rimbaud, nous supposerons que les cinq voyelles peuvent être associées à n'importe laquelle des cinq couleurs:

$$\Omega = \left\{ \begin{array}{c} \text{A noir, A blanc, A rouge, A vert, A bleu, E noir, E blanc,} \\ \text{...,I noir,...,O noir,...,U noir,...} \end{array} \right\}$$

- Un individu reconnaît seulement les voyelles, un autre seulement les couleurs: c'est une application particulière du modèle des partitions.

- L'individu i reconnaît toujours les voyelles et il reconnaît les couleurs seulement quand celles-ci correspondent aux couleurs suivant la correspondance voulue par Rimbaud. Ainsi,

P_i(A noir) = {A noir}, P_i(A blanc)= {A noir, A blanc, A rouge, A vert, A bleu} , etc.

Ce cas viole (P3), mais respecte (P1) et (P2).

- L'individu i associe toujours deux couples (voyelle, couleur) quand ils correspondent à la même voyelle et se succèdent dans la liste qui a servi à décrire Ω. Ainsi,

P_i(A noir)= {A noir, A blanc} , P_i(A blanc)= {A blanc, A rouge}, etc.
Dans ce cas, seul (P1) est vérifié.

- Supposons en outre que l'individu i lise E à la place de A, I à la place de E,..., A à la place de U. On a donc:

P_i(A noir)= {E noir, E blanc} ,
P_i(A blanc)= {E blanc, E rouge}, etc.
Aucune des trois conditions n'est plus respectée.

Les trois derniers exemples ne sont pas seulement le produit de la fantaisie poétique. Ils illustrent trois contextes banals pour lesquels le modèle des partitions devient inadéquat:

- L'individu *n'est pas au fait* de certains traits de la réalité, par quoi l'on veut dire qu'il n'est pas en mesure de se prononcer sur eux et ne sait même pas qu'il ne le peut pas. (Les anglo-saxons emploient alors un mot commode: UNAWARENESS, qui n'a pas d'équivalent direct en français.) Ainsi, le trait "couleur" n'est pas présent à l'esprit de i quand l'état réel est "A blanc". Le même individu est en revanche bien au fait de la couleur dans d'autres états comme "A noir". Inacceptable pour le modèle des partitions, cette dissymétrie n'a pourtant rien de surprenant. Il arrive banalement que nous prenions conscience d'un aspect de la situation à l'instant seulement où cet aspect se modifie - comme nous pouvons découvrir la présence d'un bruit de fond quand celui-ci s'interrompt.

- L'individu est *imprécis*. Les données que l'individu traite comme identiques dans l'état ω se confondent en partie seulement avec les données qu'il traite comme identiques dans l'état ω', et vice-versa. La perception ordinaire des grandeurs physiques fournit des exemples analogues: un habitué des bains de mer est susceptible de réagir semblablement à des températures de 16°

et de 18°, à des températures de 18° et 20°, mais non pas, sans doute, à des températures de 16° et 20°. La perception découpe dans le continuum physique des intervalles *qui se chevauchent*. La même observation élémentaire a conduit les théoriciens de la décision à nier que la relation de préférence engendre une relation d'indifférence transitive.

- L'individu est à la fois *imprécis* et *inexact*. Non seulement les données qu'il traite comme identiques se chevauchent entre deux états, mais elles ne comprennent même pas toujours la valeur exacte. Voici un autre exemple: j'apprécie le prix des objets avec une marge de 10% et en outre, je prends les euros pour des francs.

Les correspondances de possibilité permettent de parler des croyances et des connaissances à l'aide, uniquement, d'ensembles et d'opérations sur les ensembles, donc en particulier sans faire référence aux probabilités. On dira que, dans un état ω, l'individu *i croit à l'événement* $E \subseteq \Omega$ si

$$P_i(\omega) \subseteq E$$

et l'on dira qu'il *connaît l'événement* E si, en outre,

$$\omega \in \Omega.$$

Suivant cette dernière définition, la connaissance n'est rien d'autre qu'une croyance qui se trouve être vraie dans l'état du monde considéré. C'est une manière sommaire et discutable de distinguer les deux concepts, mais elle est bien établie dans une partie de la littérature et nous nous y tiendrons dans ce qui suit.[1] Nous ferons bientôt le lien entre la notion ensembliste de croyance et les notions probabilistes plus courantes. Pour l'instant, il suffit de remarquer que la notion ensembliste a un répondant intuitif. L'individu ne peut *croire* que ce dont il est susceptible d'être *informé*. En ω, il perçoit $P_i(\omega)$ ou il ne perçoit rien du tout; pour qu'il croie à E, il faut donc l'inclusion précédente. La propriété réciproque est sans doute plus discutable. Mais on peut dire en première approximation que si l'individu est informé de $P_i(\omega)$, il tient cet événement pour réalisé, donc il le croit. On peut étendre cette observation à tous les E incluant $P_i(\omega)$ au motif que l'individu ne fait pas la différence entre eux.

[1]Les épistémologues ont à juste titre développé des notions plus subtiles à partir de l'idée que la connaissance serait une croyance vraie *justifiée*. Cette idée, elle-même trop simple, fait l'objet d'un examen dans le recueil [2].

3. Les modèles de Kripke

En proposant de remplacer les partitions par des correspondances de possibilité, les économistes ont rejoint les logiciens sur un terrain que ceux-ci occupaient depuis longtemps. La relation R_i que nous avons définie par (*) s'appelle, en logique modale, *relation de Kripke*. Elle y apparaît comme un concept primitif de l'analyse. La correspondance de possibilité P_i s'obtient alors comme un concept dérivé:

$$(**)\ P_i(\omega) = \{\omega' \mid \omega R_i \omega'\}$$

Il n'y a pas d'avantage particulier à prendre une notion plutôt que l'autre: seule la tradition de la discipline fixe l'usage. On retrouve toutes les intuitions précédentes relatives à l'information lorsqu'on interprète $\omega R_i \omega'$ de la manière suivante: en ω, l'individu i regarde comme possible - il n'exclut pas - l'occurrence de ω'.

Les conditions (P1), (P2), (P3) se reformulent de manière évidente:
(P1') R_i est *réflexive.*
(P2') R_i est *transitive.*
(P3') R_i est *euclidienne*, c'est-à-dire telle que si $\omega R_i \omega'$ et $\omega R_i \omega''$, alors $\omega' R_i \omega''$ et $\omega'' R_i \omega'$.

On observe au passage que les propriétés constitutives du modèle des partitions sont redondantes: la première et la troisième impliquent la deuxième. La vérification se fait encore plus aisément à partir des relations de Kripke que des correspondances de possibilité: une relation réflexive et euclidienne est symétrique, une relation symétrique et euclidienne est transitive. Ainsi, (P1') et (P3') $\Longrightarrow$(P2'), et de même, (P1) et (P3) $\Longrightarrow$(P2).

Entre les mains des logiciens, les relations de Kripke ne jouent pas seulement le rôle d'un outil de modélisation intuitif, mais celui, rigoureusement formel, que l'on peut attendre d'une *sémantique:* elles permettent de dire si les formules d'un langage logique sont vraies ou fausses. C'est l'aspect que nous développerons maintenant.

Notre langage logique sera construit à partir des unités de base qui suivent:

- des variables atomiques $p_1, ..., p_k, ... \in \mathcal{P}$;
- les connecteurs propositionnels ordinaires $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\longleftrightarrow$ (qui se lisent respectivement "non", "et", "ou", "implique", "biimplique");

- des opérateurs épistémiques B_i, $i \in I$.

Des règles que nous ne détaillons pas indiquent ce que sont les *formules bien formées* à partir de ces constituants. Par exemple, $B_2 p_3 \rightarrow (p_2 \vee \neg B_1 B_3 p_5)$ est une formule bien formée, et $\neg p_1 B_1$ n'en est pas une. Soit $\mathcal{L}$ l'ensemble des formules bien formées ou, plus brièvement, des *formules*. C'est notre langage logique; il est très simple mais suffisamment expressif déjà pour qu'on puisse décrire la croyance et la connaissance avec son aide. Les variables atomiques serviront à représenter les faits objectifs du monde, et les formules moléculaires impliquant des opérateurs épistémiques, les faits de croyance de niveau quelconque. Pour certains développements supplémentaires, il faudrait représenter aussi les actions; on y parviendrait en enrichissant le langage $\mathcal{L}$.

