Intellectica, 2004/1, 38, pp. 291-323

Modélisations symboliques du raisonnement causal

Daniel Kayser, François Lévy*

Résumé : Pour l'Intelligence Artificielle, le raisonnement causal se compare à et se différencie du raisonnement déductif, et il est intéressant de le rapprocher du raisonnement sur l'action. Le problème de la modélisation est : un ensemble de relations cause/effet génériques étant connu, comment peut-on calculer effectivement l'évolution du monde quand plusieurs événements y adviennent. Nous décrivons deux grandes catégories de modélisations de la cause. La première développe des outils de type logique parmi lesquels nous nous attachons au calcul des situations, à la logique modale et aux chroniques. La seconde s'appuie sur une analyse des rapports entre description probabiliste et action, pour aboutir à une alliance entre graphes et probabilités. Une dernière partie étudie le problème de l'apprentissage, i.e. l'évolution du monde étant connue, comment retrouver les relations causales.

Mots-clés : Raisonnement causal, action, modalités, chronique, réseaux bayésiens. influence causale.

Abstract: **Symbolic models of causal reasoning**. In the AI view, causal reasoning can be compared to – and should be differentiated from – deductive reasoning, and it is interesting to examine its relationship with the reasoning on actions. The modeling problem is: generic causal relations being known, and some events occurring in some state of the world, how can its evolution be computed. Two main categories of modeling techniques are described. The first one uses logic-oriented tools: situation calculus, modal logics and chronicles. The second approach relies on the analysis of the relations between probabilistic description and action. It leads to a tool mixing together graphs and probabilities. A last part studies the learning problem: a collection of facts about events and worlds being known, how can causal relations be discovered?

Key-words : causal reasoning, action, modalities, chronicles, Bayesian networks, causal influence.

* LIPN, UMR 7030 du CNRS, Institut Galilée – Université Paris-Nord, 99 Av. Jean-Baptiste Clément – 93430 Villetaneuse.
E-mail : dk@lipn.univ-paris13.fr ou fl@lipn.univ-paris13.fr

I. INTRODUCTION

Cet article envisage le raisonnement causal du point de vue de l'Intelligence Artificielle (IA), discipline dont l'un des buts essentiels est de faire effectuer par la machine des raisonnements que les humains jugent utiles.

Il est évident qu'une part importante de notre activité, qu'elle se place dans le cadre professionnel ou dans celui de la vie quotidienne, consiste à anticiper, planifier, remédier, et que ces opérations reposent essentiellement sur des schémas cause-effet :

- si j'observe *a*, j'anticipe *b* parce que je sais que *a* cause *b* ;
- si je veux obtenir *d*, j'effectue *c* si je sais que *c* cause *d* ;
- si je constate *f*, que cela me déplaît et que je sais que *e* cause *f*, je tente d'agir sur *e*.

Le raisonnement basé sur de telles relations cause-effet est un de ceux qui nous sont le plus utiles ; il entre donc dans la vocation de l'IA de le modéliser, et de le faire d'une façon efficace.

Nous nous intéresserons ici aux modélisations symboliques. Il serait concevable de travailler sur des modèles analogiques des phénomènes en jeu, et lorsqu'on veut anticiper de façon précise une situation entièrement définissable par des paramètres physiques, c'est évidemment la meilleure solution. Cependant, nous disposons beaucoup plus fréquemment de données qualitatives, les conclusions qui nous intéressent sont également le plus souvent qualitatives, et les modèles symboliques sont généralement mieux à même de travailler sur le qualitatif.

La forme de raisonnement qui a été le mieux modélisée est incontestablement le raisonnement déductif, et l'IA s'est très largement consacrée à ce problème. Une première question est de savoir dans quelle mesure une modélisation **déductive** du raisonnement causal est adéquate et efficace. Concernant l'adéquation, l'inaptitude de la logique du premier ordre à rendre compte de la causalité a été un moteur majeur, il y a un siècle, du développement de différentes logiques modales. Nous discuterons plus loin si ces développements, pour féconds qu'ils aient été, ont réellement fait progresser la représentation de la causalité. Pour l'instant, rappelons simplement que les problèmes de décision en logique modale sont au moins aussi complexes, et souvent strictement plus complexes que leurs homologues en logique classique, et que ces derniers se trouvent déjà en dehors de la classe des problèmes solubles en temps polynomial, la seule qui corresponde à nos besoins d'efficacité, et ceci quelle que soit la

rapidité actuelle ou prévisible, même à long terme, des dispositifs de calcul.

Il est donc important de garder en mémoire que seuls de « petits » fragments d'une logique, quelle qu'elle soit, peuvent convenir à l'exigence d'efficacité, inséparable du besoin de modélisation du raisonnement causal.

Raisonnement causal et déduction

Avant d'envisager une modélisation, il importe de cerner le phénomène à modéliser. Comme on l'a déjà dit, la forme de raisonnement qui a été la plus étudiée est le raisonnement déductif. Le raisonnement causal en est-il une forme particulière ?

L'archétype du raisonnement déductif est le *modus ponens* : « si (si A alors B) est vrai et si A est vrai, alors B est vrai » que l'on peut rapprocher de : « si A cause B et que l'on constate A, alors on doit constater B aussi ». Ce rapprochement autorise-t-il à conclure, puisque le raisonnement déductif utilise le *modus ponens*, que tout raisonnement déductif est causal ?

Évidemment non ! On trouve dans tous les livres de logique la preuve de $a \supset a$, et on voit bien que ni le raisonnement suivi, ni son résultat ne correspondent à l'intuition de cause et d'effet.

Réciproquement, le raisonnement causal est-il une forme de raisonnement déductif ? Si, comme c'est l'usage, on considère qu'une des caractéristiques essentielles du raisonnement déductif est sa validité, c'est-à-dire le fait que partant de prémisses vraies, il aboutisse à des conséquences vraies, une réponse positive placerait la barre très haut ! En effet, le raisonnement quotidien s'appuie sur des observations incomplètes, souvent imparfaites, bien éloignées des canons de la vérité logique.

Ceci n'empêche pas de concevoir une idéalisation du raisonnement causal, de même qu'on peut idéaliser toute autre forme de raisonnement. Cette idéalisation mettrait pour ainsi dire entre parenthèses ces incomplétudes et ces imperfections et reposerait sur le schéma suivant : en négligeant les différences entre la situation présente et une situation dans laquelle les prémisses sont vraies, se borner à des déductions valides garantit que les conclusions sont vraies à cette approximation près. L'exigence de validité, c'est-à-dire de préservation de la vérité au cours du raisonnement, pourrait être remplacée sans trop de dommage par une exigence plus faible : que l'éventuelle « dégradation » de la vérité commise au cours du raisonnement reste marginale devant la liberté déjà prise avec la vérité en supposant les prémisses vraies.

Une telle idée est bien sûr incompatible avec le cadre classique, bi-valent, de la vérité, dans laquelle la seule « dégradation » concevable est la négation. Il faut remplacer ce cadre, soit par une conception multi-valuée, où la vérité elle-même possède d'autres valeurs que le vrai et le faux, soit par un niveau « méta », où la vérité des propositions est annotée par un commentaire, par exemple une probabilité.

Plus pragmatiquement, on peut chercher à caractériser le raisonnement causal. Ce terme recouvre au moins deux types bien distincts de raisonnements, même si on peut les mener simultanément :

- un certain nombre de relations causales étant postulées, comment les *utiliser* dans une situation donnée pour parvenir à un but plus ou moins nettement défini ;
- un certain nombre d'observations étant données, et dans un domaine sur lequel on dispose déjà d'un certain nombre de connaissances, *déterminer de nouvelles relations causales* ou renforcer celles déjà présentes (on notera que l'observation active, c'est-à-dire le choix des expériences et de leur angle d'observation en vue d'être en mesure d'inférer des relations causales relève plutôt de l'objectif précédent).

L'aspect déductif discuté ici est plus marqué dans le premier cas ; le second s'apparente davantage à une induction, puisqu'il s'agit de dégager une loi générale à partir d'observations et de connaissances.

Relations causales et action

Les objectifs que nous venons de mentionner utilisent tous deux la notion de *relation causale*, qui se trouve ainsi naturellement au cœur de toute entreprise de modélisation. Or les très nombreuses tentatives pour définir cette notion se heurtent soit à une circularité, soit à des contre-exemples ; ces difficultés sont un indice de ce que cette notion est une sorte de primitive cognitive, grâce à laquelle nous organisons nos autres connaissances. Qu'il en soit ainsi n'interdit cependant pas d'analyser un peu plus finement nos intuitions sur ce qu'est une relation causale.

Le mot « cause », comme beaucoup de mots de la langue, possède un « noyau de sens » autour duquel gravitent des acceptions qui le rendent **polysémique**. Dans ce noyau doit figurer l'idée d'un déroulement non déterminé du temps : en effet, si le fait qu'un événement advienne ou non est entièrement déterminé dès l'origine des temps, il est futile de s'interroger sur ses causes.

Remarquons en passant que ce noyau rend la notion de cause incompatible avec la Physique classique, dans laquelle le temps est

représenté comme l'ensemble des réels, et non comme une structure (arborescence, treillis, …) possédant des points d'indétermination.

Au delà de ce noyau, il est loisible de donner priorité à différents aspects des relations causales :

- la temporalité (la cause doit précéder l'effet),
- la nécessité (tout effet doit avoir une cause),
- la contiguïté (une causalité non « contiguë » serait alors un abus de langage pour désigner la fermeture transitive de relations causales élémentaires, qui seules seraient « véritables »).

Ces différentes options - et la liste n'est sûrement pas exhaustive - se décomposent elles-mêmes : les notions de nécessité et de contiguïté offrent par exemple de nombreuses et intéressantes variantes.

