

Actions et causalité : essais de formalisation en logique

Robert Demolombe*
ONERA Toulouse

1 Introduction

Les actions peuvent être étudiées selon de nombreux points de vue [16]. Ici nous nous intéressons à la notion de causalité, c'est-à-dire à la relation entre un agent i et une propriété p qui est vraie dans l'état du monde où on se trouve après que l'agent ait réalisé une certaine action. L'objectif est de définir à quelles conditions on peut dire que c'est l'action réalisée par i qui est la cause du fait que l'on a p .

Par exemple, quand on dit informellement que “Paul a cassé la vitre”, on veut dire que la vitre est cassée à cause de l'action réalisée par Paul. On ne s'intéresse pas à la manière dont Paul a cassé la vitre, c'est-à-dire au type d'action qu'il a réalisé. Il peut, par exemple, avoir fermé brutalement la fenêtre, ou bien avoir heurté la fenêtre en déplaçant une table. Il peut aussi l'avoir cassée en lançant un caillou. Dans ce cas on voit que la relation entre l'agent et l'effet observé n'est pas si évidente qu'il semble au premier abord. On pourrait en effet se demander si c'est Paul ou si c'est le caillou qui a cassé la vitre. Il faut donc préciser le sens qu'on donne à cette relation de causalité. La même question se pose si Paul déclenche l'exécution d'un programme qui détruit un fichier. Dans quel sens c'est le programme, ou Paul, qui est la cause de la destruction?

Avant d'essayer de répondre à cette question il est intéressant de dire dans quels contextes elle a été posée. Depuis des siècles elle a été posée dans le domaine de la morale et de la religion pour savoir si une personne avait vraiment commis une faute. Par exemple, une faute vis-à-vis de la règle “tu ne tueras pas”. Si une personne a tué pour obéir à un ordre, ou sous la contrainte, ou involontairement, est-ce que l'on peut dire qu'elle est la cause de la mort? Les hommes de loi se sont aussi posé la question pour déterminer si une personne est la cause d'un délit, et donc si elle est condamnable. On comprend alors pourquoi cette question a fait l'objet depuis longtemps de nombreuses réflexions qui nous

*ONERA-DTIM, 2 Avenue E. Belin B.P. 4025, Toulouse Cedex. e-mail : Robert.Demolombe@cert.fr.

sont très utiles, même si notre motivation ici est différente.

En effet, ici on s'intéresse à des contextes qui sont en rapport avec l'informatique. Soit parce qu'on considère des agents qui sont des logiciels, ou des personnes qui interagissent avec des logiciels. Soit parce qu'on veut modéliser des interactions entre des agents quelconques, mais qu'on veut automatiser les raisonnements faits sur ces modèles. Dans un cas comme dans l'autre on a besoin de définitions formelles.

Par exemple, la notion de causalité intervient dans la modélisation de l'organisation de systèmes socio-techniques où certains agents sont des personnes et d'autres sont des agents artificiels (par exemple, un robot ou un logiciel). Dans ce contexte on définit des obligations, des permissions, et des interdictions, qui peuvent porter sur les actions des agents. On peut avoir, par exemple, les deux règles :

Il est permis que (l'agent i ouvre le coffre $C1$)

Il est interdit que (l'agent j ouvre le coffre $C1$)

Dans ce cas, si le coffre $C1$ est ouvert il faut savoir qui en est la cause. Et la réponse n'est pas évidente si l'agent j , par exemple, a demandé (ou a ordonné, ou a contraint) à l'agent i d'ouvrir le coffre.

On peut aussi prendre en exemple une situation où la crédibilité des informations stockées dans une base de données dépend des agents qui y insèrent ces informations. Supposons qu'on assigne à une information p des niveaux de crédibilité différents selon qu'elle est insérée par l'agent i ou par l'agent j . Dans le cas où i demande à j d'insérer p , il n'est pas clair que l'on doive assigner à p le niveau de crédibilité de i ou de j . Par exemple, si j se contente de transmettre l'information comme le fait un facteur, il est vraisemblable que l'on assignera à p le niveau de i , car ce n'est pas j qui est la cause du fait que la base croit p . Par contre, si j ne transmet l'information que s'il la croit vraie, et si le niveau de j est supérieur à celui de i , il est vraisemblable que l'on assignera à p le niveau de j .

Nous allons présenter dans les sections suivantes trois types de formalisations qui correspondent à des étapes chronologiques, et qui vont de propositions simples vers des propositions plus affinées ¹. Elles ont toutes les trois été formalisées dans des logiques modales classiques ou normales (voir [3]).

2 Formalisation proposée par G.H. von Wright

L'idée principale dans la proposition de von Wright [18] est que l'agent i est la cause du fait qu'on a p ssi i a influencé l'évolution du monde de telle sorte qu'on ait p . Plus précisément l'attitude d'un agent est caractérisée par 3 états du monde :

¹Pour ce travail de synthèse nous nous sommes beaucoup inspiré de [17] et [6].

1. l'état initial, avant que l'action ait commencé,
2. l'état qui résulte de l'action réalisée par l'agent,
3. l'état qui aurait été atteint si l'agent n'avait pas influencé l'évolution du monde.