Revenant à la sémantique, nous définirons un *modèle de Kripke* comme un uplet:

$$m = \langle \Omega, R_i, i \in I, v \rangle$$

avec Ω un ensemble, R_i des relations binaires sur Ω, et

$$v : \Omega \times \mathcal{P} \rightarrow \{0, 1\}$$

une *fonction de valuation*. L'application v est la seule notion nouvelle. Elle permet au modèle m d'attribuer une valeur de vérité aux variables atomiques p selon l'état ω qui se réalise.

Formellement parlant, on pose:

- $m, \omega \models p$ si et seulement si $v(\omega, p) = 1$.

(Le membre de gauche se lit comme "p est vrai dans l'état ω du modèle m", .ou encore "le modèle m et l'état ω valident p".) Cette clause règle le cas des formules élémentaires de $\mathcal{L}$.
Les clauses qui suivent portent sur les différents cas de formules moléculaires de $\mathcal{L}$:

- $m, \omega \models \neg\phi$ si et seulement si non $m, \omega \models p$;
- $m, \omega \models \phi_1 \wedge \phi_2$ si et seulement si $m, \omega \models \phi_1$ et $m, \omega \models \phi_2$;

...

- $m, \omega \models B_i\phi$ si et seulement si pour tout ω' tel que $\omega R_i \omega'$, $m, \omega \models \phi$.

Prises ensemble, ces clauses sémantiques énoncent une définition par récurrence de la *relation de validation sémantique* $m, \omega \models \varphi$ ("ϕ est vraie dans l'état ω du modèle m") pour toute formule concevable $\phi \in \mathcal{L}$. Implicitement, la récurrence se fait sur la complexité des formules en partant des atomiques; nous omettons les détails. Les premières clauses interprètent les connecteurs propositionnels à la manière évidente; seule la dernière est conceptuellement significative. Elle incorpore dans la sémantique la notion ensembliste de croyance qui vient d'être présentée à l'aide des correspondances de possibilité. En ω, l'agent croit à la proposition correspondant à la formule ϕ si cette formule est vraie dans tous les états ω' dont il n'exclut pas l'occurrence quand ω se produit.

Par construction même, les modèles de Kripke satisfont à certaines propriétés sémantiques interprétables au point de vue de la croyance: pour tout modèle de Kripke m et tout $\omega \in \Omega$,

$$m, \omega \models B_i\varphi \wedge B_i\psi \rightarrow B_i(\phi \wedge \psi)$$

$$m, \omega \models B_i\varphi \wedge B_i(\phi \rightarrow \psi) \rightarrow B_i\psi$$

La vérification est immédiate à partir de la clause B_i de la définition par récurrence. La propriété suivante n'est pas plus difficile à vérifier:

$$\begin{array}{rl} \text{si, pour tout } m \text{ et tout } \omega & \in \ \Omega,\ m, \omega \models \phi, \\ \text{alors pour tout } m \text{ et tout } \omega & \in \ \Omega,\ m, \omega \models B_i\phi \end{array}$$

La première propriété indique en substance que la croyance respecte la conjonction logique, et la seconde, qu'elle respecte l'implication logique (on dit alors qu'elle est *monotone*). Suivant la troisième propriété, si une formule est une vérité sémantique universelle, c'est aussi une vérité sémantique universelle que cette formule est objet de croyance. Ces exigences épistémiques ne peuvent pas être innocentes. Les spécialistes d'intelligence artificielle les regroupent volontiers sous le chef de l'*omniscience logique*, ce qui en dit déjà long. Nous y reviendrons.

A ce point, nous n'avons introduit que les constituants du langage formel, c'est-à-dire les formules de $\mathcal{L}$. Pour achever de présenter la syntaxe, il faut introduire un *système d'axiomes et de règles d'inférence*. Comme toujours la discussion sémantique éclaire les choix syntaxiques. En nous laissant donc guider par le paragraphe précédent, nous retiendrons comme axiomes:

- (C) $B_i\varphi \wedge B_i\psi \rightarrow B_i(\phi \wedge \psi)$ (axiome de *conjonctivité*)

et comme règles:

- (RM) De $\phi \rightarrow \psi$, on peut inférer $B_i\phi \rightarrow B_i\psi$ (règle de *monotonie*)
- (RN) De ϕ, on peut inférer $B_i\phi$ (règle de *nécessitation*[2]).

Il est par ailleurs entendu que le système contient:

- les axiomes et règles d'inférence du calcul propositionnel ordinaire.

Un système d'axiomes et de règles d'inférence étant fixé, il en découle une notion de *théorème* propre à ce système. En substance, un théorème est une formule que l'on peut inférer en appliquant les règles aux axiomes ou à des théorèmes déjà obtenus. Nous ne donnons pas la définition technique; elle s'énonce encore une fois comme une récurrence finie. La notation $S \vdash \phi$ signifie que ϕ est un théorème du système S.

Nous venons de présenter le système le plus célèbre de la logique modale, celui de Kripke ou système K. Il se trouve que K *axiomatise* la classe des modèles de Kripke, au sens techniquement exigeant que les logiciens donnent à ce mot: toute vérité sémantique générale est un théorème du système, et réciproquement. En notation logicienne: pour toute formule $\phi \in \mathcal{L}$,

$$[(\forall m)\ (\forall \omega \in \Omega)\ m,\omega \models \phi\ ] \Leftrightarrow\ K \vdash \phi$$

Suivant la terminologie reçue, l'implication de gauche à droite énonce l'adéquation (*soundness*), et l'implication de droite à gauche la complétude (*completeness*), du système K par rapport à la sémantique de Kripke. Démontrer l'adéquation est élémentaire, mais il n'en va de même de la complétude. Le lecteur trouvera les démonstrations et d'autres précisions techniques dans le manuel de Chellas ([3], ch.4-5). Lorsque par la suite nous parlerons d'axiomatique ou d'axiomatisation à propos d'autres systèmes et d'autres classes de modèles, nous sous-entendrons toujours que le logicien a démontré l'adéquation et la complétude de son système pour la classe de modèles choisie.

[2]Cette appellation s'est imposée à cause de l'interprétation de B_i comme opérateur de nécessité, qui a précédé les interprétations épistémiques en logique modale.

Quel est l'intérêt, demandera-t-on peut-être, d'introduire un langage logique à côté du langage ensembliste si, finalement, les deux doivent servir à exprimer les mêmes faits universels? Une première réponse découle de ce qui vient d'être dit sur les axiomatisations: la correspondance entre les deux langages, entre la syntaxe et la sémantique, ne va jamais de soi; elle demande à être établie. Une deuxième réponse, plus simple dans le principe, est que la syntaxe et la sémantique sont comme deux points de vue complémentaires sur un même objet. Il y a des propriétés que l'on se représente mieux en adoptant un point de vue plutôt que l'autre. En particulier, on mesure plus exactement ce qu'exigent les modèles d'information lorsqu'on parvient à leur donner une expression syntaxique.

Le modèle des correspondances de possibilité a son répondant axiomatique dans les systèmes K et éventuellement, KD, qui est K accru de l'axiome suivant:

$$\text{(D) } B_i\varphi \rightarrow \neg B_i \neg \phi.$$

Celui-ci demande en substance que les croyances de i soient cohérentes, et il correspond sémantiquement aux hypothèses:

(P0) $P_i(\omega) \neq \emptyset$ ou (P0'): R_i est *sérielle*, c.à.d. chaque ω a au moins un successeur par R_i.

Quant au modèle des partitions, il a son répondant axiomatique dans le système $KT45$, appelé aussi $S5$, qui est K accru des trois axiomes suivants:

(T) $B_i\varphi \rightarrow \phi$ correspondant à (P1)-(P1').

(4) $B_i\varphi \rightarrow B_iB_i\phi$ correspondant à (P2)-(P2').