Nous ne voyons aucune raison de disqualifier l'une ou l'autre de ces options, bien qu'elles mènent à des différences sensibles dans l'analyse des relations causales à l'œuvre dans une situation déterminée. C'est pourquoi nous préférons parler de polysémie : suivant le contexte, une acception ou une autre de la notion de causalité sera privilégiée.

Qu'en est-il pour nous, dans le domaine de l'IA ? Notre objectif est de raisonner sur des situations du monde réel, afin d'aider les humains à anticiper, à planifier, à réparer. Dans tous ces cas de figure, il s'agit de déterminer, et éventuellement d'effectuer, des actions propres à prévenir des évolutions jugées nuisibles, à favoriser des évolutions souhaitées, ou à remédier à une situation dommageable.

C'est donc une conception interventionniste de la cause qui nous intéresse.

Encore une fois, cette conception n'a pas la prétention d'épuiser l'intuition de cause : on a à juste titre fait remarquer que des assertions comme « la lune *cause* les marées », « la rotation de la terre sur elle-même *cause* la séquence jour-nuit », paraissent conformes à l'intuition, même si elles n'évoquent aucune intervention (ôter la lune ?? arrêter la terre ??).

Si nous considérions l'interventionnisme comme faisant partie du « noyau de sens » du mot « cause », nous pourrions légitimer ces assertions en les considérant comme des extensions du sens premier, phénomène très courant dans les langues : il s'agirait d'une figure de style, à rapprocher de la métaphore, par laquelle « la lune » ou « la rotation de la terre » seraient perçues comme agentives. Mais nous

préférons ne pas nous engager dans cette voie, et nous en tenir à l'idée qu'il existe différentes variantes non hiérarchisées dans la sémantique du mot « cause », et que nous faisons le choix ici de nous intéresser à celle qui incorpore la notion d'action.

En d'autres termes, dans le contexte de l'IA, la restriction de « cause » à « cause sur laquelle on peut agir » nous paraît fondée.

II. MODÉLISATION AVEC DES OUTILS LOGIQUES

Le raisonnement causal est un cas particulier d'inférence, et les logiques ont développé une large panoplie d'outils pour modéliser les inférences. Il est donc normal de considérer d'abord comment on peut modéliser cette forme de raisonnement avec des outils logiques.

Utilisation de relations causales

Quelles inférences autorise la connaissance du fait que *a* cause *b* ? On distingue d'habitude trois domaines d'utilisation :

- la *prédiction* du comportement d'un système lorsque certaines causes s'exercent dans un état initial connu (pronostic) ; pour cela, il suffit d'utiliser les règles causales en allant des causes vers les effets (inférences progressives) ;
- la *planification* en vue d'accéder à un but, lorsqu'on connaît l'état initial ; on cherche un ensemble d'actions qui, si elles s'exercent dans le bon ordre, causeront le passage d'un état à l'autre, l'état final satisfaisant le but recherché ;
- le *diagnostic* : on dispose d'observations et il faut retrouver les causes qui ont produit ces observations ; l'état initial est l'état normal (sans panne), et on peut assimiler l'état actuel, constaté, à un état final.

Ces trois types de problèmes ne se heurtent pas aux mêmes obstacles : en planification, la difficulté vient souvent de la longueur de la série d'événements à envisager, alors qu'en diagnostic, elle réside plutôt dans l'incomplétude des observations et dans l'enchaînement des effets indirects. Par rapport à l'orientation « cause → effet » du modèle causal, ces deux derniers domaines se distinguent de la prédiction en ce qu'ils utilisent des inférences régressives.

Une différence majeure avec la notion classique d'inférence est que, quelque définition que l'on donne de la cause, on ne peut énumérer avec certitude ni l'ensemble des causes de *b*, ni l'ensemble des effets de *a*. Dès lors que la tâche à réaliser concerne le monde réel, il est impossible en pratique de décrire exhaustivement :

- les conditions nécessaires à la réalisation d'une action (problème souvent appelé *qualification* : Ginsberg et Smith, 1988).
- ses effets dans les différentes circonstances où l'action peut être exécutée (problème appelé *ramification* : Thielscher, 1997) ; ce problème inclut la description des circonstances pouvant empêcher l'effet de se produire, bien que l'action devant le causer ait été accomplie.

Ainsi par exemple, si la tâche consiste à ce qu'un robot transporte une charge d'un lieu à un autre, même si la topographie des lieux est parfaitement connue et que l'action des différentes commandes sur le robot est certaine, le plan conçu peut échouer : un obstacle inattendu peut bloquer le chemin projeté, le sol peut avoir été rendu glissant, l'objet à déplacer être plus lourd que prévu, etc. etc.

D'autre part, alors qu'en logique, l'existence d'une règle d'inférence suffit à légitimer le passage d'une prémisse à une conclusion, en raisonnement causal, il peut être nécessaire de prendre en considération les relations dans leur ensemble : dans une inférence régressive, il est en effet important de savoir si *a* est la seule cause connue de *b*.

Les règles d'inférence peuvent aussi être incertaines ou imprécises par la nature même de la situation modélisée. Ainsi en matière de diagnostic, les comportements de panne des composants sont en règle générale très mal connus, et même les domaines de validité des descriptions du comportement normal sont incertains.

Enfin, l'effet de plusieurs causes agissant ensemble est rarement l'union de leurs effets : il faut savoir comment les causes interagissent, se complètent ou bloquent les effets l'une de l'autre.

Dans le meilleur des cas, on ne peut donc raisonner que sous réserve d'informations ultérieures de nature à modifier les conclusions déjà tirées. Il est donc vain d'opposer un sens progressif (déductif, donc valide) à un sens régressif (abductif, éventuellement non valide) dans l'utilisation des relations causales. Sauf à connaître toutes les causes possibles, remonter de l'effet à la cause ne peut être qu'une hypothèse incertaine ; mais passer de la cause à l'effet repose aussi sur une hypothèse qui peut être remise en question, à savoir que toutes les conditions de validité de la loi causale sont connues et vérifiées.

Une modélisation : le calcul des situations

On se place ici dans une situation passablement idéalisée, où les actions se succèdent les unes après les autres sans recouvrement :

chaque action a fini de produire son effet avant qu'une autre n'intervienne. Cette simplification est implicite dans certaines formulations courantes. Le calcul des situations dans sa version initiale (Situation Calculus, McCarthy et Hayes, 1969) considère des *situations* s_0, …, s_n dans lesquelles certaines propriétés sont valides et d'autres pas : `vrai`(p, s_0) signifie que p est vrai dans la situation s_0 alors que ¬`vrai`(q, s_0) signifie que q n'y est pas vrai. Une fonction `résultat`(a, s) fournit la situation s' que l'on obtient après que l'action a a produit ses effets dans la situation s. On écrira par exemple :

`vrai`(`lieu`(R,Notre-Dame), s_0) & s_1 =

`résultat`(`déplace`(R, Nord, 6 km), s_0)

⊃ `vrai`(`lieu`(R,Saint-Ouen), s_1)

pour dire que le déplacement d'un robot R cause un changement de position. Chaque action produit / cause une nouvelle situation, sans considération de durée, et la seule façon de combiner deux actions a_0 et a_1 est de les enchaîner pour produire la situation `résultat` (a_1, `résultat` (a_0,s)). Implicitement, les effets sont donc traités séquentiellement sans recouvrement.

Même en se restreignant à des séquences d'actions totalement ordonnées, la description des effets à l'aide de règles causales demeure un point difficile. Une règle causale est une formule générale permettant d'exprimer l'effet d'une action. Considérons par exemple :

`vrai`(`lieu`(obj, <x,y>), s_0) & s_1 = `résultat`(`déplace`(obj, dir, dist), s_0)

⊃ `vrai` (`lieu`(obj, <x + dist.`cos`(dir), y + dist.`sin`(dir)>), s_1)

L'utilisation d'une telle règle suppose qu'on assimile l'action de commander un déplacement au succès de l'opération ainsi entreprise. En d'autres termes, cette modélisation n'envisage même pas qu'une action puisse échouer ! Pour se rapprocher du monde réel, on devrait ajouter une prémisse (ici en caractères gras), et écrire :

`vrai`(`lieu`(obj, <x,y>), s_0) & s_1 = `résultat`(`déplace`(obj, dir, dist), s_0)
& `réussit`(`déplace`(obj, dir, dist), s_0)

⊃ `vrai` (`lieu`(obj, <x + dist.`cos`(dir), y + dist.`sin`(dir)>), s_1)

mais le prédicat `réussit`, qui porte par hypothèse la conjonction de toutes les conditions nécessaires au succès de l'action, sous-entend qu'on peut énumérer ces conditions, et que leur valeur de vérité peut

être connue au moment du déclenchement de l'action, deux hypothèses fort peu réalistes !

Une autre difficulté pour *formuler* des règles causales vient de la quasi impossibilité de les décrire hors contexte, i.e. de façon modulaire. Il s'ensuit qu'une formulation apparemment correcte du fonctionnement d'un objet se trouve fréquemment en défaut dès que le contexte change. Prenons l'exemple jouet d'un interrupteur : une modélisation assez naturelle consiste à considérer deux événements `ouvrir_interrupteur` et `fermer_interrupteur`, et des relations causales :

`fermer_interrupteur` [cause] `allumé`,

`ouvrir_interrupteur` [cause] ¬`allumé`.