Sa définition formelle est donnée dans le cadre de structures de Kripke de la forme : $M = \langle W, d, e, T \rangle$, où W est un ensemble de mondes possibles, d et e ² sont des fonctions de W dans W qui font correspondre à un monde w dans l'état 1, un monde $d(w)$ dans l'état 2, et un monde $e(w)$ dans l'état 3, T est une fonction qui fait correspondre à chaque nom de proposition atomique l'ensemble des mondes dans lesquels cette proposition est vraie.

La fonction T est étendue aux formules du calcul des propositions classique de façon habituelle :

- $T(\neg p) = W - T(p)$
- $T(p \vee q) = T(p) \cup T(q)$

les autres connecteurs logiques sont définis à partir des disjonctions (\vee) ou des négations (\neg) comme d'habitude. La condition de satisfaisabilité des formules est alors :

$$M, w \models p \text{ ssi } w \in T(p)$$

Le langage du calcul des propositions est étendu avec deux modalités notées Br_i et Ss_i ³. Intuitivement $Br_i p$ signifie que l'agent i a fait en sorte que l'on ait p , et $Ss_i p$ signifie que l'agent i a maintenu le monde dans un état où on a p .

Pour définir dans quelles conditions on peut dire que $Br_i p$ et $Ss_i p$ sont vraies ou fausses, von Wright considère toutes les attitudes possibles de l'agent i vis-à-vis de p . Comme on peut avoir soit p soit $\neg p$ en w , $d(w)$ et $e(w)$, il y a 8 attitudes possibles.

Supposons, par exemple, que la proposition p signifie "la porte est ouverte". On dira que dans le monde w il est vrai que l'agent i a fait en sorte que la porte soit ouverte (en abrégé, i a ouvert la porte), si en w la porte était fermée (on a $\neg p$ en w), et en $d(w)$ la porte est ouverte (on a p en $d(w)$), et en $e(w)$ la porte est fermée (on a $\neg p$ en $e(w)$) (voir figure 1). On voit qu'il y a deux conditions pour caractériser le fait que c'est l'action réalisée par i qui est la cause du fait que la porte est ouverte.

²Ces fonctions sont notées respectivement d et e pour rappeler l'expression "doing" et "empty action".

³Les notations Br_i et Ss_i correspondent respectivement aux expressions "to bring it about that" et "to sustain".

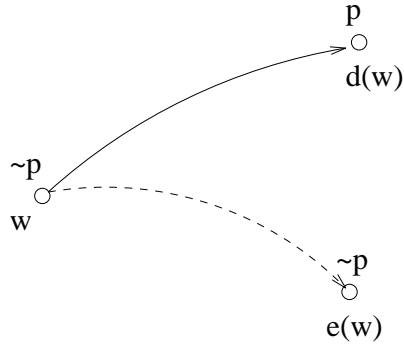


Figure 1: $Br_i p$: faire en sorte que p .

La première condition, qui vient tout de suite à l'esprit, est qu'après que i ait agi la porte est ouverte. Mais la deuxième condition est essentielle, elle exprime que c'est bien à cause de ce qu'a fait i que la porte est ouverte. Car si i n'avait pas fait ce qu'il a fait la porte serait restée fermée.

En effet, si l'action de i avait consisté simplement, par exemple, à fumer une cigarette, ou à chanter une chanson, et que la porte soit ouverte en $d(w)$ parce qu'un courant d'air l'a poussée, ou bien parce qu'un autre agent j l'a ouverte, alors la deuxième condition n'aurait pas été satisfaite. En effet, même si l'agent n'avait pas fumé, ou n'avait pas chanté, la porte aurait été néanmoins ouverte par le courant d'air, ou par l'agent j , et en $e(w)$ on aurait p au lieu de $\neg p$.

Cette deuxième condition, appelée "counter action condition" est fondamentale pour caractériser la causalité. Plus simplement, quand on dit que c'est l'action réalisée par i qui est la cause du fait que l'on a p , on veut dire, entre autre, que si i n'avait pas fait cela on n'aurait pas p .

Il peut y avoir d'autres situations où on peut dire que c'est ce qu'à fait i qui est la cause du fait qu'on a p . Supposons, par exemple, qu'en w la porte soit ouverte (on a p en w) et que souffle un courant d'air qui pourrait fermer la porte si celle-ci n'était maintenue ouverte par i . Alors en $d(w)$ la porte est encore ouverte (on a p en $d(w)$) et si l'agent n'avait rien fait elle aurait été fermée (on a $\neg p$ en $e(w)$) (voir la figure 2). Dans ce cas on dit que i a maintenu la porte ouverte, et formellement on a $Ss_i p$ vraie en w .

Pour caractériser certaines situations où il n'est pas vrai que i a fait en sorte que p , formellement, où on a $\neg Br_i p$, von Wright a introduit un opérateur logique, noté om , qui a un sens plus spécifique que la négation. Cet opérateur exprime que l'agent se trouve dans une situation où il aurait pu faire en sorte que p , mais il ne l'a pas fait.