(5)$\neg B_i\phi \rightarrow B_i\neg B_i\phi$ correspondant à (P3)-(P3').

Comme les conditions sémantiques correspondantes, l'axiome (4) est redondant, mais la décomposition procurée par $KT45$ est conceptuellement éclairante. L'axiome de *véridicité* (T) rend manifeste que le modèle des partitions permet de traiter de connaissance, au sens où l'on a distingué cette notion de celle, plus générale, de croyance. Les deux axiomes (4) et (5) montrent que, si l'on modélise l'information par des partitions, on soumet la connaissance à une double exigence d'*introspection positive* et d'*introspection négative*. La dernière exigence, en particulier, paraît exorbitante. Nous nous tournons maintenant vers ce point critique.

4. Objections contre l'introspection négative

Cette section reprend tout d'abord les travaux de Shin ([21]) et de Shin et Williamson ([22]), qui fondent la critique de l'introspection négative sur des considérations spécifiquement logiques. Nous exposerons ensuite une critique d'un genre plus intuitif.

En suivant plus particulièrement ([22]), on se donne un ensemble dénombrable de variables propositionnelles p, de sorte que l'ensemble des formules $\mathcal{L}$ est lui-même dénombrable, et l'on fixe un modèle de Kripke $m = \langle \Omega, R_i, i \in I, v \rangle$ vérifiant le système KD. Pour chaque $\omega \in \Omega$, on définit la *théorie de* i en ω comme:

$$\text{Déf. } \Delta_i(\omega) = \{\phi \mid m, \omega \models B_i\phi\}$$

Cet ensemble décrit à la manière syntaxique tout ce que l'individu i croit dans l'état ω. Parce que (D) est satisfait, il est *cohérent*, c'est-à-dire ne contient jamais à la fois ϕ et $\neg\phi$. On établit alors la propriété suivante.

$$\begin{array}{rcl} \text{Si } m, \omega & \models & B_i\phi \rightarrow \phi \text{ et } m, \omega \models \neg B_i\phi \rightarrow B_i\neg B_i\phi, \\ \text{alors} & : & (*)\ \Delta_i(\omega) = \{\phi \mid \neg B_i\phi \notin \Delta_i(\omega)\} \end{array}$$

La preuve est très simple; nous la donnons pour faire sentir le rôle des deux hypothèses:

- (De gauche à droite.) Si $\phi \in \Delta_i(\omega)$, alors $m, \omega \models B_i\phi$, non $m, \omega \models \neg B_i\phi$, et, puisque la première hypothèse implique: $m, \omega \models B_i\neg B_i\phi \rightarrow \neg B_i\phi$, on a: non $m, \omega \models B_iB_i\neg\phi$, d'où $\neg B_i\varphi \notin \Delta_i(\omega)$.

- (De droite à gauche.) Si $\neg B_i\varphi \notin \Delta_i(\omega)$, non $m, \omega \models B_i\neg B_i\phi$, et on utilise la seconde hypothèse, récrite comme: $m, \omega \models B_i\phi \vee B_i\neg B_i\phi$, pour conclure que $m, \omega \models B_i\phi$, d'où $\phi \in \Delta_i(\omega)$.

(Tel qu'il est présenté, le raisonnement a un caractère local: il concerne un état donné ω qui se trouve satisfaire aux deux hypothèses de véridicité et d'introspection négative. On aurait pu raisonner globalement et se donner un modèle pour $KT45$, (D) devenant alors redondant.)

La propriété qui vient d'être établie comporte une conséquence logique importante: un ensemble $\Delta_i(\omega)$ qui satisfait aux deux hypothèses est récursif. Nous

rappelons qu'un ensemble est *récursif* s'il vérifie simultanément les deux conditions: il existe une procédure qui donne, en un nombre fini d'étapes, une réponse affirmative à la question: "L'élément x appartient-il à l'ensemble?" lorsque x appartient à l'ensemble; et de même, il existe une procédure qui donne, en un nombre fini d'étapes, une réponse négative à cette question lorsque x n'appartient pas à l'ensemble. Un ensemble est *récursivement énumérable* si la première des deux conditions est remplie. En l'occurrence, la procédure "positive" pour $\Delta_i(\omega)$ consiste simplement à en énumérer les formules: si $\phi \in \Delta_i(\omega)$, ϕ apparaîtra dans la liste après un nombre fini d'étapes, puisque l'ensemble est dénombrable. L'existence d'une procédure "négative" pour $\Delta_i(\omega)$ résulte de l'équivalence (*) lorsque les deux hypothèses sont vérifiées. Le complémentaire de $\Delta_i(\omega)$ s'écrit alors

$$\{\phi \mid \neg B_i\phi \in \Delta_i(\omega)\}$$

et l'on repère $\phi \notin \Delta_i(\omega)$ lorsqu'on arrive à $\neg B_i\phi$ dans la liste des formules de $\Delta_i(\omega)$, donc encore fois après un nombre fini d'étapes. Le raisonnement a un caractère local: il concerne un état ω donné et la théorie de i qui lui correspond.

Une conséquence importante, que nous exprimerons informellement, découle de ce qui vient d'être dit: *on peut engendrer des impossibilités pourvu que l'on enrichisse convenablement la théorie de l'agent* $\Delta_i(\omega)$. Admettons que l'on ait enrichi le langage $\mathcal{L}$ de manière qu'il puisse exprimer une théorie de l'arithmétique et que cette théorie soit contenue dans $\Delta_i(\omega)$, sans que $\Delta_i(\omega)$ soit devenu pour autant incohérent. Supposons réalisées les deux hypothèses de véridicité et d'introspection négative, et admettons que la modification de $\Delta_i(\omega)$ ait préservé son caractère dénombrable, ou tout au moins récursivement énumérable, de manière qu'on puisse encore conclure, en reproduisant le raisonnement précédent, que $\Delta_i(\omega)$ est *récursif.* Alors, on pourra invoquer le théorème de Gödel-Rosser, qui dit en substance qu'une extension cohérente d'une théorie de l'arithmétique ne peut pas être récursive. On aura produit une contradiction.

Ce résultat négatif est dirigé contre l'une ou l'autre des deux conditions supposées vraies en ω pour produire la contradiction. Un individu i dont les croyances vérifient ces conditions ne peut croire aux lois de l'arithmétique. Le paradoxe dépend évidemment de la glose assez particulière - purement syntaxique - donnée à l'expression "croire aux lois de l'arithmétique". Il n'en est pas moins troublant.

Nous présenterons maintenant les arguments de type intuitif que suscite l'introspection négative. On peut finalement distinguer trois cas, dont deux correspondront à une violation très nette de cette condition.

- Un premier cas de violation survient quand la proposition examinée met en relation des concepts que l'individu ne possède pas. Ainsi, la proposition qu'exprime la phrase: "Peter Wiles a fourni la première démonstration correcte de la conjecture de Fermat" est pratiquement dénuée de signication aux yeux de l'homme de la rue, qui n'a ni le concept de conjecture de Fermat, ni, véritablement, ceux de démonstration et de Peter Wiles (même s'il peut en repérer quelques attributs, comme le fait qu'une démonstration est de l'ordre de la science et que Peter Wiles est un anglo-saxon). Non seulement l'homme de la rue ignore la proposition précédente, mais il ne sait pas qu'il l'ignore.

- Le second cas se distingue du précédent en ce que l'individu possède les concepts requis; toutefois, il les possède *seulement en puissance et non pas actuellement*, ce qui conduit de nouveau à une violation de l'introspection négative. Ce cas correspond à l'exemple suivant, popularisé par Geanakoplos ([7]). Dans l'état du monde où le chien n'a pas aboyé le soir du crime, Sherlock Holmes sait qu'il ne sait pas que le chien a aboyé, tandis que le naïf Watson n'a rien remarqué: non seulement il ne sait pas que le chien a aboyé, mais il ne sait pas qu'il ne sait pas que le chien a aboyé. Bien évidemment, Watson comprend les concepts impliqués dans cette proposition mais il n'en fait pas activement usage. Ce second cas de violation correspond à l'*unawareness* dans sa pureté.