Mais si la lumière est commandée par deux boutons montés en série, agir sur un d'eux peut n'avoir aucun effet immédiat sur la lumière. On devra passer à une modélisation plus complexe ; par exemple, pour $i = 1, 2$:

`fermer_interrupteur_`*i* [cause] `circuit_`*i*`_fermé`,

`ouvrir_interrupteur_`*i* [cause] ¬`circuit_`*i*`_fermé`,

`circuit_`*1*`_fermé` & `circuit_`*2*`_fermé` [cause] `allumé`,

¬`circuit_`*1*`_fermé` | ¬`circuit_`*2*`_fermé` [cause] ¬`allumé`.

Remarquons au passage (avec Thielscher, 1997) qu'il faut se garder d'assimiler le connecteur [cause] à un connecteur logique : si on représentaitle montage par :

`circuit_`*1*`_fermé` & `circuit_`*1*`_fermé` ⇔ `allumé`,

dans la situation où le circuit_1 est fermé et où on ferme l'interrupteur_2, on conclurait ou bien que la lumière s'allume … ou que `circuit_1_fermé` devient faux !

Le niveau de modélisation ci-dessus convient également pour des boutons montés en parallèle, en échangeant les connecteurs & et |. Mais si les interrupteurs sont montés en va-et-vient, il faut encore une fois changer le formalisme : il n'est pas intéressant de distinguer une position ouverte et fermée des interrupteurs, car l'effet dépend de l'état antérieur de la lumière, si bien qu'il n'y a plus qu'un seul événement `bascule` par interrupteur, et qu'il faut ajouter une notation d'effet conditionnel ; pour $i = 1,2$:

`bascule_`*i* [cause] `allumé` [quand] ¬`allumé`

`bascule_`*i* [cause] ¬`allumé` [quand] `allumé`

De plus, si l'on admet la possibilité d'événements simultanés, `bascule_1` et `bascule_2` peuvent se produire en même temps, par exemple quand la lumière est allumée, mais la combinaison des deux règles ci-dessus ne produit pas dans ce cas le bon résultat.

Modélisation de la causalité en logiques modales

Les discussions qui ont conduit à la création des logiques modales ont beaucoup à voir avec la causalité, malgré un vocabulaire différent hérité d'une tradition philosophique plus ancienne. La notion de vérité nécessaire pose un distinguo entre des énoncés qui sont vrais « par hasard » (*contingents*), dont la vérité n'a pas de signification autre que d'être constatée, et des vérités *nécessaires* qui sont le reflet ou l'effet d'une structure plus profonde qui les rend, justement, nécessaires. Cette structure profonde peut être assimilée à de la causalité : dans cette optique, c'est parce que la formule est la manifestation d'une loi causale qu'elle est nécessaire. Nous reviendrons sur ce point un peu plus loin.

Techniquement, on ajoute au langage un symbole de nécessité (noté ❑) afin d'être en mesure de distinguer la proposition A (A est vraie de façon contingente) de la proposition ❑A (A est nécessairement vraie). On cherchera alors quelles formules logiques ou quels axiomes rendent compte de la notion de nécessité, par exemple : ce qui est nécessaire est vrai (❑A ⊃ A). On peut avec ce dispositif former une grande variété de logiques modales ; seule une partie d'entre elles a à voir avec la notion de nécessité et ce qu'elle peut emporter de causalité.

Les logiques modales ont une interprétation en termes de modèles. Au lieu d'un monde, on considère une famille de mondes possibles dont chacun peut interpréter différemment les « atomes » du langage. De chaque monde, on peut en envisager d'autres qui sont dits *accessibles*. Une formule ❑F est vraie dans un monde *m* si F est vraie dans chacun des mondes accessibles depuis *m*.

L'idée la plus simple de la nécessité est obtenue quand chaque monde voit tous les autres – une formule est alors nécessaire si elle est vraie dans tous les mondes possibles[1]. Dans ce cas, l'empilement des modalités (on peut écrire ❑❑❑...) est limité : ❑❑A équivaut à ❑A.

Ceci dit, quand on examine ses propriétés formelles, la notion de nécessité ne réussit qu'imparfaitement à traduire la causalité :

- Une formule logiquement toujours vraie (une tautologie, par exemple 3 = 3) se trouve par là même valide dans tous

[1] La logique s'appelle alors S5.

les mondes, donc nécessaire. Mais une tautologie ne représente pas une relation causale ! Philosophiquement, on peut raffiner l'analyse des raisons pour lesquelles une proposition est vraie, distinguant dans la tradition de Kant les raisons internes à la notion (l'analytique) de celles qui sont construites (le synthétique), les raisons qui préexistent par nature à l'expérience (l'a priori) de celles qui sont constatées (a posteriori). L'aspect proprement philosophique de la discussion sort de notre compétence.

- L'implication « matérielle » : $(p \supset q)$ de la logique classique ne correspondant manifestement pas, comme on l'a déjà dit, à l'intuition qu'on a de la notion de cause, on a introduit une implication « stricte » qui aurait, elle, valeur causale : $(p \succ q)$. La plupart des logiques modales assimilent cette implication à $\Box(p \supset q)$; mais cette formule, qui correspond à la vérité de $(p \supset q)$ dans tous les mondes possibles, traduit bien mal la causalité : l'implication matérielle est telle que si p est faux, $(p \supset q)$ est vrai, donc si p est par exemple la négation d'une tautologie, $(p \supset q)$ acquerra le statut de nécessité indépendamment de q. Un énoncé tel que « 3 = 4 est cause de ce que Noël se fête le 25 décembre » ne saurait pourtant être déclaré conforme à notre intuition de la causalité !!

- Les discussions propres à l'IA ont souligné un autre point, déjà mentionné : la notion de cause s'accommode assez mal de la validité dans *tous* les mondes accessibles. Pour reprendre un exemple de J. McCarthy, on considère généralement que « tourner la clé de contact cause le démarrage du moteur », tout en sachant que ceci présuppose que les circonstances ne soient pas trop exceptionnelles : si le réservoir d'essence a été siphonné, ou la batterie vidée, le moteur ne démarrera pas. La notion de cause est, par rapport à la loi physique, une simplification qui décrit de façon économique un comportement normal.

Cette idée de conséquence normale a donné lieu à une traduction en logique modale dans Delgrande (1988). Pour cela, on ajoute au langage de la logique modale un connecteur conditionnel noté $\Rightarrow$[2] : $A \Rightarrow B$ se lit « si A, normalement B ». L'enchâssement des connecteurs $\Rightarrow$ est interdit. L'accessibilité entre mondes possibles sert à définir une proximité entre mondes qui à son tour fournit une traduction sémantique de la notion de normalité. Pour chaque ensemble E de mondes accessibles depuis un certain monde m, on

[2] à ne confondre ni avec l'implication matérielle $\supset$, ni avec l'implication stricte $\succ$ évoquées plus haut.

définit un sous-ensemble de mondes de E « les plus proches » de *m*. Quand on prend pour E l'ensemble des mondes où A est valide, on obtient les mondes normaux pour *m* et A. Autrement dit, A ⇒ B est vrai en *m* si B est vrai dans les mondes normaux pour *m* et A.

La logique N obtenue ainsi permet de raisonner sur les règles de comportement normal sans interdire les exceptions : il peut exister des mondes accessibles depuis *m* (mais non normaux) dans lesquels ces règles ne s'appliquent pas. A cause B se traduira par « si A se produit à l'instant t, normalement on a B à l'instant t+1 ». On peut combiner les règles de comportement normal pour obtenir, par exemple, que si on met le contact, puis qu'on accélère en embrayant, normalement la voiture se met en route.

Les difficultés de modélisation ressurgissent quand on veut conclure sur les individus. En effet, rien ne permet de savoir si l'occurrence de A sur laquelle on veut raisonner relève d'une situation normale ou non. Le problème de la fusillade de Yale[3] a été fréquemment utilisé pour illustrer ce type de paradoxe : on a un pistolet et une victime possible ; le propriétaire du pistolet l'a chargé, a attendu puis a tiré. Dans le formalisme de la logique N, le problème est décrit par :

i. `Vrai` (`Vivant`, s_0)

la victime est vivante en s_0

ii. `Vrai` (`Chargé`, `Résultat` (`Charger`, *s*))

après avoir chargé un pistolet, il est chargé

iii. `Vrai` (`Chargé`, *s*) ⇒ `Vrai` (`Mort`, `Résultat` (`Tirer`, *s*))

tirer sur une victime avec un pistolet chargé entraîne généralement sa mort

iv. `Vrai` (f, *s*) ⇒ `Vrai` (f, `Résultat` (e, *s*))

les propositions demeurent généralement vraies (persistance)

On notera que les deux derniers axiomes utilisent le connecteur conditionnel, donc admettent les exceptions. On considère donc la situation :

s_3 = `Résultat`(`Tirer`, `Résultat`(`Attendre`, `Résultat`(`Charger`, s_0)))

[3] « Yale Shooting Problem » dans la littérature anglo-saxonne.

(on a chargé, puis attendu, puis tiré) : `Vivant` est normalement vrai dans s_3 par persistance, i.e. par application répétée de iv, et `Mort` y est normalement vrai par application de iii et par la persistance de `Chargé` après `Résultat`(`Charger`, s_0).

Dans un tel cas de conflit (en précisant bien sûr qu'on ne peut être à la fois vivant et mort), la logique N ne devient pas inconsistante, mais elle ne peut décider où est l'exception, et donc elle ne peut tirer aucune conclusion. Or la description initiale laisse à penser que la conclusion normale est `Mort`. On ne sait pas résoudre autrement que par des règles ad hoc le conflit entre l'effet des événements et la persistance des propriétés non affectées.