Par exemple, si en w la porte était fermée (on a $\neg p$ en w) et quand l'agent i n'agit pas la porte est encore fermée (on a $\neg p$ en $e(w)$), alors, si après qu'il

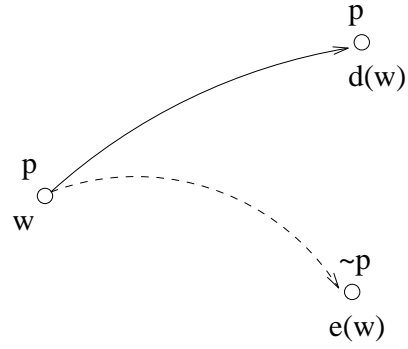


Figure 2: $Ss_i p$: maintenir p .

ait agit la porte est encore fermée (on a $\neg p$ en $d(w)$), on peut dire que i avait l'opportunité d'ouvrir la porte, mais qu'il n'a pas saisi cette opportunité. Intuitivement, on peut dire que i a omis, ou s'est abstenu, d'ouvrir la porte (voir figure 3).

Il y a des cas où, bien que i n'ait pas fait en sorte que p (où on a $\neg Br_i p$), il n'est pas vrai que i a omis de faire en sorte que p (on a $\neg om Br_i p$). C'est le cas, par exemple, quand la porte a été ouverte par un courant d'air, c'est-à-dire où on a $\neg p$ en w , p en $d(w)$ et p en $e(w)$. Dans ce cas i n'avait pas l'opportunité d'être la cause de l'ouverture de la porte, car la porte a été ouverte indépendamment de ce qu'il a fait.

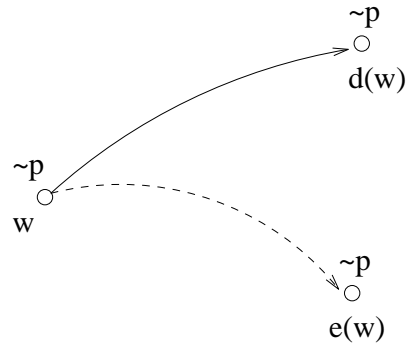


Figure 3: $om Br_i p$: omettre de faire en sorte que p .

Les conditions de satisfaisabilité des opérateurs Br_i et Ss_i sont définies dans le cas général par :

- $M, w \models Br_i p$ ssi $M, w \models \neg p$ et $M, d(w) \models p$ et $M, e(w) \models \neg p$.

- $M, w \models Ss_i p$ ssi $M, w \models p$ et $M, d(w) \models p$ et $M, e(w) \models \neg p$.

L'ensemble des 8 attitudes possibles de l'agent i par rapport à p sont présentées dans le tableau ci-dessous.

	w	d(w)	e(w)	
Action1	$\neg p$	p	$\neg p$	$Br_i p$
Action2	p	p	$\neg p$	$Ss_i p$
Action3	p	$\neg p$	p	$Br_i \neg p$
Action4	$\neg p$	$\neg p$	p	$Ss_i \neg p$
Action5	$\neg p$	p	p	$omSs_i \neg p$
Action6	p	p	p	$omBr_i \neg p$
Action7	p	$\neg p$	$\neg p$	$omSs_i p$
Action8	$\neg p$	$\neg p$	$\neg p$	$omBr_i p$

La première partie du tableau, Action1 à Action4, caractérise les attitudes de l'agent qui ont une influence sur le fait que la porte est ouverte ou fermée. En effet, la valeur de vérité de p n'est pas la même selon qu'il agit (en $d(w)$), ou qu'il n'agit pas (en $e(w)$). Son influence peut consister à empêcher une certaine évolution du monde, cas de Action 2 et Action4, ou au contraire, à provoquer une certaine évolution du monde, cas de Action1 et Action3.

La deuxième partie du tableau, Action5 à Action8, caractérise les attitudes de l'agent qui n'ont pas d'influence sur le fait que la porte est ouverte ou fermée. En effet, la valeur de vérité de p est la même qu'il agisse (en $d(w)$), ou qu'il n'agisse pas (en $e(w)$). Dans le cas de Action6 et Action8, il avait l'opportunité de faire en sorte que la porte soit ouverte ou fermée.

L'opérateur d'action Br_i que nous venons de voir pose le problème suivant. D'après les conditions de satisfaisabilité, on peut vérifier facilement que pour n'importe quel monde w de n'importe quelle structure M on a : $M, w \models Br_i p \rightarrow \neg p$, et donc on a $\models Br_i p \rightarrow \neg p$. Ceci ne correspond pas du tout à l'intuition, ou bien alors il faudrait interpréter $Br_i p$ comme : "l'agent i est sur le point de faire en sorte que p ". Il serait plus naturel de définir un opérateur dont le sens intuitif serait "l'agent i a fait en sorte que p " pour lequel $Br_i p \rightarrow p$ serait une formule valide. C'est pour cette raison que von Wright a proposé les opérateurs Br_i^1 et Ss_i^1 dont les conditions de satisfaisabilité sont :

- $M, w \models Br_i^1 p$ ssi il existe un monde u dans W tel que $w=d(u)$ et $M, u \models \neg p$ et $M, w \models p$ et $M, e(u) \models \neg p$.
- $M, w \models Ss_i^1 p$ ssi il existe un monde u dans W tel que $w=d(u)$ et $M, u \models p$ et $M, w \models p$ et $M, e(u) \models \neg p$.

Pour l'opérateur Br_i^1 on a effectivement la propriété attendue : $\models Br_i^1 p \rightarrow p$.

3 Formalisation proposée par S. Kanger et I. Pörn

Kanger dans [10] puis Pörn dans [15] ont proposé des opérateurs modaux pour exprimer la notion de causalité qui sont plus généraux que ceux proposés par von Wright, dans la mesure où l'état du monde qui résulte du fait que l'agent a agi, ou n'a pas agi, n'est pas unique. Dans un certain sens on peut dire que les actions réalisées par les agents ne sont pas déterministes.