- Le troisième cas couvre les circonstances, finalement limitées, où l'introspection négative s'applique. Sherlock Holmes y satisfait dans l'exemple précédent, contrairement à Watson. Voici une illustration plus réaliste. Début octobre 2001, les opérateurs boursiers ne savaient pas quelles devaient être les conséquences macro-économiques des attentats du 11 septembre, et l'on peut sans doute ajouter qu'ils savaient qu'ils ne les connaissaient pas. Ils possèdaient les concepts pertinents et les mobilisaient, tout en ne parvenant qu'à un savoir négatif de second ordre. Néanmoins, les raisonnements des opérateurs les rapprochent parfois plus souvent de Watson que de Holmes.

5. Peut-on formaliser l'*awareness*?

Le langage logique dont nous disposons ne permet pas de formaliser la distinction, pourtant substantielle, du premier et du second cas. Pour représenter convenablement l'appréhension des concepts par les individus, il faudrait une logique plus raffinée qui analyserait les constituants internes des propositions. Nous pouvons toutefois essayer de discuter la propriété consistant à être au fait d'une proposition; nous ignorerons les raisons diverses qui ont pu conduire à cet état épistémique. Les exemples de la section précédente suggèrent d'eux-mêmes la définition à retenir: on est au fait des propositions que l'on connaît et aussi de celles dont on sait qu'on ne les connaît pas. Nous ferons donc figurer dans le langage $\mathcal{L}$ des opérateurs supplémentaires A_i ("A" pour *awareness*) dérivés des opérateurs de croyance par la définition:

$$(\text{Def.}A_i)\ A_i\phi \leftrightarrow B_i\phi \vee B_i\neg B_i\phi$$

Modica et Rustichini ([18]) ont étudié l'opérateur A_i en proposant de le soumettre à une contrainte de *symétrie par rapport à la négation*:

$$(\text{A})\ A_i\phi \leftrightarrow A_i\neg\phi$$

Cette condition apparemment plausible crée une difficulté inattendue pour l'analyse de la croyance à la manière de Kripke, comme nous allons le montrer en suivant les deux auteurs.

On se donne un langage $\mathcal{L}$ comportant une seule proposition atomique p (représentant "le chien aboie") et un seul opérateur de croyance B_W (W pour Watson). On se donne aussi un modèle de Kripke à deux états, $m = \{\{\omega, \omega'\}, R_w, v\}$, à l'aide duquel on s'efforcera de décrire les deux états d'esprit que l'on peut prêter à l'individu Watson. Dans l'état ω, le chien aboie et Watson le sait. Dans l'état ω', le chien n'aboie pas, Watson ne le sait pas et il ne sait pas qu'il ne le sait pas. On voudra donc que les deux états vérifient les conditions:

$$\begin{aligned} m, \omega &\models p, B_W p \\ m, \omega' &\models \neg p, \neg B_W p, \neg B_W \neg B_W p \end{aligned}$$

Ces données reflètent simplement l'objectif de modélisation. Il reste à voir ce qu'elles impliquent pour la relation de Kripke R_W définie sur $\{\omega, \omega'\}$.

On admettra que R_W est *sérielle*, ce qui revient à prendre comme système logique KD au lieu de K et, plus informellement, à supposer la cohérence de l'individu Watson. Les données précédentes imposent alors que ω soit relié à lui-même et non pas à ω', tandis que ω' doit être relié *à la fois* à lui-même et à ω' (la sérialité excluant que ω' ne soit relié à aucun état). La relation R_W est maintenant définie. Elle viole la condition (P3'), ce qui illustre une nouvelle fois l'échec du modèle partitionnel dans les cas d'*unawareness*.

L'exemple comporte un enseignement plus complexe. La définition obtenue pour R_W livre automatiquement d'autres informations sémantiques:

$$m, \omega \models \neg B_W \neg p \text{ et } m, \omega' \models \neg B_W \neg p$$

d'où:

$$m, \omega' \models B_W \neg B_W \neg p$$

Or c'est là une violation de l'axiome (A). La définition de l'opérateur A_W et la dernière ligne impliquent en effet:

$$m, \omega' \models A_W \neg p$$

alors que l'on a obtenu précédemment:

$$m, \omega' \models \neg A_W p$$

La sérialité de R_W mise à part, nous n'avons utilisé dans les raisonnements précèdents que *la définition* de la relation $\models$. Le modèle m manifeste donc un conflit entre la propriété de symétrie de l'*awareness* et la sémantique de Kripke (et tout aussi bien, les correspondances de possibilité si l'on préfère ce langage.)

La difficulté précédente appelle différentes réponses possibles:

- On peut contester le choix du modèle m. Supposons que l'on introduise un autre état ω'' relié par R_W seulement à lui-même et doté des propriétés sémantiques suivantes:

  $$m, \omega'' \models \neg p, \neg B_W p, B_W \neg p$$

  Il suffit alors de poser $\omega' R_W \omega''$ pour éviter la violation de (A). La prise en compte de cet axiome impose peut-être seulement de choisir des modèles

de Kripke de cardinalité convenable. Pour voir ce qu'il en est, il faudrait disposer d'une caractérisation de (A) par une propriété relationnelle, sur le modèle de la caractérisation de (5) par la propriété euclidienne; or une telle caractérisation fait défaut.

- Au-delà de l'exemple particulier, on peut contester la définition de l'*awareness*, ainsi, éventuellement, que l'axiome (A). La définition (Def.A_i) est trop limitative parce qu'elle ne tient pas compte de la croyance d'ordre supérieur à 2. On voudrait par exemple que la formule suivante soit un théorème du système:

$$B_i \neg B_i \neg B_i \phi \rightarrow A_i \phi$$

ce qui n'est pas le cas dans $KDDef.A_i$. Quant à l'axiome (A), les deux raisons pour lesquelles on n'est pas au fait d'une proposition, l'incompétence conceptuelle et l'inadvertance, ne le justifient pas au degré: la symétrie de l'opérateur A_i par rapport à la négation semble convenir dans le premier cas mais non dans le second. Dekel, Lipman et Rustichini ([4]) ont repris l'analyse de l'*awareness* dans une définition différente, qui tient compte de ces objections intuitives. Ils retrouvent la conclusion que le modèle de Kripke est inadéquat.[3]

- Il reste donc à contester la sémantique de Kripke elle-même. D'autres arguments plus directs, qui sont liés à l'omniscience logique, nous incitent à aller justement dans ce sens. Nous explorerons cette voie dans la dernière section de l'article.

6. La connaissance commune dans le modèle partitionnel

On attribue l'idée de connaissance commune à Lewis dans *Convention* ([13]), où il tente d'expliquer la conformité à une convention par un raisonnement de type strictement individualiste et qui exclut toute possibilité de se lier par un accord préalable. Comme Lewis le montre, il faut alors supposer chez chaque acteur non seulement la connaissance d'un intérêt qu'a l'autre à agir d'une certaine manière

[3]La définition de l'*awareness* est loin de faire l'objet d'un accord unanime. Fagin et Halpern ([5]) en propose une autre encore dans le cadre du système K.

(par exemple, "rouler à droite"), mais aussi la connaissance de la connaissance qu'a l'autre d'un intérêt chez lui-même à agir d'une certaine manière, et ainsi de suite. Cette régression infinie de la connaissance conditionne la coordination efficace d'acteurs qui, par hypothèse, ne communiquent pas entre eux; toute interruption de la chaîne peut constituer une raison pour l'un d'eux de s'écarter de l'action considérée.

En faisant abstraction du problème particulier de coordination qui préoccupait Lewis, on peut décrire ainsi la connaissance commune d'un événement F: l'individu 1 sait que F; l'individu 2 sait que F; 1 sait que 2 sait que F; 2 sait que 1 sait que 2 sait F; et ainsi de suite. En ajoutant un élément d'introspection positive à cette idée première, on dira que F est de *connaissance commune* si tous les individus savent que F, tous les individus savent que tous les individus savent F, et ainsi de suite. Cette notion légèrement renforcée par rapport à la précédente est aussi plus facile à manier, et c'est elle que nous emploierons exclusivement. Par ailleurs, il n'y a pas la moindre difficulté à étendre les définitions au cas d'une population finie quelconque d'individus; c'est pour simplifier les notations que nous en tenons au cas de deux individus.