Une modélisation du déroulement temporel : les chroniques

Beaucoup d'auteurs ont noté, à juste titre selon nous, le caractère « contrefactuel » de la causalité. Lorsqu'on affirme que « *a* a causé *b* », on a en tête que « toutes choses égales par ailleurs » si *a* n'avait pas eu lieu, *b* ne se serait pas produit. On veut donc être en mesure de comparer dans un même système deux cas de figure incompatibles : *a* a lieu au temps *t*, et *a* n'a pas lieu au temps *t*.

Or la modélisation habituelle du temps, inspirée de la Physique, consiste à voir *t* comme un nombre réel. Comme aucun raisonnement cohérent ne peut s'appuyer sur des propositions contradictoires telles que `se_produit`(*a*,*t*) et ¬`se_produit`(*a*,*t*), le raisonnement causal ne peut se satisfaire d'une telle modélisation. McDermott (1982) propose de garder la continuité du temps, tout en ajoutant une « ouverture vers le futur » ; pour cela il introduit une notion d'*état* et une relation de précédence antisymétrique[4] et transitive[5] $\preceq$ entre états ($\prec$ si on retire l'égalité), à laquelle il impose un axiome de densité :

$$\prec (e_1, e_2) \supset \exists e_3\ (\prec (e_1,e_3)\ \&\ \prec (e_3,e_2)),$$

et il ajoute : $(\preceq (e_1, e)\ \&\ \preceq (e_2, e)) \supset (\preceq (e_1, e_2) \mid \preceq (e_2, e_1))$.

Cette dernière formule (la linéarité à gauche) signifie que le passé est unique, puisque si deux états sont antérieurs à un même troisième, ils sont comparables entre eux. L'ouverture vers le futur se traduit par l'absence d'un axiome symétrique : deux états peuvent être postérieurs à un même troisième sans être comparables, ce qui veut dire qu'un même état peut être suivi de futurs distincts.

En chaque état, comme dans une logique modale ordinaire, les propositions possèdent une valeur de vérité : on notera comme

[4] $(\preceq (e_1, e_2)\ \&\ \preceq (e_2, e_1)) \supset = (e_1, e_2)$

[5] $(\preceq (e_1, e_2)\ \&\ \preceq (e_2, e_3)) \supset \preceq (e_1, e_3)$

précédemment `vrai`(p, e) le fait que la valeur de la proposition p dans l'état e soit « vrai ».

Chaque état possède une *date* qui est un réel, et la précédence des états correspond à la relation d'ordre sur les réels ; en d'autres termes, $\prec (e_1, e_2) \supset < (\text{date}(e_1), \text{date}(e_2))$, mais la réciproque est fausse : il ne suffit pas que la date d'un état e_1 soit antérieure à la date d'un autre e_2 pour que e_1 précède e_2. Encore faut-il qu'ils appartiennent à une même *chronique*.

La notion de chronique est l'élément clé de cette modélisation : on pourra avoir `se_produit`(a,e_1) et ¬`se_produit`(a,e_2) avec $\text{date}(e_1) = \text{date}(e_2)$ pourvu que e_1 et e_2 appartiennent à des chroniques différentes. McDermott définit donc une chronique c comme un ensemble d'états totalement ordonné, et en correspondance avec l'axe des réels, c'est-à-dire :

$$(\in (e_1,c) \,\&\, \in (e_2,c)) \supset (\preceq (e_1, e_2) \mid \preceq (e_2, e_1))$$

$$(\forall t)\ (\exists e)\ (\in (e,c) \,\&\, =(\text{date}(e), t)).$$

La notion suivante est celle d'*événement* ; une occurrence d'événement est un intervalle d'états appartenant à une même chronique. On note `Occ`(e_1, e_2, ev) le fait que l'événement ev se produit exactement entre les états e_1 et e_2.

La causalité est traduite par deux notions distinctes : la première est une *causalité entre événements*, notée `Ecause`(ev_1, ev_2, p, pt, d) qui exprime l'idée suivante : toute occurrence de l'événement ev_1 sera suivie d'une occurrence d'ev_2 après un délai d, pourvu que la proposition p reste vraie, faute de quoi la relation de cause à effet serait inhibée. Le paramètre pt indique à partir de quand le délai d commence à courir (pt varie de 0 à 1 selon que le délai part du commencement ou de la fin de l'événement ev_1). Cette idée se traduit formellement par l'expression :

$$(\texttt{Ecause}(ev_1, ev_2, p, \text{pt}, \text{d}) \,\&\, \texttt{Occ}(e_1, e_2, ev_1)) \supset (\forall c)\ (\in (e_2,c) \supset \exists e_3\ (\in (e_3,c) \,\&\, \texttt{dans\_les\_délais}(e_3, \text{pt}, \text{d}, e_1, e_2) \,\&\, (\forall e)\ (e_2 \preceq e \preceq e_3 \supset \texttt{vrai}(p, e)) \supset (\exists e_4)\ (\in (e_4,c) \,\&\, \texttt{Occ}(e_3, e_4, ev_2)))),$$

le prédicat `dans_les_délais` étant défini par :

$$\texttt{dans\_les\_délais}(\text{e}, \text{pt}, \text{d}, e_1, e_2) \Leftrightarrow \text{début}(\text{d}) \leq \text{date}(\text{e}) - (1\text{-}\text{pt}).\text{date}(e_1) - \text{pt}.\text{date}(e_2) \leq \text{fin}(\text{d}).$$

Commentaires : la première formule exprime que si ev_1 cause ev_2, alors toute chronique c comportant une occurrence d'ev_1 et dans laquelle la proposition p se maintient à vrai au moins de la fin d'ev_1 jusqu'à un état e_3 se situant dans un certain délai, contiendra également une occurrence d'ev_2 commençant en cet état e_3.

La deuxième formule précise les limites du délai ; si l'événement ev_2 est réputé se produire après un intervalle d de 25 à 35 minutes à partir d'un point de déclenchement pt se situant au tiers de l'événement ev_1, si ev_1 s'est produit de 12h à 12h30, un état e sera « dans les délais » s'il survient entre 12h35 et 12h45 (car ce sont les dates qui satisfont 0h25 ≤ date(e) - 2/3 12h - 1/3 12h30 ≤ 0h35).

La deuxième forme de causalité, notée `Pcause`, rend compte des cas où c'est une proposition qui est causée par un événement. Mais McDermott signale ici une difficulté : inférer qu'une proposition est vraie en un état est de peu d'utilité pratique, car cela n'apprend rien sur sa valeur de vérité en tout autre état. On se sert donc de connaissances sur le monde concernant la *persistance* des propositions pour « extrapoler » les résultats ainsi obtenus : de ce qu'on a placé un meuble dans un appartement, on est en droit d'inférer qu'il va rester en place plusieurs mois ; mais de ce qu'on a placé un journal sur ses genoux, on ne saurait tirer pareille conclusion. Définir la persistance *per* d'une proposition *p* par le fait que si *p* est vraie en e, elle sera vraie dans toute chronique incluant e en tout état dont la date est comprise dans l'intervalle [date(e), date(e) + *per*] n'est pas satisfaisant pour au moins deux raisons :

- d'une part, la chronique peut inclure un événement qui invalide *p* ; donc il faut préciser « dans toute chronique incluant e et n'incluant aucun événement connu pour causer ¬*p* avant l'expiration de la durée *per* » ;
- d'autre part et surtout, une telle définition manque sa cible : si on associe à une certaine proposition *p* une persistance *per*, disons d'une heure, et qu'on sait *p* vraie à midi, cela veut dire qu'en l'absence d'événement contraire, *p* sera vraie à une heure, d'où on pourra inférer qu'en l'absence d'événement contraire, elle sera encore vraie à deux heures, à trois heures … Donc la valeur de *per* n'a aucune signification, puisque quelque valeur positive qu'on lui donne, on aboutit au même résultat : la proposition reste vraie pour l'éternité dans toute chronique n'incluant pas d'événement la rendant fausse.

La solution choisie par McDermott consiste à causer non pas une proposition, mais la persistance d'une proposition à partir d'un certain état. Techniquement, cela donne :

$$(\texttt{Pcause}(ev, p, q, \text{pt}, \text{d}, per) \,\&\, \texttt{Occ}(\text{e}_1, \text{e}_2, ev)) \supset (\forall c)\,(\in(\text{e}_2, c) \supset \exists \text{e}_3\,(\in(\text{e}_3, c) \,\&\, \texttt{dans\_les\_délais}(\text{e}_3, \text{pt}, \text{d}, \text{e}_1, \text{e}_2) \,\&\, (\texttt{vrai}(p, \text{e}_3) \supset \texttt{Persiste}(\text{e}_3, q, per))))).$$

La définition de `Persiste`(e, *p*, *per*) nécessite d'introduire un événement `Fait_cesser`(*p*) et exprime que *p* est vraie en e, et que si aucun événement `Fait_cesser`(p) ne se produit pendant le délai *per*, alors *p* est vrai en tout état e' vérifiant e $\prec$ e' et date(e') $\leq$ date(e) + *per*.

Commentaires : la formule ci-dessus exprime que l'événement *ev* cause la persistance de *q* pour une durée *per*, pourvu que la proposition *p* se maintienne à vrai jusqu'à un certain état e_3 se situant dans les délais expliqués précédemment. En définissant la persistance à partir d'un certain état, et non à partir de la constatation de la vérité d'une proposition, on évite l'effet itératif signalé plus haut. Enfin, considérer qu'un événement cause non la vérité instantanée, mais la persistance d'une proposition, permet de tirer des conclusions utiles.

Remarquons pour terminer que les formules explicitant `Ecause` et `Pcause` sont des implications simples et non des équivalences. Il ne s'agit donc pas de définitions : on peut en effet observer que toute chronique contenant un événement ev_2 contient également auparavant un événement ev_1 sans vouloir en inférer une relation causale. Nous discuterons le problème de l'inférence de relations causales dans la section IV.