Les structures dans lesquelles ces opérateurs sont définis sont des n -uplets de la forme $M = \langle W, R_i, R'_i, T \rangle$ où W est un ensemble de mondes possibles, mais ici un élément de W correspond à une histoire du monde, et non à un état du monde à un instant donné comme pour von Wright, R_i et R'_i sont deux relations définies sur $W \times W$, T est défini comme précédemment.

Les relations R_i et R'_i correspondent respectivement aux fonctions d et e , dans la mesure où elles sont utilisées pour caractériser les situations où l'agent i a agi, et celles où il n'a pas agi. Elles sont définies de la façon suivante.

Soit A l'ensemble des actions réalisées par l'agent i en w . On a :
 wR_iu ssi en u l'agent i a fait A , et éventuellement d'autres choses,
 wR'_iv ssi en v l'agent i n'a pas fait A , mais éventuellement d'autres choses.

D'après la définition de R_i on a toujours wR_iw , donc R_i est réflexive. De plus, si on a wR_iu et uR_iu' , en u' i a fait au moins ce qu'il a fait en u , donc il a fait au moins A , et on a wR_iu' . Donc R_i est aussi transitive.

La relation R'_i est anti-réflexive car on n'a jamais wR'_iw . On impose en plus qu'elle soit sérielle. Cette propriété est très importante car elle exprime que l'agent a la possibilité de ne pas faire A . Par exemple, si i a agi sous la contrainte, il n'a pas la possibilité de ne pas faire A .

Considérons maintenant une propriété p qui est toujours vraie quand i a fait au moins A , et qui est toujours fausse quand i n'a pas fait A (voir figure 4), soit p telle que :

- (1) $\forall u(wR_iu \Rightarrow M, u \models p)$
- (2) $\forall v(wR'_iv \Rightarrow M, v \models \neg p)$

La condition (1) exprime que l'action réalisée A est suffisante pour obtenir la propriété p , et la condition (2) exprime que l'action A est nécessaire pour obtenir p . Ici "nécessaire" est pris dans le sens où i) si on ne fait pas A , en aucun cas on n'obtient p .

Kanger a introduit deux modalités D_i et D'_i qui expriment respectivement la condition (1) et la condition (2). La signification intuitive de $D_i p$ est "ce qu'a fait i est suffisant pour que l'on ait p ", et celle de $D'_i p$ est "ce qu'a fait i est nécessaire pour que l'on ait p ". Les conditions de satisfaisabilité correspondantes

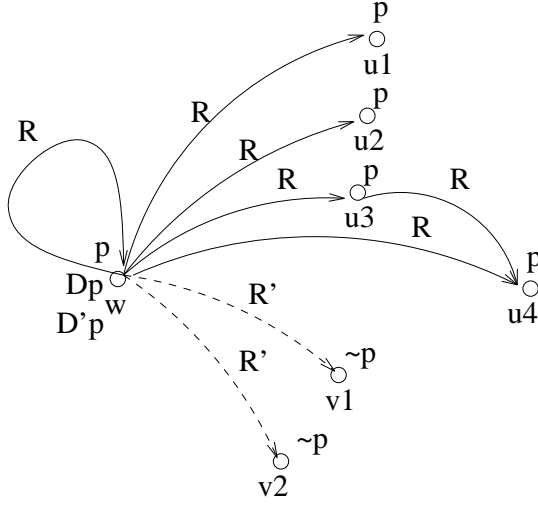


Figure 4: Sémantique de $D_i p$ et de $D'_i p$.

sont :

- $M, w \models D_i p$ ssi pour tout u tel que wRu on a $M, u \models p$.
- $M, w \models D'_i p$ ssi pour tout v tel que wRv on a $M, v \models \neg p$.

L'opérateur d'action KE_i est alors défini par :

$$KE_i p \stackrel{\text{def}}{=} D_i p \wedge D'_i p$$

La signification intuitive de $KE_i p$ est donc que l'action réalisée par i est suffisante et nécessaire pour obtenir p . On peut noter, comme pour les opérateurs d'action définis par von Wright, que cette définition ne fait pas référence à l'action particulière qui a été réalisée,

Andrew J.I. Jones a fait remarquer que selon cette définition $KE_i(p \rightarrow q)$ et $KE_i p$ sont inconsistants. En effet on peut facilement vérifier que $D'_i(p \rightarrow q)$ et $D'_i p$ sont inconsistants, ce qui intuitivement n'est pas acceptable. Cette remarque a conduit Pörn à proposer une autre définition.