On doit à Aumann, dans l'article célèbre "Agreeing to Disagree" ([1]), d'avoir fait passer les notions informelles qui précèdent au stade d'une mathématisation opératoire. Il suppose que l'information individuelle se conforme au modèle des partitions:

$$\langle \Omega, \Pi_1, \Pi_2 \rangle$$

et il propose la définition suivante: un événement $E \subseteq \Omega$ est *connaissance commune en* $\omega \in \Omega$ si $\Pi_C(\omega) \subseteq E$, avec $\Pi_C =$ le plus fin grossissement commun des partitions individuelles Π_1 et Π_2. On appellera Π_C la *partition de connaissance commune.*

On dit qu'une partition Π' est un grossissement d'une autre partition Π si chaque cellule de Π est incluse dans une cellule de Π'; il est équivalent de dire que Π est un raffinement de Π'. Ces deux notions expriment dans le langage des partitions le fait que l'individu dispose d'un pouvoir de distinction diminué ou, au contraire, accru. L'existence d'un grossissement *mimimal*, c'est-à-dire d'une partition de connaissance commune au sens voulu par Aumann, est automatiquement assurée lorsque l'espace d'états Ω est fini. Soit par exemple:

$$\Omega = \{\omega_1, ..., \omega_{10}\}$$

$$\Pi_1 = \{\{\omega_1, \omega_2\}, \{\omega_3, \omega_4, \omega_5\}, \{\omega_6, \omega_7\}, \{\omega_8, \omega_9\}, \{\omega_{10}\}\}$$
$$\Pi_2 = \{\{\omega_1\}, \{\omega_2, \omega_3, \omega_4\}, \{\omega_5\}, \{\omega_6\}, \{\omega_7, \omega_8, \omega_9\}, \{\omega_{10}\}\}$$

Alors:

$$\Pi_C = \{\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}, \{\omega_6, \omega_7, \omega_8, \omega_9\}, \{\omega_{10}\}\}$$

et l'on vérifie par exemple que $E = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$ est connaissance commune en n'importe lequel de ses points, tandis que $E \setminus \{\omega_5\}$ n'est connaissance commune en aucun point de Ω.

Aumann justifie sa définition de la connaissance commune en signalant qu'elle est équivalente à une autre, qui repose sur une construction itérative plus facile à interpréter: F est connaissance commune en $\omega \in \Omega$ si

- $\Pi_1(\omega) \subseteq F$, $\Pi_2(\omega) \subseteq F$ (1 sait F, 2 sait F),
- $\forall \omega' \in \Pi_1(\omega)$, $\Pi_2(\omega') \subseteq F$ (1 sait que 2 sait F)
- $\forall \omega' \in \Pi_2(\omega)$, $\Pi_1(\omega') \subseteq F$ (2 sait que 1 sait F),
- $\forall \omega' \in \Pi_1(\omega)$, $\forall \omega'' \in \Pi_2(\omega'')$, $\Pi_1(\omega'') \subseteq F$ (1 sait que 2 sait que 1 sait F), et ainsi de suite.

A cause du modèle partitionnel retenu, la définition inclut automatiquement l'aspect d'introspection positive évoqué plus haut; ainsi, la première condition $\Pi_1(\omega) \subseteq F$ (1 sait F) équivaut à:

$$\forall \omega' \in \Pi_1(\omega),\ \Pi_1(\omega') \subseteq F \text{ (1 sait que 1 sait } F)$$

C'est grâce à la notion partitionnelle de connaissance commune qu'Aumann démontre le théorème fameux: "il est impossible d'être d'accord sur un désaccord". Mathématiquement: *si 1 et 2 ont des probabilités a priori identiques et que leurs probabilités a posteriori en un état sont de connaissance commune en cet état, alors ces probabilités a posteriori sont identiques.*

Le théorème a des implications déconcertantes pour l'analyse théorique des paris et de l'échange entre deux individus rationnels. Prenons, par exemple, deux individus susceptibles de parier sur le résultat de l'élection présidentielle de 2002. Nous ferons abstraction de l'attitude par rapport au risque et de l'utilité de la

monnaie, de sorte que la valeur du pari se ramène pour chaque individu à une espérance de gain. Nous admettrons en outre qu'ils forment au départ les mêmes probabilités a priori, tandis que leurs informations privées sont susceptibles de différer. On peut imaginer que le premier ait ses amis au Parti Socialiste, et le second les ait au R.P.R., ce qui vaut à chacun de bénéficier d'informations dont l'autre est privé. Le pari se produira-t-il entre les deux individus? *D'après le théorème qui précède, c'est impossible.* Si le pari se produit, les deux espérances de gain diffèrent, ce qui suppose que les probabilités a posteriori - c'est-à-dire les probabilités a priori communes révisées pour tenir compte de chaque information privée - diffèrent d'un individu à l'autre. Or le théorème dit que les a posteriori ne peuvent différer si elles sont connaissance commune, et l'on peut admettre que si les individus parient, leurs a posterori deviennent implicitement connaissance commune. D'où une contradiction.

Voici comment on peut formaliser le raisonnement précédent. Nous représenterons le pari par une variable aléatoire $f : \Omega \to R$. Pour chaque réalisation $\omega \in \Omega$, les deux individus réactualisent leur a priori commune p au vu des informations privées $\Pi_1(\omega)$ et $\Pi_2(\omega)$, respectivement. La transaction ne se produira que s'ils parviennent à des espérances différentes conditionnellement à ces informations. Formons donc:

$$F = \{\omega \in \Omega \mid E(f \mid \Pi_1(\omega)) > E(f \mid \Pi_2(\omega))\},$$

et supposons que F soit connaissance commune en un $\omega \in \Omega$. Alors, la définition d'Aumann permet d'écrire:

$$\Pi_C(\omega) = \sum_j \Pi_1^j = \sum_k \Pi_2^k \subseteq F$$

où $\sum_j \Pi_1^j$ et $\sum_k \Pi_2^k$ représentent deux unions disjointes de cellules de partition appartenant à 1 et 2 respectivement. Par construction de F, on a:

$$E(f \mid \Pi_1^j) > E(f \mid \Pi_2^k) \text{ si } \Pi_2^k \cap \Pi_1^j \neq \emptyset$$

Donc,

$$E(f \mid \Pi_1^j) > \sum_{\{\Pi_2^k \cap \Pi_1^j \neq \emptyset\}} \frac{P(\Pi_2^k \cap \Pi_1^j)}{P(\Pi_1^j)} E(f \mid \Pi_2^k)$$

On ne change pas la valeur de la somme en la faisant porter sur tous les Π_2^k. De la sorte:

$$P(\Pi_1^j)E(f \mid \Pi_1^j) > \sum_k P(\Pi_2^k \cap \Pi_1^j)E(f \mid \Pi_2^k)$$

En sommant ces inégalités sur j, on obtient à gauche:

$$\sum_j P(\Pi_1^j) E(f \mid \Pi_1^j) = E(f \mid \Pi_C(\omega))$$

et à droite, après avoir permuté les sommes:

$$\sum_k \sum_j P(\Pi_2^k \cap \Pi_1^j) E(f \mid \Pi_2^k) = \sum_k P(\Pi_2^k) E(f \mid \Pi_2^k) = E(f \mid \Pi_C(\omega)),$$

ce qui constitue la contradiction recherchée.