Les idées qui sont à la base de ce formalisme ont été développées dans différents contextes. Une généralisation récente en est donnée dans Bennett et Galton (2004).

Causalité et non-monotonie

On a utilisé dans les paragraphes précédents des expressions telles que « sauf si certain événement survient » ou « pourvu que toutes les conditions normales soient satisfaites », ou encore « toutes choses égales par ailleurs ». C'est le genre d'expression facile à écrire, mais dont la signification n'est pas toujours très claire.

La première semble non problématique : l'événement survient ou ne survient pas. Mais elle cache une modalité : le raisonnement causal donne des résultats en fonction des données dont il dispose ; s'il ne dispose pas d'informations sur la survenue de l'événement, doit-il conclure qu'il s'est produit ou qu'il ne s'est pas produit ? La validité logique impose une prudente abstention. Mais l'exigence d'utilité du raisonnement causal ne peut se satisfaire de cette abstention, et, en utilisant implicitement l'adage « si c'était vrai, cela se saurait », on aura tendance à postuler qu'il ne s'est pas produit.

Ce même mode de raisonnement est encore plus visible dans les autres cas : il est rare de disposer d'une liste exhaustive des « conditions normales », et très invraisemblable d'avoir

suffisamment d'informations pour montrer que chacune de ces conditions est satisfaite (c'est le problème de la *qualification* mentionné plus haut). Pour avancer dans le raisonnement, on se contente d'utiliser implicitement l'idée que s'il y avait une anomalie, ou si les conditions avaient été modifiées d'une façon significative, cette information aurait été donnée.

Ce procédé est légitime, car, comme nous l'avons vu, l'exigence de validité est incompatible avec l'efficacité du raisonnement. Cependant, il faut être conscient de la fragilité des conclusions auxquelles on parvient ainsi : si de nouvelles informations nous parviennent, ou si un raisonnement nous amène à conclure que tel événement est bien survenu, des conclusions déjà tirées devront être invalidées. Cette fragilité n'est pas spécifique du raisonnement causal ; on la retrouve, sous le nom de non-monotonie dans tous les secteurs de l'IA, et elle a donné lieu à des traitements variés[6].

Dans Kayser et Mokhtari (1998) nous nous sommes inspirés de la modélisation de McDermott pour introduire la notion d'*action*, notion dont nous avons expliqué dans l'introduction du présent article pourquoi nous la jugions fondamentale dans le traitement du raisonnement causal par l'IA. Dans ce travail, nous insistons également sur l'importance de la *norme* dans la causalité. Nous y reviendrons en conclusion.

III. MODÉLISATION À BASE DE GRAPHES

Causalité et probabilité

> « Le mot *cause* ne fait pas partie du vocabulaire standard de la théorie des probabilités. C'est un fait étrange mais indéniable que la théorie des probabilités, le langage mathématique officiel de nombreuses sciences expérimentales, ne permet pas d'exprimer des phrases telles que : « la boue ne cause pas la pluie » (…) et doivent surajouter au langage des probabilités un vocabulaire pour la causalité permettant de différencier cette phrase de

[6] Pour l'anecdote, citons McDermott (1982) à ce sujet : « I try to appeal to non-monotonic deductions as seldom as possible. This is because the logics they are based on are still rather unsatisfactory (…). In attempting to represent things, it is helpful to be as formal as we can, but if the formal systems cannot keep up with the inferences we want to make, so much the worse for the formal systems. In the long run, I am confident that non-monotonic logics will be developed that capture the inferences we need. (…) If representation designers make it clear what they need, logicians will make it work. » (p.122) On peut considérer l'article de Giunchiglia, Lee, Lifschitz, McCain et Turner (2004) comme une confirmation récente de cette prédiction.

> « la boue est indépendante de la pluie ». Bizarrement, cette distinction reste à incorporer dans l'analyse scientifique classique »
>
> J. Pearl, 2000, p.134

Considérons un exemple classique : Monsieur Dupont se réveille un beau matin et constate que sa pelouse est mouillée. Il se demande s'il a plu durant la nuit, ou si son arroseur automatique s'est mis en marche d'une façon intempestive. Le raisonnement qu'il devrait suivre pour évaluer s'il a plu est le suivant :

$$p(il\ a\ plu \mid ma\ pelouse\ est\ mouillée) = \frac{p(ma\ pelouse\ est\ mouillée \mid il\ a\ plu) \times p(il\ a\ plu)}{p(ma\ pelouse\ est\ mouillée)}$$

c'est-à-dire une application directe du théorème de Bayes. Mais s'il constate que l'arroseur s'est effectivement mis en route pendant la nuit, le fait que sa pelouse soit mouillée ne lui est d'aucune utilité pour savoir s'il a plu, et il ne peut alors qu'appliquer :

$$p(il\ a\ plu \mid ma\ pelouse\ est\ mouillée) = p(il\ a\ plu)$$

Cet exemple est utilisé par Judea Pearl, dont nous adopterons l'analyse dans cette section, pour illustrer la différence entre observation et intervention. Le calcul de la probabilité pour qu'une grandeur Y ait la valeur *y* sachant que la grandeur X vaut *x* ne donne pas, en général, le même résultat selon que *x* est une valeur *observée*, ou que c'est une valeur *imposée*. Cette distinction, fondamentale pour le raisonnement causal, n'est généralement pas représentée en statistiques. C'est pourquoi Pearl propose de différencier :

$p(Y = y \mid X = x)$ d'une part, et $p(Y = y \mid \text{do}(X = x))$ d'autre part.

Si on dispose d'une système d'équations décrivant comment la valeur de chaque grandeur X_i dépend de son environnement, c'est-à-dire un système de la forme :

$$p(X_i = x_i) = f(p(X_j = x_j)) \text{ pour } j \neq i$$

[la fonction f a autant d'arguments que le nombre de grandeurs qu'il faut connaître pour évaluer la probabilité de X_i], la résolution de ce système donne les probabilités lorsqu'on **observe** certaines valeurs. Si on **intervient** sur la grandeur X_i en lui donnant la valeur x_0, il suffit de remplacer dans ce système l'équation régissant X_i par :

$$p(X_i = x_i) = 1 \text{ si } x_i = x_0,\ 0 \text{ sinon}$$

Les réseaux bayésiens : présentation et théorème fondamental

Le système d'équations dont il est question ci-dessus n'est pas d'un usage bien agréable, et les probabilités qu'il requiert sont souvent difficiles à estimer. Les réseaux bayésiens que nous allons brièvement présenter ici, sont beaucoup plus proches de l'intuition, et surtout ils permettent de raisonner **qualitativement** sur les facteurs qui causent telle ou telle variation.

L'hypothèse de base est que l'on peut, pour chaque grandeur X_i considérée, identifier ses « parents markoviens », notés $pa(X_i)$, c'est-à-dire les grandeurs dont il est nécessaire et suffisant de connaître les valeurs pour que la distribution de probabilité de X_i soit connue. Formellement, si les grandeurs $X_1, \ldots, X_n$ forment une suite, $pa(X_i)$ est un ensemble minimal, au sens de l'inclusion, tel que :

$$p\left(X_i = x_i \mid X_1 = x_1, \ \ldots \ X_{i-1} = x_{i-1}\right) = p\left(X_i = x_i \mid X_j = x_j \text{ pour } j \in pa(X_i)\right)$$

Un réseau bayésien est un graphe orienté sans cycle, obtenu en prenant pour nœuds chacune des grandeurs X_i et pour arcs les relations $X_j \rightarrow X_i$ pour $j \in pa(X_i)$. On dira qu'une distribution de probabilité est *compatible* avec un tel graphe si les égalités ci-dessus sont vérifiées.

Considérons quelques exemples (Becker et Naïm, 1999) et tout d'abord celui de la pluie et de la pelouse : pour savoir s'il a plu ou si c'est son arroseur qui a fonctionné, Monsieur Dupont peut regarder la pelouse de son voisin : si elle est également mouillée, cela renforcera notablement la probabilité qu'il ait plu. Donc si Z (*il a plu*) est parent markovien de X (*ma pelouse est mouillée*) et de Y (*la sienne aussi*), et qu'on ignore la valeur de Z, la valeur de X peut influer sur la probabilité de Y. En revanche, si la valeur de Z est connue (Monsieur Dupont sait qu'il a plu), cette influence n'existe pas (le fait que sa pelouse est mouillée ne modifie pas la probabilité que celle de son voisin le soit, dans le schéma simplifié où on est situé).

Le deuxième exemple est celui d'une transitivité : X est parent de Y, lui-même parent de Z ; par exemple X est l'*ensoleillement*, dont la valeur influe sur Y, la *quantité récoltée*, qui elle-même influe sur Z, le *prix*. Tant qu'on ignore Y, la connaissance de X modifie la distribution de probabilité de Z. Mais dès que Y est connu, X n'influe plus sur Z (dans le schéma choisi, où on a supposé que l'influence de X sur Z est indirecte).

Le troisième exemple est celui d'une confluence de causes : X et Y sont parents de Z ; par exemple, Z est le *déclenchement d'un système d'alarme* qui peut provenir soit de X, un *cambriolage*, soit de Y, un *tremblement de terre*. X et Y n'ont aucun lien a priori. Cependant, la connaissance de Z peut créer un tel lien : si le système d'alarme s'est déclenché, on affecte une probabilité à ce que la cause

soit un cambriolage ; si on apprend qu'il y a eu un tremblement de terre, cette probabilité va décroître.

Ces exemples anecdotiques et simplistes veulent illustrer la variété des circonstances dans lesquelles la connaissance d'un facteur peut influer sur la probabilité d'un autre. Essayons de formaliser cette idée, puis de voir quel type de raisonnement peut nous permettre de tirer des conclusions générales.

