Modéliser les interactions entre agents : un prérequis pour analyser l'éthique des systèmes complexes

Robert Demolombe

Institut de Recherche en Informatique de Toulouse (collaborateur extérieur)
robert.demolombe@orange.fr

RÉSUMÉ. *Les règles d'éthique appliquées aux systèmes complexes, tels que les systèmes d'armes faisant intervenir des drones, nécessitent une analyse détaillée des interactions entre les agents. Celles-ci font intervenir des agents institutionnels, humains, logiciels et matériels. Leurs interactions font intervenir la causalité, l'influence et les relations entre agents humains et institutionnels. En partant d'exemples, ces notions sont analysées informellement puis formalisées en logique modale. Cette formalisation met en évidence des règles de raisonnement différentes qui sont souvent confondues : le raisonnement sur la conséquence logique, le raisonnement sur la causalité et sur l'influence, et le raisonnement sur les relations entre actions des agents institutionnels et actions des agents qui agissent en leur nom. Cette formalisation peut aider à concevoir des règles d'éthique cohérentes.*

ABSTRACT. *Ethical rules applied to complex systems, such as weapons systems involving drones, require a detailed analysis of the interactions between the agents. These rules involve institutional, human, software and hardware agents. Their interactions involve causality, influence and relationships between human tutional. Starting from examples, these notions are analyzed informally then formalized in modal logic. This formalization highlights different rules of reasoning which are often confused: the reasoning on the logical consequence, the reasoning on causality and influence, and the reasoning on the relations between institutional agents' actions and actions of agents acting on their behalf. This formalization can help to design consistent ethical rules.*

MOTS-CLÉS : *agents, causalité, influence, éthique, logique modale.*

KEYWORDS: *agents, causality, influence, ethics, modal logic.*

DOI:10.3166/RIA.32.683-703

1. Introduction

L'éthique est une notion dont la définition demande à être précisée. En particulier, la différence entre morale et éthique n'est pas toujours claire parce qu'à l'origine on leur donnait le même sens. Mais, au cours du temps, on leur a donné des sens

différents qui peuvent être exprimés de la façon suivante : la morale définit ce qui est bien et l'éthique définit comment il faut se comporter pour satisfaire ce qui est bien, c'est-à-dire ce qu'il faudrait faire et ce qu'il ne faudrait pas faire.

Si on accepte cette distinction on voit que l'éthique fait nécessairement référence aux notions d'agent et aux actions qu'ils réalisent. C'est la raison pour laquelle le but de cet article est de proposer une analyse approfondie des différents types d'agents et des propriétés des actions auxquelles on se réfère pour définir une éthique, en particulier l'influence et la causalité.

De plus, l'éthique n'est pas définie en général, mais pour certains types d'activités. Ici, nous nous intéressons, en particulier, à l'éthique concernant la mise en oeuvre de systèmes complexes tels que des systèmes d'armes qui font intervenir des drones. Il ne s'agit pas de définir les règles d'éthique qui devraient s'appliquer dans ce domaine, mais d'aider à la définition de ces règles en précisant la structure des agents qui interviennent et la nature de leurs interactions.

Dans la section 2 nous donnons une présentation semi-formelle des agents qui interagissent entre eux, puis dans la section 3 des définitions formelles, en logique modale, des concepts les plus importants, en les commentant intuitivement pour que des lecteurs non familiers avec ce formalisme puissent en comprendre l'essentiel. De plus, nous donnons un sens précis au fait que l'action de tel agent a causé tel effet ou a influencé tel effet, ainsi qu'au fait que, vis-à-vis des normes d'une institution, l'action de tel agent compte comme une action de cette institution.

2. Analyse semi-formelle

Les agents qui interviennent dans la mise en oeuvre de systèmes d'armes complexes peuvent être regroupés en trois types : les agents institutionnels (par exemple : un certain corps d'armée), les agents humains (par exemple : un lieutenant ou un commandant) et les agents artificiels. Parmi ceux-ci nous distinguerons les agents logiciels (par exemple : un logiciel utilisé pour commander un drone) et les agents matériels qui peuvent réaliser des actions physiques [1] (par exemple : un drone).