La formalisation qui précède portait sur une *application* du théorème d'Aumann au cas des paris. Mais elle implique une démonstration du théorème lui-même: il suffit de prendre la variable aléatoire f égale à la fonction caractéristique d'un ensemble donné. On s'est servi d'une propriété de la connaissance commune qui résulte directement de la définition: la cellule $\Pi_C(\omega)$ se *partitionne* relativement aux cellules individuelles ($\Pi_C(\omega) = \sum_j \Pi_1^j = \sum_k \Pi_2^k$). Comme l'a montré Samet ([19]), on peut, sous certaines conditions limitatives, étendre le théorème aux modèles d'information satisfaisant (P1) et (P2), mais pas (P3); la démonstration donnée ici ne peut évidemment plus convenir. Quant aux conditions (P1) et (P2), elles apparaissent inéluctables.[4] En particulier, on ne peut pas transposer le théorème à la croyance commune: implicitement, chaque individu déduit une information *sur le monde* à partir de l'information qu'il obtient *sur l'information de l'autre*, et cette déduction n'est possible que dans un contexte de croyance vraie et sue comme telle. L'hypothèse des probabilités a priori communes est aussi nécessaire. En effet, le théorème d'Aumann admet une réciproque, à peine moins frappante que l'énoncé initial: l'impossibilité d'être en accord sur un désaccord *implique* l'hypothèse d'a priori communes. Parmi les nombreuses démonstrations de cette propriété, la plus simple est celle de Samet ([20]).

Le raisonnement sur les paris s'étend à des situations plus complexes où interviennent des marchés et des prix. Tel est l'objet des nombreux théorèmes de *non-transaction* démontrés à partir de celui d'Aumann (par exemple, chez Milgrom et Stokey[14]). Le raisonnement consiste encore une fois à produire une absurdité en supposant que l'accord se produit, ce qui revient à "se placer à l'équilibre" et à en tirer les conséquences.

[4] La généralisation que Samet ([19]) mentionne à propos de (P1) n'est pas significative.

7. La croyance commune plus généralement

Le modèle des correspondances de possibilité permet de définir une notion de *croyance commune,* et non de connaissance commune, puisque la condition de véridicité n'est pas nécessairement remplie. On dira donc que F est croyance commune en $\omega \in \Omega$ si

- $P_1(\omega) \subseteq F$, $P_2(\omega) \subseteq F$ (1 croit F, 2 croit F),
- $\forall \omega' \in P_1(\omega), P_1(\omega') \subseteq F$ et $\forall \omega' \in P_2(\omega), P_2(\omega') \subseteq F$ (chacun croit qu'il croit F)
- $\forall \omega' \in P_1(\omega), P_2(\omega') \subseteq F$ et $\forall \omega' \in P_2(\omega), P_1(\omega') \subseteq F$ (chacun croit que l'autre croit F)
- $\forall \omega' \in P_1(\omega)$, $\forall \omega'' \in P_2(\omega'')$, $P_1(\omega'') \subseteq F$ (1 croit que 2 croit que 1 croit F), et ainsi de suite.

Dans le modèle de Kripke, équivalent à celui-ci, la définition se formule plus brièvement: F est croyance commune en $\omega \in \Omega$ si

$$\forall \omega' \in F,\ \omega R_C \omega'$$

avec R_C = la *relation de Kripke de la croyance commune* , qui est définie comme la clôture transitive de R_1 et R_2 (autrement dit, la plus petite relation transitive contenant $R_1 \cup R_2$). Le parallèle avec la section précédente est manifeste: la définition itérative de la croyance commune et la définition statique par R_C généralisent les définitions correspondantes chez Aumann. S'agissant de la définition par R_C, il est facile de voir que la clôture transitive de deux relations d'équivalence définit une nouvelle relation d'équivalence associée à la partition Π_C.

Au point de vue logique, la question se pose d'axiomatiser les *modèles de Kripke pour la connaissance commune*:

$$m = \langle \Omega, R_i, i \in I, R_C, v \rangle$$

avec la clause sémantique suivante en sus des clauses déjà indiquées (section 3):

- $m, \omega \models C\phi$ si et seulement si pour tout ω' tel que $\omega R_C \omega'$, $m, \omega \models \phi$.

Nous introduirons dans la syntaxe un opérateur auxiliaire E pour la *croyance partagée*:

$$(\text{Def.}E)\ E\phi \longleftrightarrow \wedge_{i \in I} B_i \phi$$

Si C est l'opérateur de croyance commune et que l'on note $E^k = E...E$ (k *fois*), on voudrait pouvoir écrire:

$$(\#)\ C\phi \longleftrightarrow E\phi \wedge EE\phi \wedge ... \wedge E^k\phi \wedge ...$$

Mais toute la difficulté vient précisément de ce qu'on ne peut pas écrire cette formule! Les langages formels employés ici n'autorisent que des formules finies. On pourrait certes envisager de changer de cadre. Mais on perdrait le bénéfice de la théorie existante des relations de Kripke, et il se trouve aussi que les logiques infinitaires - celles dont le langage autorise des formules infinies dénombrables - ne sont pas faciles à utiliser techniquement.[5]

Pour résoudre la difficulté, on sélectionnera des propriétés de la croyance commune *qui ont un caractère fini*, et on les transcrira syntaxiquement. Si les propriétés ainsi récrites fournissent une axiomatisation des modèles de Kripke pour la croyance commune, on sera pleinement rassuré sur le fait que l'on a reproduit par des moyens finis l'effet de la pseudo-formule (#). Car la sémantique de la croyance commune fournie par R_C correspond elle-même à l'idée d'itération infinie. Démontrer que la syntaxe choisie correspond bien à la sémantique, c'est se convaincre que la syntaxe rend l'effet de la formule que l'on ne peut pas écrire.

Voici quatre propriétés qui se présentent assez naturellement:

1. *Le point fixe*: la croyance commune d'un événement implique la croyance partagée de la croyance commune de cet événement. On obtient la même notion si l'on poursuit l'itération de la croyance partagée au-delà du stade atteint pour la croyance commune. Syntaxiquement, on voudra que $C\phi \rightarrow EC\phi$ puisse être prouvé dans la logique.

[5] Néanmoins, Kaneko ([12]) choisit la voie des logiques infinitaires pour définir la croyance commune et analyser la structure de certains jeux. Heifetz ([10]) compare en détail les formulations finitaires et infinitaires de la croyance commune.

2. *Une propriété itérative limitée*: la croyance commune implique la croyance partagée de niveau quelconque. Syntaxiquement, il faut que $C\phi \rightarrow E^k\phi$ puisse être prouvé pour tout entier k. Si l'on réexamine la pseudo-formule (#), on voit que l'implication de gauche à droite peut se comprendre de cette façon: elle est informellement équivalente à une suite infinie d'implications. On peut rassembler l'exigence précédente et celle qui vient d'être énoncée dans une formule de point fixe améliorée:

$$\text{(PF) } C\phi \rightarrow E(\phi \wedge C\phi)$$

Compte tenu de la monotonie de l'opérateur E, que celui-ci hérite de la monotonie des B_i (règle (RM) du système K), (PF) permet de prouver à la fois $C\phi \rightarrow EC\phi$ et $C\phi \rightarrow E^k\phi$; la vérification est facile.

3. *La propriété d'induction*: certains événements, qu'on peut appeler *intrinsèquement publics*, ne peuvent pas survenir sans être connus de tous: par exemple, le niveau atteint par le prix dans une enchère à la criée. On peut penser que les événements intrinsèquement publics sont objet de croyance (et même de connaissance) commune dès qu'ils surviennent; ainsi, le prix devient connaissance commune entre les enchérisseurs dès qu'il est crié. Syntaxiquement, on fera en sorte que la règle

$$\text{(RI) } \frac{\phi \rightarrow E\phi}{E\phi \rightarrow C\phi}$$

s'applique dans la logique choisie. Il se trouve que cette règle peut se prévaloir d'une autre intuition, qui permet de faire à nouveau le lien avec (#). Dans une logique monotone, c'est-à-dire comportant (RM), la règle suivante de monotonie s'applique pour tout entier k:

$$\frac{\phi \rightarrow E\phi}{E\phi \rightarrow E^k\phi}$$

Or (RI) s'apparente à un passage à la limite de ces règles multiples, et de ce point de vue prend en compte l'implication de la droite vers la gauche dans la formule impossible. Enfin, grâce à (RI) et à (PF) ensemble, on prouve des propriétés supplémentaires de la croyance commune qui s'accordent avec l'idée du point fixe:

$$C\phi \rightarrow CC\phi, ..., C\phi \rightarrow C^k\phi, ...$$

4. *La monotonie*: puisque la croyance partagée est monotone dans K, on peut s'attendre que la croyance commune hérite de cette propriété. On demandera donc que la règle suivante:

$$(\mathrm{RM}_C)\ \frac{\phi \rightarrow \psi}{C\phi \rightarrow C\psi}$$

s'applique dans la logique. En l'appliquant en même temps que (FP) et (RI), on peut donner des réciproques à certaines des propriétés précédentes, notamment celle-ci, qui achève de préciser l'idée de point fixe:

$$E(\phi \wedge C\phi) \rightarrow C\phi$$

Ajoutées à K, les quatre composantes (Def.E), (FP), (RI) et (RM_C) constituent une axiomatisation des modèles de Kripke pour la croyance commune. La syntaxe finitaire et la sémantique infinitaire se correspondent donc rigoureusement. Le théorème d'adéquation et de complétude est démontré par Halpern et Moses ([8]) pour un système logiquement équivalent à celui-ci; en même temps que d'autres variantes, il est discuté dans la synthèse de Lismont et Mongin ([15]).