La notion sous-jacente est celle d'**indépendance conditionnelle** ; on dit que X et Y sont indépendants, conditionnellement à Z ssi[7] $p(X=x \mid Y=y \ \& \ Z=z) = p(X=x \mid Z=z)$, c'est-à-dire que la distribution de probabilité de X ne dépend plus de la valeur de Y dès que la valeur de Z est connue.

Un théorème de Verma et Pearl (1988) permet de déterminer s'il y a ou non indépendance conditionnelle par le seul examen du graphe d'un réseau bayésien, sans aucun calcul.

Son énoncé requiert la définition de quelques notions de théorie des graphes : un *chemin* entre X et Y est une suite de nœuds $X_0=X$, …, $X_n=Y$ telle que pour tout i entre 0 et *n*-1, l'arc $X_i \rightarrow X_{i+1}$ appartienne au graphe ; s'il existe un chemin entre X et Y, on dit que Y est un *descendant* de X ; une *chaîne* entre X et Y est une suite de nœuds $X_0=X$, …, $X_n=Y$ telle que pour tout i entre 0 et *n*-1, au moins l'un des deux arcs $X_i \rightarrow X_{i+1}$ ou $X_{i+1} \rightarrow X_i$ appartient au graphe. En d'autres termes, une chaîne est une liaison entre deux nœuds sans tenir compte de l'orientation des arcs. Une chaîne *converge* en X_i ssi les arcs $X_{i-1} \rightarrow X_i$ et $X_{i+1} \rightarrow X_i$ appartiennent au graphe. Elle *diverge* en X_i ssi les arcs $X_i \rightarrow X_{i-1}$ et $X_i \rightarrow X_{i+1}$ appartiennent au graphe. Elle est *série* en X_i ssi les arcs $X_{i-1} \rightarrow X_i$ et $X_i \rightarrow X_{i+1}$ appartiennent au graphe.

Dans les exemples précédents, la chaîne entre *tremblement de terre* et *cambriolage* converge en *système d'alarme*, celle entre *ma pelouse est mouillée* et *celle de mon voisin aussi* diverge en *il a plu*, et celle entre *ensoleillement* et *prix* est série en *quantité récoltée*.

On dit qu'une chaîne est *bloquée* par un ensemble de nœuds Z ssi elle diverge ou est série en un nœud appartenant à Z, ou si elle converge en un nœud qui n'appartient pas à Z et dont aucun descendant n'appartient à Z. Deux ensembles de nœuds X et Y sont dits *séparés* par Z si toute chaîne reliant un élément de X à un élément de Y est bloquée par Z.

Le théorème dit la chose suivante : si X et Y sont séparés par Z, alors X et Y sont indépendants conditionnellement à Z dans toute

[7] ssi est l'abréviation usuelle de « si et seulement si »

distribution de probabilité compatible avec le graphe ; réciproquement, si X et Y ne sont pas séparés par Z, il existe au moins une distribution de probabilité compatible avec le graphe dans laquelle X et Y ne sont pas indépendants conditionnellement à Z.

Ce théorème affirme donc que les trois exemples ci-dessus sont représentatifs de l'ensemble des cas de figure :

- la pluie bloque (divergence) la chaîne reliant les deux pelouses : sachant s'il a plu, l'état de ces deux pelouses est indépendant ;
- la quantité récoltée bloque (série) la chaîne reliant l'ensoleillement au prix : connaissant la quantité récoltée, le prix est indépendant de l'ensoleillement ;
- la chaîne reliant le tremblement de terre au cambriolage n'est pas bloquée par l'alarme (la convergence se fait en un nœud de Z), donc il y a dépendance entre ces paramètres quand le statut de l'alarme est connu.

Causalité et réseaux bayésiens

L'insuffisance des moyens classiques pour tirer des conséquences de données statistiques est illustrée avec brio dans la discussion que Pearl consacre au paradoxe de Simpson (Pearl, 2000, pp.174 sqq.). Rappelons ce dont il s'agit en examinant le tableau statistique ci-dessous (fabriqué pour les besoins de la cause) :

population	soin		guéri	non guéri	% guérison
80			36	44	45%
	OUI	40	20	20	50%
	NON	40	16	24	40%

Le taux de guérison est donc légèrement meilleur chez les patients qui se soignent.

Supposons que la population est répartie par parts égales entre hommes et femmes, et que les chiffres soient les suivants :

hommes	Soin		guéri	non guéri	% guérison
40			25	15	62,5%
	OUI	30	18	12	60%
	NON	10	7	3	70%
femmes	Soin		guéri	non guéri	% guérison
40			11	29	27,5%
	OUI	10	2	8	20%
	NON	30	9	21	30%
total (pour mémoire)	*Soin*		*guéri*	*non guéri*	*%guérison*
80			*36*	*44*	*45%*
	OUI	*40*	*20*	*20*	*50%*
	NON	*40*	*16*	*24*	*40%*

Le taux de guérison est donc meilleur chez les hommes qui ne se soignent pas … ainsi que chez les femmes qui ne se soignent pas ! L'amélioration constatée sur la population globale est en fait une illusion due au phénomène suivant : les hommes sont naturellement beaucoup plus enclins à guérir de cette maladie, et comme ils sont plus nombreux à se soigner, on a l'impression que ce sont les soins qui guérissent.

Mais considérons maintenant le même tableau avec un autre scénario. Nous avons toujours 80 patients, dont la moitié se soigne et les taux de guérison sont toujours ceux indiqués. Mais en cours de traitement, on mesure un paramètre, disons la pression artérielle, et les mesures donnent les résultats suivants :

soin	amélioration de la pression artérielle		guéri	non guéri	% guérison
40			20	20	50%
	OUI	30 (75%)	18	12	60%
	NON	10	2	8	20%
Pas de soin	amélioration de la pression artérielle		guéri	non guéri	% guérison
40			16	24	40%
	OUI	10 (25%)	7	3	70%
	NON	30	9	21	30%
total (pour mémoire)	*amélioration de la pression artérielle*		*guéri*	*non guéri*	*%guérison*
80			*36*	*44*	*45%*
	OUI	*40*	*25*	*15*	*62,5%*
	NON	*40*	*11*	*29*	*27,5%*

Les malades qui se soignent ont beaucoup plus de chances de voir leur pression artérielle s'améliorer (75% contre 25%), et ceux dont la pression artérielle s'est améliorée ont beaucoup plus de chances de guérir (62,5% contre 27,5%). Les soins sont donc efficaces !

Donc avec les mêmes chiffres[8], on conclut (correctement) dans un cas que les soins sont néfastes, et dans l'autre qu'ils sont utiles. Ces conclusions ne sont donc pas dues seulement aux chiffres, mais également au schéma causal sous-jacent, et l'impression de paradoxe est due une fois encore à la confusion entre : $p(Y = y \mid X = x)$ et

[8] Les deux tableaux diffèrent par leur mode de présentation : le premier regroupe les patients par sexe, le deuxième les regroupe selon qu'ils se soignent ou non, mais le lecteur pourra vérifier qu'en changeant l'étiquetage *homme* / *femme* en *amélioration* / *pas d'amélioration* de la pression artérielle, ce sont exactement les mêmes chiffres qui sont fournis dans les deux cas.

$p(Y = y \mid \textbf{\textit{do}}\ (X = x))$. Ce qui importe en effet n'est pas la valeur de p(guéri = OUI | soin = OUI), mais celle de p(guéri = OUI | ***do*** (soin = OUI)), et cette dernière s'évalue différemment suivant le schéma causal postulé : on suppose que les soins n'influent pas sur le sexe du patient, alors qu'on suppose qu'ils influent sur l'amélioration de sa pression artérielle.

Calcul des influences

On a vu plus haut comment calculer la modification d'une probabilité due à une intervention lorsqu'on dispose d'un système d'équations. Le problème qui se pose concerne l'adaptation de ce calcul lorsqu'on représente les probabilités dans un réseau bayésien. Pearl l'a résolu dans deux cas de figure importants : pour énoncer ses résultats, il nous faut ajouter quelques définitions.

Une chaîne est dite relier X et Y *par l'arrière* ssi elle contient un arc orienté vers X. Un ensemble de nœuds Z satisfait le *critère arrière* par rapport à une paire de nœuds X, Y ssi aucun nœud de Z n'est un descendant de X, et si Z bloque toutes les chaînes reliant X et Y par l'arrière. Il satisfait le *critère avant* si les trois conditions suivantes sont réunies : tout chemin reliant X à Y contient un nœud de Z, il n'y a aucune chaîne reliant X par l'arrière à un nœud de Z, et toutes les chaînes reliant par l'arrière un nœud de Z à Y sont bloquées par X.

On a les théorèmes suivants : si Z satisfait le critère arrière par rapport à X et Y, alors le calcul de l'intervention sur x ne requiert que la connaissance de Z, et plus précisément, on a :

$$p(Y = y \mid \textbf{\textit{do}}(X = x)) = \sum_{\text{toute valeur } z \text{ de } Z} \left[p(Y = y \mid X = x,\ Z = z)\, p(Z = z) \right]$$

Si Z satisfait le critère avant, on a :

$$p(Y = y \mid \textbf{\textit{do}}(X = x)) = \sum_{\text{toute valeur } z \text{ de } Z} \left[p(Z = z \mid X = x) \sum_{\text{toute valeur } x' \text{ de } X} p(Y = y \mid X = x',\ Z = z)\, p(X = x') \right]$$

Dans le schéma causal : soin → amélioration de la pression artérielle → guérison, le critère avant est vérifié, et le calcul donne p(guéri | ***do*** (soin = OUI)) = 0,485 tandis que p(guéri | ***do*** (soin = NON)) = 0,175. Le bénéfice des soins est ainsi rendu très visible.