Dans cette définition, la condition qui exprime que ce qu'a fait i est nécessaire pour obtenir p est plus faible. Dans sa définition le fait que l'action A est nécessaire pour obtenir p est pris dans le sens où ii) si on ne fait pas A , il est possible que l'on ait $\neg p$. Autrement dit, pour garantir d'avoir toujours p il est nécessaire de faire A . Formellement la condition ii) s'exprime par :

$$(2') \exists v (wR'_i v \text{ et } M, v \models \neg p)$$

En définissant C'_i comme le dual de D'_i , soit $C'_i p \stackrel{\text{def}}{=} \neg D'_i \neg p$, le nouvel opérateur E_i est défini par :

$$E_i p \stackrel{\text{def}}{=} D_i p \wedge C'_i p$$

On peut vérifier que l'on a les schémas d'axiomes suivants :

$$(K) \models E_i(p \rightarrow q) \rightarrow (E_i p \rightarrow E_i q)$$

$$(C) \models E_i p \wedge E_i q \rightarrow E_i p \wedge q$$

D'autre part, comme la relation R'_i est sérielle, la condition $C'_i p$ impose qu'il existe toujours un monde v accessible depuis w par R'_i où p est faux. Donc on n'a jamais $M, w \models C'_i(\text{true})$, ni $M, w \models E_i(\text{true})$ ⁴. On a donc

$$(\neg N) \models \neg E_i(\text{true})$$

Cette propriété correspond tout à fait à l'intuition, elle exprime qu'en aucun cas un agent i ne peut être la cause du fait qu'une tautologie ("true") est vraie. Ceci montre une différence essentielle avec la logique dynamique [5] où pour n'importe quelle action A on a $[A](\text{true})$. En effet en logique dynamique $[A]p$ signifie que p est vraie après que l'action A ait été réalisée, mais on ne peut pas en général en déduire que c'est à cause de A que p est vraie.

On peut facilement vérifier que l'on a $D_i p \rightarrow p$ et donc on a

$$(T) \models E_i p \rightarrow p$$

L'évaluation de la valeur de vérité de $E_i p$ dans un monde w se ramène à l'évaluation de p dans les mondes accessibles par R_i ou R'_i . L'évaluation d'une formule quelconque se réduit donc finalement à l'évaluation des connecteurs logiques \neg et \vee en fonction de la valeur de vérité des formules atomiques. Il s'en suit que si p et q sont logiquement équivalentes, alors $E_i p$ et $E_i q$ le sont aussi. On a donc la règle d'inférence :

$$(RE) \models p \leftrightarrow q \quad \Rightarrow \quad \models E_i p \leftrightarrow E_i q$$

Par contre **on n'a pas** la règle d'inférence $\models p \rightarrow q$ implique $\models E_i p \rightarrow E_i q$. Ceci peut surprendre si on fait le raisonnement suivant. Si $E_i p$ est vraie alors p est vraie, et si on a $\models p \rightarrow q$ alors q est vraie. Mais le fait que q soit vraie n'implique pas que $E_i q$ est vraie. Il se peut que la cause du fait que q est vraie ne soit pas la même que la cause du fait que p est vraie, et que cette cause soit indépendante de ce qu'a fait i .

Par exemple si p signifie "la porte est ouverte" et p' signifie "la terre tourne", on a $\models p \rightarrow (p \vee p')$, mais on n'a pas $\models E_i p \rightarrow E_i(p \vee p')$. On peut d'abord le vérifier formellement. En effet on peut avoir $M, w \models C'_i p \wedge \neg C'_i(p \vee p')$, soit $\exists v(wR'_i v \text{ et } M, v \models \neg p)$ et $\forall v'(wR'_i v' \Rightarrow M, v' \models p \vee p')$, et donc $M, w \models E_i p \wedge \neg E_i(p \vee p')$. En effet, si i ne fait pas l'action A il se peut que dans un certain monde v la porte ne soit pas ouverte, mais même si i ne fait pas l'action A la terre continue de tourner, et donc p' est vraie dans tous les mondes v' , et

⁴On utilise la notation "true" pour désigner n'importe quelle tautologie.

en particulier en v .

Du fait qu'on a $\models \neg E_i(\text{true})$, **on n'a pas** la règle d'inférence de nécessité $\models p$ implique $\models E_i p$. Pour cette raison l'opérateur E_i est un opérateur modal classique mais pas normal, au sens de Chellas [3].

On **n'a pas** non plus le schéma d'axiome (M) $E_i p \wedge q \rightarrow E_i p \wedge E_i q$. En effet, on peut avoir $M, w \models (E_i p \wedge q) \wedge \neg E_i p$, car on peut avoir $M, w \models (C'_i p \wedge q) \wedge \neg C'_i p$, soit $\exists v(wR'_i v \text{ et } M, v \models \neg p \vee \neg q)$ et $\forall v'(wR'_i v' \Rightarrow M, v' \models p)$. En effet, $\forall v'(wR'_i v' \Rightarrow M, v' \models p)$ est consistant avec $\exists v(wR'_i v \text{ et } M, v \models \neg q)$.

Il y a des cas où la définition de $E_i p$ n'est pas satisfaisante, ou pas suffisamment précise. En effet, la définition des mondes v où l'agent i n'a pas fait ce qu'il a fait en w peut poser un problème. En particulier quand d'autres agents sont susceptibles d'agir quand i n'agit pas.

Considérons, par exemple, le cas de deux personnes i et j qui sont au bord de la plage et qui observent un baigneur qui est en train de se noyer. L'agent i décide de se porter à son secours et évite la noyade. Appelons p la proposition qui signifie "le baigneur n'est pas noyé". Intuitivement il paraît acceptable de dire que l'agent i a fait en sorte que le baigneur ne soit pas noyé, soit $E_i p$ est vraie.