8. Modèles minimaux de la croyance et de la croyance commune

La modélisation précédente de la croyance commune présente tous les défauts ordinaires des modèles de Kripke. Nous avons signalé que ceux-ci incorporaient une exigence pesante d'omniscience logique. Elle prend les formes suivantes, qu'il est commode de rattacher aux trois éléments constitutifs de K:

- La règle de nécessitation (N) demande que les individus croient toutes les vérités logiques du système (y compris bien entendu celles qui font référence aux croyances).
- La règle de monotonie (RM) demande qu'ils tirent les conséquences logiques de ce qu'ils croient.

- L'axiome de conjonctivité (C) demande qu'ils regroupent leurs croyances conjonctivement.

Le dernier axiome s'avère incompatible avec l'analyse *probabiliste* de la croyance à moins de considérer que seuls les événements de probabilité 1 soient objets de croyance. On peut illustrer l'incompatibilité à l'aide du prétendu *paradoxe de la loterie.*[6] Supposons la définition suivante: un individu croit à un événement F s'il lui attribue une probabilité supérieure ou égale à $1-\varepsilon$, avec, par exemple, $\varepsilon = 0.001$. Considérons une urne contenant mille billets qui ont une chance égale d'être tirés. On note G_k l'événement "le billet k ne gagne pas"; il a probabilité 0.999. L'événement $\cap_k G_k$ représente alors le fait qu'aucun billet ne gagne et il a probabilité 0. Si l'on pose que le joueur croit à l'événement F lorsqu'il lui attribue une probabilité égale ou supérieure à 0.999, il faut conclure que le joueur croit à chaque événement G_k alors même qu'il ne croit pas à l'événement conjonctif $\cap_k G_k$. Si l'on juge inadmissible une pareille configuration des croyances, on conclura que celles-ci ne peuvent se définir par l'attribution d'une probabilité supérieure ou égale à $1-\varepsilon$, si petit que soit $\varepsilon > 0$. La position la plus courante est en effet celle-ci. Mais elle suppose évidemment que l'on ne discute pas la conjonctivité de la croyance, alors que l'on peut retourner l'exemple et contester cet axiome, tout en maintenant la définition de la croyance par la probabilité $1-\varepsilon$. L'intérêt de cette résolution différente du "paradoxe" est qu'elle rend compte de la pratique ordinaire de la décision statistique: on rejette une hypothèse qui sort de l'intervalle de confiance prévu *parce que l'on n'y croit pas.*

D'un point de vue formel, la solution proposée a l'avantage de ménager une transition entre les analyses de la croyance qui reposent sur l'information et les formalismes probabilistes qui ont cours en théorie des jeux.[7]. Si l'on se rappelle enfin la critique du système K menée à partir de l'"awareness", il semble que l'on dispose d'un faisceau d'arguments pour envisager d'autres modèles d'information que celui de Kripke. Les économistes qui s'en tiennent aux correspondances de possibilité se sont en quelque sorte arrêtés en chemin dans leur réflexion.

[6]Enoncé par Kyburg et souvent discuté par la suite (voir par exemple van Fraassen [6]).

[7]Heifetz et Mongin ([11])précisent la transition. Ils développent une *logique modale probabiliste* qui revient à quantifier les structures ensemblistes d'information. La sémantique probabiliste qu'ils donnent à leur syntaxe inclut les modèles de Kripke comme un simple cas particulier, celui des événements de probabilité 1.

La logique modale épistémique propose justement une notion de *modèle minimal* (parfois attribuée au logicien Scott), qui se définit ainsi:

$$m^v = \langle \Omega, N_i, i \in I, v \rangle$$

Les relations de Kripke R_i cèdent la place aux *fonctions de voisinage*:
$N_i : \Omega \rightarrow 2^{2^\Omega}$, $\omega \longmapsto N_i(\omega) = \emptyset$ ou $\{A, A', ...\}$ avec $A, .A', ... \subseteq \Omega$,
qui peuvent être a priori quelconque, la définition n'imposant même pas la non-vacuité. Intuitivement, les $A, A', ...$, s'il en existe, se comprennent comme les événements auxquels i croit en ω. La clause sémantique essentielle s'écrira donc:

- $m, \omega \models B_i\phi$ si et seulement si $\{\omega' \in \Omega \mid m, \omega' \models \phi\} \in N_i(\omega)$.

Si peu contraignant que soit ce formalisme, il impose une forme d'omniscience logique encore. C'est la syntaxe qui la met à jour, en manifestant donc une fois encore son pouvoir d'explicitation en matière épistémique. On montre que la classe des modèles m^v est axiomatisée par les axiomes et règles du calcul propositionnel ordinaire et la *règle d'équivalence*:

$$\text{(RE)} \quad \frac{\phi \longleftrightarrow \psi}{B_i\phi \longleftrightarrow B_i\psi}$$

(Voir [3], ch.7-9.)
(RE) constitue un affaiblissement notable de (RM), qui se formule avec l'implication au lieu de la bi-implication. Néanmoins, à cause de (RE), le modèle dit minimal ne peut vraiment passer pour tel du point de vue de l'omniscience logique. L'individu qui croit à un événement décrit par une certaine formule ϕ est astreint à croire à cet événement dans *toutes* ses descriptions équivalentes. En particulier, s'il adhère à une vérité logique particulière, par exemple $p \vee \neg p$, il traitera de même *toutes les autres vérités logiques*. Dans ce cas, on retrouve indirectement l'effet de (N), dont on cherchait à se débarrasser.

Cette réserve exprimée, il reste que les modèles minimaux offrent un formalisme particulièrement commode par leur généralité: ils permettent de ressaisir comme autant de cas particuliers les restrictions précédemment imposées à la croyance. Celles-ci se traduiront par des propriétés ensemblistes naturelles, énoncées à propos des fonctions de voisinage, chaque combinaison de propriétés menant à une axiomatisation en bonne et due forme. (Voir [3].) Ainsi:

1. pour (RM): pour tout $\omega \in \Omega$, $N_i(\omega)$ est clos par surensembles;
2. pour (C): pour tout $\omega \in \Omega$, $N_i(\omega)$ est clos par intersections;
3. pour (N): pour tout $\omega \in \Omega$, $N_i(\omega)$ contient Ω;
4. pour (D): pour tout $\omega \in \Omega$, $N_i(\omega)$ ne contient pas $\emptyset$;
5. pour (T): pour tout $\omega \in \Omega$, si $A \in N_i(\omega)$, alors $\omega \in A$.

On voit notamment que les modèles minimaux ne préjugent pas des spécifications, probabilistes ou non, que l'analyse ultérieure imposerait. Si l'on veut retrouver les modèles de Kripke comme cas particuliers de modèles minimaux, il faut naturellement imposer les contraintes qui correspondent à (RM), (C) et (N). Les correspondances de possibilité s'obtiennent alors ainsi:

$$P_i(\omega) = \cap \{A \mid A \in N_i(\omega)\}$$

L'interprétation la meilleure de cet ensemble en fait la *croyance totale* de l'agent: la conjonction ensembliste des croyances A est elle-même objet de croyance et à partir duquel retrouve toutes les autres par inclusion. La définition d'un modèle minimal ne suppose rien de tel.