En faisant référence à la figure 1, nous décomposerons un système complexe de la façon suivante. L'agent H_1, qui est un commandant, réalise l'action de donner l'ordre à l'agent H_2, qui est un lieutenant, d'activer le drone désigné comme l'agent M_2.

On note a le type d'action réalisé par H_1 et on note $H_1 : a$ la réalisation d'une action de type a par H_1; on appellera par la suite "actes" les couples tels que $H_1 : a$.

1. Nous utilisons cette terminologie pour simplifier car il est bien clair que les agents logiciels sont aussi matériels. Mais la différence avec ceux que nous appellerons par la suite "agents matériels" est que la fonction des agents logiciels est de traiter de l'information.

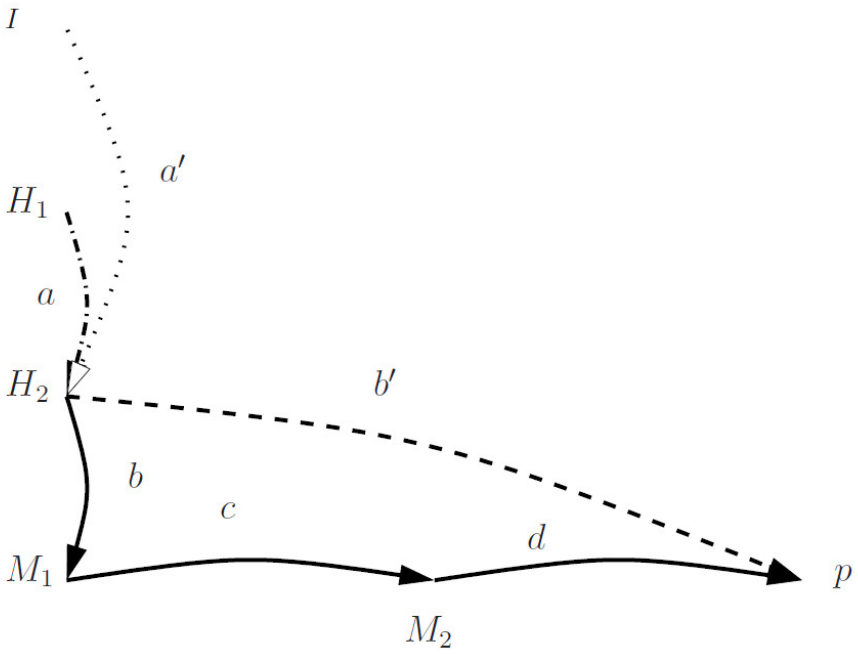

Figure 1. Actions des agents institutionnels humains et artificiels

L'agent H_2 réalise une action de type b qui consiste à appuyer sur un bouton qui a pour effet d'activer l'agent logiciel M_1. L'agent logiciel M_1 réalise une action de type c qui consiste à activer le drone M_2. Le drone M_2 réalise une action de type d qui a pour effet, par exemple, la mort de certaines personnes ou la destruction d'un hôpital [2].

Au regard des règles qui régissent l'institution du Corps d'Armée I l'action réalisée par H_1, à condition qu'elle soit réalisée conformément à ses règles, [3] "compte comme" une action a' réalisée par I. On dira alors que, pour l'institution I, l'acte $H_1 : a$ compte comme l'acte $I : a'$.

Une question essentielle du point de vue de la définition de l'éthique d'un tel système est de préciser quelle est la part de libre arbitre de chacun des agents.

Pour cela on peut distinguer les actes qui causent des effets, dans le sens où chaque réalisation d'un acte de ce type par le même agent, et dans le même contexte, suffit

2. Pour simplifier l'écriture on utilisera parfois des expressions telles que : "l'acte $H_2 : b$ qui consiste à appuyer sur un bouton" au lieu de : "l'acte $H_2 : b$ qui consiste à ce que l'agent H_2 appuie sur le bouton".