Mais si dans ce scénario on suppose que si i ne s'était pas porté à son secours, j ce serait porté à son secours, alors dans tous les mondes v accessibles par R'_i où i n'a pas agi le baigneur n'est pas noyé, c'est-à-dire p est vraie. Donc en w $C'_i p$ est fausse, et formellement $E_i p$ est fausse en w .

On voit avec cet exemple que les mondes v doivent être définis de façon plus précise. Il faut que ce soit des mondes où i n'a pas fait ce qu'il a fait en w , et où les autres agents doivent avoir le même comportement qu'en w . Dans notre exemple ce devrait être des mondes possibles où l'agent j n'aurait pas agi, comme c'est le cas en w (voir [4] pour plus de détails).

La définition de la condition $C'_i p$ nécessite également d'être affinée quand en w l'agent i a fait simultanément plusieurs actions. C'est ce que nous allons voir dans la section suivante.

4 Formalisation proposée par R. Hilpinen

Pour illustrer le problème que pose la "counter action condition" dans le cas où un agent i a réalisé simultanément plusieurs actions on considère l'exemple suivant.

Supposons que l'agent i soit en train de conduire à grande vitesse sur une autoroute. Il voit soudain un panneau qui annonce une sortie dans la direction où il a l'intention d'aller. Pour prendre cette sortie, il tourne tout en freinant, et il en résulte que sa voiture dérape et quitte la chaussée.

Considérons, comme von Wright, des mondes possibles qui correspondent à un état du monde à un instant donné. Soit w le monde dans lequel se trouve i avant qu’il ait commencé à agir. En w il avait la possibilité de réaliser le type d’action tourner (noté A), ou freiner (noté B) ou les deux simultanément (noté AB), ou aucune des deux.

On appelle v le monde possible correspondant à la situation dans laquelle se trouve l’agent une fois qu’il a terminé l’action. Hilpinen appelle le couple $\langle w, v \rangle$ une instance de réalisation de l’action de type AB . Si on appelle p la proposition qui signifie “la voiture est en dehors de la chaussée”, en v p est vraie.

On considère maintenant l’énoncé contre-factuel [14] cf dont la signification est “si i n’avait pas freiné, alors la voiture n’aurait pas quitté la chaussée”. Si le fait de quitter la chaussée est dû au fait de freiner tout en tournant, alors on accepte généralement qu’en v le contre-factuel cf soit vrai. Hilpinen dit que la raison pour laquelle on accepte que cf est vrai en v est fondée sur plusieurs hypothèses qui sont implicitement acceptées.

En appelant v' un monde possible dans lequel on se place quand i n’a pas fait AB , ces hypothèses sont :

1. les opérations réalisées en $\langle w, v' \rangle$ sont différentes de celles réalisées en $\langle w, v \rangle$; ici $\langle w, v' \rangle$ n’est pas une instance de réalisation de AB ,
2. les opérations réalisées en $\langle w, v' \rangle$ sont consistantes avec les hypothèses contre-factuelles; ici $\langle w, v' \rangle$ n’est pas une instance de réalisation de B ,
3. les opérations réalisées en $\langle w, v' \rangle$ “ressemblent le plus possible” aux opérations réalisées en $\langle w, v \rangle$; ici l’agent i n’a pas freiné mais il a tourné.

La troisième hypothèse est essentielle pour donner un sens à l’hypothèse contre-factuelle. C’est parce qu’elle est satisfaite qu’on peut affirmer qu’en v' on a $\neg p$. Si on n’avait pas cette hypothèse on pourrait considérer, par exemple, qu’en $\langle w, v' \rangle$ l’agent i n’a pas freiné, mais qu’il a téléphoné avec son téléphone portable (action de type C) tout en tournant, donc que $\langle w, v' \rangle$ est une instance de réalisation de AC .

Dans ce cas on peut considérer que cette action a fait perdre à i le contrôle de la voiture, et qu’elle a quitté la route. On aurait alors p en v' , et cf serait faux en v .

Hilpinen a proposé un cadre formel [6] qui permet de donner un sens précis à “ $\langle w, v' \rangle$ ressemble le plus possible à $\langle w, v \rangle$ ”. Ce cadre en lui-même ne donne pas la définition de $\langle w, v' \rangle$ mais il permet de l’exprimer clairement dans un contexte donné.

Pour cela Hilpinen définit une fonction $f_i(A, w)$ qui a un type d’action A et un monde w fait correspondre un ensemble de mondes qui résultent de la

réalisation de A , et uniquement de A . On a donc $v \in f_i(A, w)$ ssi $\langle w, v \rangle$ est un instance de A .

Il définit une fonction de même type $g_i(A, w)$ telle que $v \in g_i(A, w)$ ssi $\langle w, v \rangle$ est un instance de réalisation de A , et éventuellement d'autres actions.

Enfin, il définit une fonction $s(A, \langle w, v \rangle)$ qui a une action de type A , et à une instance de réalisation $\langle w, v \rangle$ d'un certain type d'action, fait correspondre un ensemble de mondes v' tels que $\langle w, v' \rangle$ est une instance de réalisation de A qui ressemble le plus possible à $\langle w, v \rangle$.