Dans le schéma causal : sexe → guérison, sexe → soin → guérison, les critères ci-dessus ne s'appliquent pas et on doit recourir au système d'équations : lorsqu'on remplace les probabilités de soin

par p(soin = OUI) = 1, on obtient p(guéri | ***do*** (soin = OUI)) et la valeur numérique est 0,4 ; de même, si on les remplace par p(soin = NON) = 1, on obtient p(guéri | ***do*** (soin = NON)) dont la valeur numérique est 0,5. La nocivité des soins est ainsi établie.

IV. ACQUISITION DE RELATIONS CAUSALES

Introduction

Même si la notion de causation est considérée comme une primitive cognitive, on ne saurait se dispenser de chercher d'où vient notre connaissance de relations causales particulières. Il n'entre pas dans nos compétences de présenter ce que le comportement humain révèle sur les étapes d'acquisition de ces relations, et nous nous limiterons ici à discuter quelles données peuvent être exploitées.

Il est bien connu que la simple observation de co-occurrences ou de successions régulières ne peut suffire à expliquer comment se construisent les relations causales : sinon, pour reprendre un exemple connu, nous croirions que lorsque le soleil est en face de nous, l'arrivée de l'ombre d'une personne cause l'arrivée de cette personne ! Cette insuffisance a pu faire croire à l'impossibilité d'une méthode objective d'acquisition de ces relations.

Le problème est le suivant : supposons que la nature obéisse à des règles déterministes stables représentées sous forme d'un graphe sans cycle entre des variables. Dans quelle mesure est-il possible de re-constituer ce graphe à partir de l'observation ?

Approche statistique

Becker et Naïm (1999) récapitulent l'approche purement statistique de la reconstruction du réseau causal. Celle-ci repose sur la distance de Kullback-Leibler (KL-distance) entre deux distributions de valeurs f et g (probabilité théorique ou fréquence mesurée sur un échantillon) pour la même variable aléatoire X :

$$l(f, g) = \sum_{x \in X} f(x) \log(\frac{f(x)}{g(x)})$$

Cette mesure vient de la théorie de l'information : si f et g proviennent de distributions identiques, la seconde n'apporte aucune information par rapport à la première.

Il s'agit donc de produire le réseau bayésien dont la KL-distance aux observations soit minimale. Employé seul, ce principe conduirait à sur-générer des relations causales pour améliorer l'approximation. L'approche purement statistique propose deux solutions à cette difficulté. La première restreint la recherche à des réseaux en forme

d'arbre ; on a alors une solution générale en théorie des graphes (un arbre couvrant de poids maximal). La seconde modifie la distance en ajoutant un terme de pénalité en fonction de la complexité du réseau.

L'approche bayésienne utilise plutôt la formule de Bayes de probabilité des causes pour comparer la probabilité de deux structures étant donné les observations. Si l'on fait l'hypothèse que toutes les structures ont la même probabilité *a priori* (avant toute observation) et que les observations sont indépendantes l'une de l'autre[9], on peut identifier dans le calcul une formule g(Ci) représentant la qualité d'un jeu de parents donné. L'algorithme K2 de Cooper et Herskovits (1992) implante alors une heuristique qui prend en entrée un ordre sur les nœuds tel que toute cause précède son effet, et fournit en sortie un réseau causal en général proche de l'optimal.

Approche graphique

Pearl (2000) construit une vision plus locale de la causalité, appuyée sur l'hypothèse d'un mécanisme sous-jacent générant les fréquences observées (une chaîne de Markov) : il s'agit de montrer que ce mécanisme peut être au moins partiellement reconstruit à partir des observations.

Nous illustrerons notre discussion par l'exemple jouet suivant : on observe trois sorties d'un montage électronique A, B et C, et celles-ci valent 0 ou 1. Le montage est ainsi fait que A et B sont indépendants avec une probabilité p_A (resp. p_B) de valoir 1, alors que C vaut 1 ssi A = B, mais l'observateur n'a pas accès au schéma du dispositif et ne peut se fier qu'à ses observations. A-t-il un moyen d'inférer la loi qui régit ce montage ?

Il consigne ses observations dans le tableau ci-dessous :

[9] Un jeu d'observation ne représente donc pas une série temporelle, i.e. une succession de mesures prises sur un système dynamique. Il s'agit seulement de tirages au sort répétés.

A	B	C	Proba[10]
1	1	1	p_1
1	0	0	p_2
0	1	0	p_3
0	0	1	p_4

ainsi que les 4 combinaisons ci-contre qui, par la nature du montage, ne se produisent jamais

A	B	C	Proba
1	1	0	0
1	0	1	0
0	1	1	0
0	0	0	0

On a évidemment $p_1 + p_2 + p_3 + p_4 = 1$. La probabilité constatée de A=1 est $p_1 + p_2$, celle de A=0 est $p_3 + p_4$,celle de B=1 est $p_1 + p_3$, etc.

Considérons les trois graphes suivants (L, M, N représentent A, B ou C, mais on ne sait pas qui influence qui) :

(1) L ⟶ M N

(L et M sont corrélés, N est indépendant de chacun des deux autres)

(2) L ⟶ M ⟶ N

(L et M sont corrélés, M et N aussi ; quand M est connu, L et N sont indépendants)

(3) L ⟶ N ⟵ M

(L et M sont indépendants, N est corrélé à L&M. Quand N est connu, L et M sont corrélés)

Nous utilisons ces graphes pour exprimer les hypothèses que peut faire l'observateur qui ne connaît pas le mécanisme sous-jacent, et qui veut tester si elles sont compatibles avec les valeurs observées (rappelons que le montage réel correspond au graphe (3) avec C dans le rôle de N).

Si les sorties A et B sont indépendantes, on a les 4 relations :

$$p(A=i \ \& \ B=j) = p(A=i) * p(B=j) \text{ pour } i= 0 \text{ ou } 1, j= 0 \text{ ou } 1.$$

Comme $p(A=0) + p(A=1) = 1$, $p(B=0) + p(B=1) = 1$, une seule de ces relations implique les autres. Nous gardons par exemple celle obtenue pour i=j=1 :

$$p_1 = (p_1 + p_2)*(p_3 + p_4).$$

[10] En toute rigueur, on ne constate pas des probabilités, mais des fréquences. Nous supposerons que les observations se font sur un nombre suffisant d'expériences pour que cette assimilation soit admissible.

De même, l'indépendance de B et C se traduit par :

$$p_1 = (p_1 + p_3)*(p_1 + p_4),$$

et celle de A et C, par : $p_1 = (p_1 + p_2)*(p_1 + p_4)$.

Supposons tout d'abord que le montage réel est tel que p_A et p_B valent 1/2.

L'observation donne alors une valeur de 1/4 pour p_1, p_2, p_3 et p_4, et donc non seulement p(A=0), p(A=1), p(B=0), p(B=1), mais aussi p(C=0) et p(C=1) valent 1/2.

Comme par exemple p(B=1 & C=1) = 1/4 = p(B=1) * p(C=1), l'observateur conclut que B et C sont indépendantes. Il conclut de même pour les autres couples de variables.

On a cependant p(C=1 | B=1& A=1) = 1, alors que p(C=1 | A=1) = 1/2, ce qui prouve la non-indépendance de B et C conditionnellement à A. Autrement dit, les variables sont deux à deux indépendantes, mais elles ne le sont pas conditionnellement à la troisième variable.

Ceci exclut les graphes (1) et (2) : dans (1), les probabilités conditionnelles et inconditionnelles de N sont identiques ; dans (2), N est indépendant de L conditionnellement à M.

Par contre, on ne peut trancher entre les 3 instanciations possibles du graphe (3).

Supposons maintenant que p_A=0,6 et p_B=0,7.

Les valeurs constatées seront donc p_1=0,42, p_2=0,18, p_3=0,28 et p_4=0,12.

L'observateur constate que p(A=1 | B=1) = 0,42/0,7 = 0,6 = p(A=1), donc il est en mesure de conclure que A et B sont indépendants. Il voit que p(C=1) = p_1 + p_4 vaut 0,54, donc que p(A=1 | C=1 & B=1) = 1 ≠ p(A=1 | C=1) = 0,42/0,54 = 7/9, i.e. A et B ne sont pas indépendants conditionnellement à C ; pour A et B la situation est identique au cas précédent.

Par contre, A et C ne sont ni indépendants puisque p(A=1 | C=1) ≠ p(A=1), ni indépendants conditionnellement à B {p(A=1 | C=1 & B=1) ≠ p(A=1 | B=1)}. De même, B et C ne sont indépendants ni inconditionnellement {p(B=1 | C=1) = 0,42/0,54 ≠ p(B=1) = 0,7}, ni conditionnellement à A {p(B=1 | C=1 & A=1) = 1 ≠ p(B=1 | A=1) = 0,42/0,6=0,7}.

Le second schéma reste donc exclu (on aurait au moins une indépendance conditionnelle), et dans le troisième les places

de L et M sont nécessairement tenues par A et B, les seules variables indépendantes.

Dans ce cas de figure, l'observateur est donc en mesure d'inférer le graphe correspondant à la réalité du dispositif.

Cet exemple jouet nous montre que, selon les valeurs de certains paramètres, l'inférence fournit la solution, ou la laisse indécise parmi un ensemble de solutions possibles.

En général, il n'est pas réaliste de supposer que tout est observé. Dans la version la plus simple, toutes les variables essentielles sont accessibles, les autres constituent un contexte dont chaque variable perturbe aléatoirement une seule fonction du système, de sorte que le contexte crée indépendamment une marge d'incertitude sur chacune des variables observées. Dans ce cas, les corrélations entre observables résultent des interactions décrites dans le graphe causal, et il s'agit de reconstituer un système minimal d'interactions expliquant ce qu'on observe.