Les avantages des modèles minimaux ont conduit Lismont et Mongin ([15], [16] et [17]) à développer une logique de la croyance commune dans ce cadre sémantique particulier. Dans le premier travail mentionné, ils considèrent des *modèles minimaux monotones*, c'est-à-dire, en termes syntaxiques, qu'ils conservent la restriction kripkéenne (RM) tout en abandonnant (C) et (N). Dans le second travail, ils se dispensent de (RM) aussi. Dans les deux cas, la construction sémantique de la croyance commune perd de sa simplicité initiale. Il y a deux manières d'y procéder:

1. On généralise la définition itérative qui était adaptée aux modèles de Kripke. Si l'on veut retrouver toutes les intuitions liées à la croyance commune, en particulier le fait qu'elle constitue un point fixe, *il faut désormais prolonger l'itération de la croyance partagée au-delà du dénombrable.* Il en va ainsi pour la raison précise que la croyance individuelle, donc la croyance partagée, ne satisfait plus la propriété de conjonctivité. Celle-ci permettait de totaliser au fur et à mesure les croyances partagées:

$$E\phi \wedge EE\phi \wedge ... \wedge E^k\phi \longleftrightarrow E(\phi \wedge E\phi \wedge ... \wedge E^{k-1}\phi)$$

En "passant à la limite" à gauche et à droite dans cette équivalence, on dérive informellement la propriété de point fixe présente dans la syntaxe:

$$C\phi \longleftrightarrow E(\phi \wedge C\phi)$$

Lorsqu'on renonce à la conjonctivité, ce raisonnement intuitif n'est plus disponible. L'analyse mathématique confirme que si l'on continue à définir la croyance commune par une suite dénombrable de conditions, celle-ci ne satisfera plus à la propriété de point fixe. Une construction non dénombrable s'impose alors; elle est expliquée notamment dans ([15]).

2. On définit la croyance commune sans faire référence à l'itération. Une telle option est en rupture apparente avec le point de vue d'Aumann, mais elle présente un avantage de simplicité par rapport à l'autre. On peut en outre la défendre ainsi: l'idée de *point fixe* et celle d'*induction* (à partir des événements intrinsèquement publics) sont peut-être conceptuellement plus fondamentales que l'idée première de croyance partagée à un niveau logique quelconque. Suivant cette heuristique, ([16] et [17]) définissent un événement P comme *épistémiquement clos* si

$$\forall \omega \in \Omega,\ P \in \cap_{i \in I} N_i(\omega)$$

et adopte la définition suivante de la croyance commune: F est croyance commune en ω s'il existe P épistémiquement clos, de croyance partagée en ω, et tel que $P \subseteq F$. La clause sémantique s'écrit:

- $m, \omega \models C\phi$ si et seulement si $\exists P \subseteq \Omega$ épistémiquement clos, $P \in \cap_{i \in I} N_i(\omega)$ et $P \subseteq \{\omega' \in \Omega \mid m, \omega' \models \phi\}$.

Cette définition reprend l'idée que la croyance commune résulte des événements intrinsèquement publics et elle recupère également une idée de point fixe.[8] On vérifie qu'elle implique la croyance partagée à tous les niveaux logiques dénombrables, et au-delà. On démontre même qu'elle est équivalente à la construction non dénombrable de ([15]), ce qui signifie que les deux points de vue se rejoignent finalement.

[8] Autour de cette même idée, différentes variantes techniques sont concevables; voir Heifetz ([9]).

Une fois apportés ces éclaircissements sémantiques, on peut étudier l'axiomatisation des modèles minimaux de la croyance commune. La plus simple d'entre elles concerne les modèles minimaux monotones:

$$m^v = \langle \Omega, N_i, i \in I, v \rangle \text{ avec les } N_i(\omega) \text{ clos par surensembles}$$

Le système qui les axiomatise est composé de (RM) et du bloc d'axiomes et de règles relatif à la croyance commune, c'est-à-dire (Def.E), (FP), (RI) et (RM$_C$).

Les modèles minimaux, monotones ou non, n'ont pas encore eu d'applications concrètes. Cela ne saurait surprendre complètement si l'on songe que, à l'étape antérieure, les modèles de possibilités de correspondance ont mis un temps significatif à s'implanter et qu'ils n'ont eux-mêmes fourni qu'un nombre limité d'applications. Mais deux directions de recherche résultent tout naturellement de ce travail. Il faudrait approfondir la signification de l'*unawareness* pour les raisonnements naturels sur l'échange, et il serait urgent de reprendre la généralisation du théorème d'Aumann en examinant de plus près les hypothèses épistémiques dont il dépend.

References

[1] Aumann, R.J. (1976), "Agreeing to Disagree", *Annals of Mathematical Statistics*, 1236-1239.

[2] Bernecker, S. & F. Dretske (2000) (eds), *Knowledge. Readings in Contemporary Epistemology*, Oxford, Oxford University Press.

[3] Chellas, B.F. (1980), *Modal Logic*, Cambridge, Cambridge University Press.

[4] Dekel, E., B.L. Lipman & A. Rustichini (1998), "Standard State-Space Models Preclude Unawareness", *Econometrica*, 66, 159-173.

[5] Fagin, R. & J.Y. Halpern (1988), "Belief, Awareness, and Limited Reasoning", *Artificial Intelligence*, 34, 39-76.

[6] van Fraassen, B. (1995), "Fine-Grained Opinion, Probability and the Logic of Full Belief", *Journal of Philosophical Logic*, 24, 349-377.

[7] Geanakoplos, J. (1992), "Common Knowledge", *Journal of Economic Perspectives*, 6, 53-82.

[8] Halpern, J.Y. & Y.O. Moses (1992), "A Guide to Completeness and Complexity for Modal Logics of Knowledge and Belief", *Artificial Intelligence*, 54, 319-379.

[9] Heifetz, A. (1996), "Common Belief in Monotonic Epistemic Logic", *Mathematical Social Sciences*, 32, 109-123.

[10] Heifetz, A. (1999), "Iterative and Fixed Point Common Belief", *Journal of Philosophical Logic*, 28, 61-79.

[11] Heifetz, A. & P. Mongin (2001), "Probability Logic for Type Spaces", *Games and Economic Behavior*, 35, 31-53.

[12] Kaneko, M. (1999), "Epistemic Considerations of Decision-Making in Games", *Mathematical Social Sciences*, 38, 105-137.

[13] Lewis, D. K. (1969), *Convention*, Cambridge, Mass., Harvard University Press.

[14] Milgrom, P. & N. Stokey (1982), "Information , Trade and Common Knowledge", *Journal of Economic Theory*, 26, 17-27.

[15] Lismont, L. & P. Mongin (1994a), "On the Logic of Common Knowledge and Common Belief", *Theory and Decision*, 37, 75-106.

[16] Lismont, L. & P. Mongin (1994b), "A Very Weak But Non-Minimal Axiomatization of Common Belief", *Artificial Intelligence*, 70, 363-374.

[17] Lismont, L. & P. Mongin (1995), "Belief Closure: A Semantics of Common Knowledge for Modal Propositional Logic", *Mathematical Social Sciences*, 30, 127-153.

[18] Modica, S. & Rustichini, A. (1994), "Awareness and Partitional Information Structures", *Theory and Decision*, 37, 107-124.

[19] Samet, D. (1990), "Ignoring Ignorance and Agreeing to Disagree", *Journal of Economic Theory*, 52, 190-207.

[20] Samet, D. (1998), "Common Priors and Separation of Convex Sets", *Games and Economic Behavior*, 24, 172-174.

[21] Shin, H. (1993), "Logical Structure of Common Knowledge", *Journal of Economic Theory*, 69, 1-23.

[22] Shin, H.S. & T. Williamson (1994), "Representing the Knowledge of Turing Machines", *Theory and Decision*, 37, 125-146.