Dans l'exemple précédent si $\langle w, v \rangle$ est une instance de réalisation de AB , on peut définir les mondes v' où i n'a pas freiné par ⁵:

$$s(omB, \langle w, v \rangle) = f_i(A, w)$$

Ce qui signifie intuitivement que si i n'avait pas freiné la seule chose qu'il aurait faite est de tourner (action A) et non, de tourner tout en faisant autre chose, ce qui serait représenté par $s(omB, \langle w, v \rangle) = g_i(A, w)$, et en particulier pas de tourner tout en téléphonant, ce qui serait représenté par $s(omB, \langle w, v \rangle) = f_i(AC, w)$.

Pour caractériser les situations où il n'aurait pas tourné on pourrait avoir $s(omA, \langle w, v \rangle) = f_i(B, w)$. Les situations qui sont simplement caractérisées par le fait qu'il n'aurait pas freiné tout en tournant pourraient être représentées par $s(omAB, \langle w, v \rangle) = f_i(A, w) \cup f_i(B, w)$.

Dans le cas général les structures proposées par Hilpinen sont des n -uples de la forme $M = \langle W, g_i, s, T \rangle$, où W est un ensemble de mondes possibles définis comme par von Wright, G_i et s sont définies comme ont vient de le voir, et T est défini comme par von Wright.

Nous avons noté ici l'opérateur d'action de Hilpinen HE_i . De façon intuitive $HE_i p$ est vraie en w si i) w résulte de la réalisation d'une certaine action de type A à partir d'un monde initial u , i.e. $\langle u, w \rangle$ est une instance de réalisation de A , et la réalisation de A , et éventuellement d'autres actions, conduit à des mondes où on a toujours p , et ii) si i n' avait pas fait A mais avait fait ce qui s'en rapproche le plus, alors il aurait pu arriver que l'on n'ait pas p (voit figure 5). Formellement la condition de satisfaisabilité de $HE_i p$ est :

$$M, w \models HE_i p \text{ ssi il existe un monde } u \text{ et un type d'operation } A \text{ tels que}$$

- i) $w \in g_i(A, u)$ et $g_i(A, u) \subseteq T(p)$, et
- ii) $s(omA, \langle u, w \rangle) \not\subseteq T(p)$.

On peut remarquer que la fonction g_i permet d'exprimer que ce qu'a fait i est suffisant pour obtenir p , elle joue donc un rôle similaire à R_i chez Pörn. La fonction s permet d'exprimer que ce qu'a fait i est nécessaire (au sens de $C'_i p$),

⁵Hilpinen utilise l'opérateur om dans le même sens que von Wright.

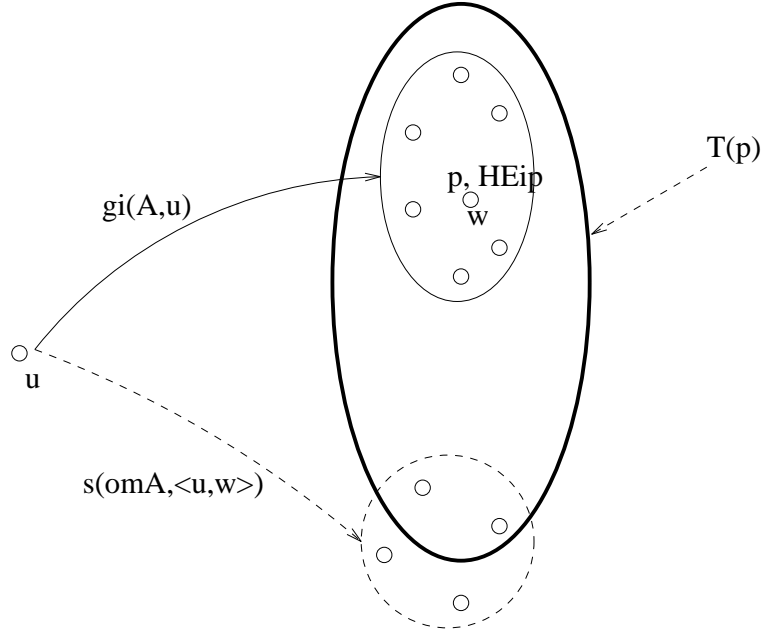


Figure 5: Opérateur HE_i .

et elle joue le même rôle que R'_i . Ce qu'apporte en plus la fonction s c'est de donner un sens précis à la "counter action condition".

Hilpinen a proposé un autre opérateur d'action, noté ici $HE_i^1 p$ dans lequel la condition i) est remplacée par : $w \in g_i(A, u) \cap T(p)$. Cette condition exprime qu'en w on a p mais que le fait de réaliser A ne garantit pas toujours d'obtenir p . La distinction entre $HE_i p$ et $HE_i^1 p$ peut servir de base pour distinguer le cas où i a la capacité de faire ce qu'il a fait, avec $HE_i p$, et le cas où il a fait une action qui a conduit à p , mais où il l'a fait sans en avoir la capacité, c'est-à-dire qu'en refaisant le même type d'action on aurait pu avoir $\neg p$. Ceci correspond, par exemple, au cas d'un agent qui a lancé une fléchette au centre d'une cible, mais qui en refaisant le même type d'action aurait pu l'envoyer en dehors du centre.

5 Conclusions

Les propositions de formalisation que nous avons vues montrent qu'il n'y a pas une définition unique de la causalité. Néanmoins le choix de telle ou telle définition dans un contexte donné permet d'éviter les malentendus, et de définir les conséquences que l'on peut tirer d'un ensemble d'hypothèses décrivant une situation donnée quand on a fait le choix d'une définition (voir, par exemple,

[13]).