Il arrive aussi que certaines corrélations demeurent ainsi inexpliquées – on en déduit que l'hypothèse était trop forte, que certaines variables non observées (les variables « latentes ») sont communes à la genèse de plusieurs variables et créent ainsi des corrélations indirectes. Il y a alors une infinité de solutions : en effet, ajouter à un schéma compatible avec les données un nombre arbitraire de variables latentes correctement conditionnées donnera encore un schéma compatible avec les données. On contraindra donc davantage la réponse en n'utilisant que le minimum de variables latentes.

Cette contrainte d'économie de moyens ne permet cependant de caractériser qu'un fragment du graphe causal. On dit qu'un graphe G' peut en *simuler* un autre G si pour toute réalisation des paramètres de G donnant aux variables observées [O] la distribution de probabilité $P_{[O]}$, il existe une réalisation des paramètres de G' donnant à O la même distribution. G est alors préféré à G' au nom de l'économie de moyens. G est *minimal* si chaque fois que G peut simuler G', G' peut aussi simuler G.

On obtient ainsi une famille de graphes minimaux indiscernables à l'aide des variables observées. On dira que la variable C a une *influence causale* sur la variable E si et seulement si, dans tout graphe minimal compatible avec la distribution de probabilité des observables, il existe un arc de C vers E.

La construction des influences causales repose sur deux propriétés :

- Les nœuds non directement reliés par un arc du graphe sont indépendants conditionnellement aux nœuds intermédiaires qui les séparent. Seuls les nœuds directement reliés n'ont aucune indépendance conditionnelle.
- Quand deux nœuds X et Y non directement reliés ont un voisin commun E, soit celui-ci fait partie d'un ensemble conditionnellement auquel X et Y sont indépendants, soit X et Y sont tous deux des causes de E (les arcs joignant X à E et Y à E peuvent alors être orientés vers E).

Pearl observe cependant que le critère de minimalité peut s'avérer insuffisant pour discriminer entre plusieurs explications causales, et il introduit une autre notion : notre préférence ne va pas toujours vers la minimalité en tant que telle, mais vers la *stabilité*. Sans entrer dans une définition technique de ce terme, nous adapterons ici un exemple (Pearl, 2000, p.49) qui nous semble bien l'éclairer : soient deux variables Y et Z dépendant linéairement d'une même grandeur X, mais indépendantes l'une de l'autre conditionnellement à X ; on peut modéliser leur comportement par le système d'équations :

$$y = ax + bu_1 \quad z = cx + du_2$$

où u_1 et u_2 sont les valeurs de variables latentes. Mais on peut également refléter l'influence potentielle de Y sur Z par le système suivant, qui ajoute l'indépendance conditionnelle comme une contrainte sur les coefficients :

$$y = ax + bu_1 \quad z = ey + fu_1 + du_2, \qquad f = -eb$$

La contrainte $f = -eb$ garantit que la valeur de z est déterminée dès que celles de x et u_2 le sont. Cette deuxième solution, qui ne postule pas plus de variables latentes que la première, est cependant *instable* car l'indépendance entre y et z n'est pas structurelle : une légère variation dans l'estimation des paramètres f, e ou b y mettrait fin. L'indépendance instable ne justifie pas pour Pearl l'absence de lien, et seule une indépendance stable doit être prise en compte pour construire les graphes minimaux.

On remarquera que, dans l'exemple jouet discuté plus haut, la configuration $p_A = p_B = 1/2$ correspondait à une instabilité, puisque certaines indépendances constatées étaient en réalité fallacieuses, dues à des valeurs particulières des paramètres (ici l'équiprobabilité généralisée).

Des algorithmes construisant les influences causales à partir de distributions stables sont décrits dans Pearl (2000). Leur résultat est un graphe possédant quatre sortes d'arcs : X est

- une cause authentique de Y,
- une cause potentielle de Y (d'autres facteurs que X peuvent expliquer la valeur de Y),
- en association illusoire avec Y (il existe en fait une variable latente qui conditionne à la fois X et Y),
- en relation indéterminée avec Y (on ne dispose pas d'information suffisante pour savoir dans lequel des trois cas précédents on se trouve).

La connaissance du déroulement temporel peut faciliter la découverte de relations causales, mais il est intéressant de noter qu'en son absence, on peut synthétiser une notion de *temps statistique*. Cette notion est définie comme un ordre des variables qui respecte au moins un graphe causal minimal compatible avec les probabilités observées. Pearl émet la conjecture que le temps « physique » coïncide avec au moins un temps statistique. Les équations physiques étant généralement réversibles, le déroulement du futur vers le passé devrait également coïncider avec un temps statistique ; mais, comme le fait observer l'auteur, le choix des variables observées n'est pas indifférent, et conduit à favoriser les schémas orientés dans un sens causal respectant l'antériorité temporelle de la cause sur l'effet.

V. CONCLUSION

Au terme d'un parcours techniquement plutôt rude, le lecteur peut se demander en quoi tous ces modèles font progresser notre compréhension de la notion de cause.

Rappelons d'abord que notre prétention n'est pas de couvrir toutes les acceptions que l'intuition confère à cette notion, et que nous nous sommes principalement intéressés aux causes sur lesquelles on peut agir. Mais nous n'avons guère discuté jusqu'ici comment modéliser cette « possibilité d'action ».

Un exemple classique (Mackie, 1974) consiste à envisager un homme qui, en allumant une cigarette, provoque une explosion. Si l'événement se produit dans une raffinerie, on dira volontiers que c'est *le fumeur* qui a causé l'explosion ; s'il a lieu chez lui, on considérera plutôt que c'est *une fuite de gaz*, elle-même due à un acte de malveillance, à un défaut d'entretien ou à tout autre facteur, qui est à l'origine de l'explosion.

Notre intuition sur la cause de l'accident diffère en raison de nos connaissances sur les actions que nous jugeons *normales* en telle ou telle circonstance. Fumer chez soi n'a rien d'extraordinaire ; dans une raffinerie, il en va différemment. Parmi les possibilités d'action qui

peuvent être retenues comme causalement responsables dans une situation considérée, nous conjecturons que la cognition humaine tend à ne considérer comme causes que les plus anormales.

Cette conjecture implique l'existence d'une échelle de normalité, qui semble difficile à mettre en évidence (on n'a d'ailleurs besoin que d'un ordre partiel). Et les outils que nous avons envisagés dans cet article ne semblent guère adéquats pour la modéliser, à l'exception peut-être de la logique de Delgrande.

C'est pourtant une question d'une grande importance cognitive ; il y a longtemps que Schank (Schank et Abelson, 1977) a montré que la compréhension des textes utilise intensivement ce qu'il appelle des *scripts*, c'est-à-dire des déroulements stéréotypés reflétant ce qu'il est habituel de faire dans telle ou telle circonstance de notre vie personnelle ou sociale. Plus récemment, les travaux sur l'anaphore associative ont confirmé qu'il faut disposer de connaissances prototypiques pour interpréter correctement de nombreuses constructions du langage. Les outils de logique non-monotone, qui sont de toute manière nécessaires pour modéliser le raisonnement causal, peuvent également être mis à contribution pour représenter ce qui est considéré comme usuel ou normal. Encore faut-il les nourrir des connaissances appropriées, c'est-à-dire exprimer correctement la relation qui relie une situation déterminée aux actions qui sont réputées normales dans cette situation.

Un programme de recherche consistant à mettre en évidence les normes implicites d'un domaine et à les hiérarchiser, à peine effleuré jusqu'à présent, devrait concerner, et profiter à un sous-ensemble important des disciplines appartenant à la constellation des Sciences Cognitives.

Bibliographie

Becker, A., Naïm, P. (1999). *Les réseaux bayésiens. Modèles graphiques de connaissance.* Eyrolles.

Bennett, B., Galton, A. P. (Mars 2004). A unifying semantics for time and events. *Artificial Intelligence* vol.153 n^{os} 1-2 pp.13-48,

Cooper, G.F., Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning* vol.9 pp.309-347.

Delgrande, J. P. (Août 1988). An Approach to Default Reasonning Based on a First-Order Conditional Logic: Revised Report. *Artificial Intelligence*, vol 36 pp.63-90.

Ginsberg, M. L., Smith, D. E. (Juillet 1988). Reasoning about action II: the qualification problem. *Artificial Intelligence* vol.35 n°3 pp.311-342.

Giunchiglia, E., Lee, J., Lifschitz, V., McCain, N., & Turner, H. (Mars 2004). Nonmonotonic causal theories. *Artificial Intelligence* vol.153 n^{os} 1-2 pp.49-104.

Kayser, D., Mokhtari, A. (Février 1998). Time in a Causal Theory. *Annals of Mathematics and Artificial Intelligence*, vol.22 n^{os} 1-2 pp.117-138.

Mackie, J.L. (1974). *The Cement of the Universe: A Study of Causation*. Oxford University Press.

McCarthy, J., Hayes, P. J. (1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence, in: *Machine Intelligence 4* (B. Meltzer & D. Michie eds.) pp.463-502, Edinburgh University Press.

McDermott, D. (1982). A Temporal Logic for Reasoning About Processes and Plans. *Cognitive Science* vol.6, 101-155.

Pearl, J. (2000). *Causality. Models, reasoning, and inference*. Cambridge University Press.

Schank, R. C., Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum Ass.

Thielscher, M. (Janvier 1997). Ramification and causality. *Artificial Intelligence* vol.89 n^{os}1-2 pp.317-364.