Il y a de nombreux aspects que, faute de place, nous n'avons pas abordés ici. Par exemple, la notion de choix délibéré. C'est-à-dire le fait qu'à un moment donné un agent a le choix de réaliser tel ou tel type d'action (voir [8, 11]). Par exemple, quand une personne a cassé une vitre en lançant un caillou, le caillou n'avait pas le choix de partir dans une autre direction, et s'il a cassé la vitre ce n'est pas le résultat de son choix. Donc, de ce point de vue, on dirait que ce n'est pas le caillou qui a cassé la vitre.

Nous n'avons pas non plus considéré les aspects temporels. Par exemple, le fait qu'une action est en train d'être réalisée, mais qu'elle n'est pas terminée, pose des problèmes qui, à notre connaissance, n'ont pas fait l'objet d'essais de formalisation. En effet, ce qu'a fait un agent à un instant donné peut parfois être interprété comme le début de réalisation d'une action de type A ou d'une action de type B . A quelles conditions peut-on dire qu'il est en train de réaliser l'une ou l'autre? Quelquefois un événement qui a eu lieu à l'instant initial permet de trancher. Par exemple, un chauffeur de poids lourd est parti de Paris. Lorsqu'il est à Lyon peut-on dire qu'il est en train de transporter sa marchandise à Milan ou à Marseille? La réponse peut être donnée en fonction de l'ordre qui lui a été donné, ou de son intention, au départ de Paris.

Nous avons à peine évoqué le lien entre causalité et compétence, ou capacité (voir [1, 12]). En fait les deux notions sont très liées, comme le mentionnent Belnap et Horty dans [8, 7]. En effet, si un agent a lancé un dé qui est tombé sur le 6, on ne peut pas dire qu'il est la cause du résultat dans le même sens qu'il est la cause du fait que la porte est ouverte. Nous n'avons pas non plus présenté les liens entre causalité et obligation alors que, comme nous l'avons dit dans l'introduction, beaucoup de travaux sur la causalité ont eu comme motivation de donner un sens précis au fait qu'un agent doit faire quelque chose (voir [9, 8]).

Enfin, nous n'avons pas présenté un aspect qui est important dans les systèmes multi-agents, c'est la notion d'action collective (voir [2]). Que veut-on dire exactement quand on dit, par exemple, que l'équipe de football de Madrid a marqué 3 buts, ou bien qu'elle doit gagner le match contre Barcelone? Les rapports entre actions collectives et actions individuelles sont au coeur de nombreux problèmes quand on veut définir l'organisation des interactions entre agents et groupes d'agents.

References

- [1] M. Brown. On the logic of ability. *Journal of Philosophical Logic*, 17:1–26, 1988.
- [2] J. Carmo and O. Pacheco. Deontic and action logics for collective agency and roles. In R. Demolombe and R. Hilpinen, editors, *Proceedings of the 5th*

- International workshop on Deontic Logic in Computer Science*. ONERA, 2000.
- [3] B. F. Chellas. *Modal Logic: An introduction*. Cambridge University Press, 1988.
 - [4] D. Elgesem. *Action Theory and Modal Logic*. PhD thesis, University of Oslo, Department of Philosophy, 1992.
 - [5] D. Harel. Dynamic logic. In D. Gabbay and F. Guenther, editors, *Handbook of Philosophical Logic*, volume 2. Reidel, 1984.
 - [6] R. Hilpinen. On Action and Agency. In E. Ejerhed and S. Lindstrom, editors, *Logic, Action and Cognition*. Kluwer, 1997.
 - [7] J. Horty. *Action and deontic logic*. Oxford University Press, 2000.
 - [8] J.F. Horty and N. Belnap. The deliberative STIT: a study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24:583–644, 1995.
 - [9] A.J.I. Jones and M. Sergot. On the role of Deontic Logic in the characterization of normative systems. In J.-J. Meyer and R.J. Wieringa, editors, *Proc. First Int. Workshop on Deontic Logic in Computer Science*, 1991.
 - [10] S. Kanger. Law and Logic. *Theoria*, 38:105–132, 1972.
 - [11] A. Kenny. *Will, Freedom and Power*. Basil Blackwell, 1975.
 - [12] A. Kenny. Human abilities and dynamic modalities. In J. Manninen and R. Tuomela, editors, *Essays on explanation and understanding: studies in the foundations of humanities and social sciences*. D. Reidel Publishing Company, 1976.
 - [13] C. Krogh. On the role of action logics and deontic logics in specifying protocols. In J. Bell and Z. Huang, editors, *Proc. of the workshop on Practical Reasoning and Rationality (IJCAI'99)*. University of London, 1999.
 - [14] D. Lewis. *Counterfactuals*. Harvard University Press, 1973.
 - [15] I. Porn. Action Theory and Social Science. Some Formal Models. *Synthese Library*, 120, 1977.
 - [16] F.A. Santos, A.J.I. Jones, and J.M. Carmo. Action Concepts for Describing Organised Interaction. In *Proc. 30th Hawaii International Conference on System Sciences*, 1997.
 - [17] K. Segerberg. Getting started: beginnings in the logic of action. *Studia Logica*, 51:347–378, 1992.
 - [18] G. H. von Wright. *Norm and Action*. Routledge and Kegan, 1963.