3. Habituellement ces règles définissent des "rôles", c'est à dire des ensembles d'obligations et de pouvoirs, les agents qui sont titulaires de ces rôles, et les circonstances dans lesquelles ils exercent leurs rôles. Voir (Cuppens, 2015 ; J. Gelati, Sartor, 2002 ; Pacheco, Santos, 2004 ; Demolombe, Louis, 2006).

à produire le même effet. Par exemple, l'acte $H_2 : b$, qui consiste à appuyer sur un certain bouton, est la cause du fait que l'agent M_1 réalise l'acte $M_1 : c$.

Mais il y a aussi des actes pour lesquels l'effet peut être parfois obtenu mais ne l'est pas toujours. Par exemple, l'acte $H_1 : a$ a toujours pour effet de créer l'obligation pour H_2 de réaliser $H_2 : b$, mais il n'a pas toujours pour effet que H_2 réalise $H_2 : b$. Dans ce cas on dira que l'acte $H_1 : a$ influence l'acte $H_2 : b$. En effet H_2, dans la mesure où il dispose d'un libre arbitre, peut choisir, ou non, d'exécuter l'ordre qui lui a été donné.

L'action d'un agent peut influencer l'action d'un autre agent d'un grand nombre de façons différentes. Par exemple, l'action du premier peut modifier l'intention du second en lui faisant savoir (ou croire) que s'il réalise cette action il y trouvera un certain avantage, ou que cela lui évitera un désavantage s'il ne la réalise pas. Une autre forme d'influence peut s'exercer quand l'action du premier rend matériellement possible, ou impossible, telle action du second. Dans ce contexte, par exemple, si H_2 ne peut activer M_1 que s'il connaît un certain code, H_1 peut influencer H_2 en lui transmettant ce code.

Pour rappel nous utilisons pour les actions les notations :

– action a : donner l'ordre d'activer le drone,

– action b : appuyer sur un bouton,

– action c : activer le drone,

– action d : tuer une personne (ou détruire un hôpital),

– action a' : donner l'ordre d'activer le drone (action réalisée par l'agent institutionnel).

Pour préciser un peu plus les interactions nous adoptons les notations suivantes :

– proposition q : il est obligatoire que H_2 réalise l'acte $H_2 : b$,

– proposition q' : l'agent H_2 connaît le mot de passe qui permet d'activer M_1,

– proposition r : l'agent M_1 a reçu le signal qui déclenche l'acte $M_1 : c$,

– proposition s : l'agent M_2 a reçu le signal qui déclenche l'acte $M_2 : d$,

– proposition p : la personne X est morte, ou bien, l'hôpital Y a été détruit.

Les différentes interactions entre agents peuvent alors être décrites de la façon suivante :

– l'acte $H_1 : a$ **compte comme** l'acte $I : a'$ [4],

– l'acte $H_1 : a$ **cause** le fait q, ou, dans d'autres circonstances, l'acte $H_1 : a$ **cause** l'effet q',

4. La distinction entre ces deux actes permet de distinguer la responsabilité d'une personne physique de la responsabilité d'une personne morale.

– le fait que q est vraie **influence** l'acte $H_2 : b$, ou bien le fait que q' est vraie **influence** l'acte $H_2 : b$,

– l'acte $H_2 : b$ **cause** la réalisation de l'acte $M_1 : c$,

– l'acte $M_1 : c$ **cause** la réalisation de l'acte $M_2 : d$,

– l'acte $M_2 : d$ **cause** le fait que p est vraie.

Nous avons distingué explicitement les types d'actions et les instances d'action. Par exemple, le type d'action b : "appuyer sur un certain bouton", désigne n'importe quelle instance d'action qui a pour effet que le bouton est appuyé. Donc le type d'action b ne distingue pas, par exemple, appuyer vite sur le bouton, ou appuyer lentement sur le bouton, ou appuyer sur le bouton avec la main droite, ou appuyer sur le bouton avec la main gauche, ... etc ...

Si le drone M_2 est doté de fonctions dites d'Intelligence Artificielle (voir le rapport numéro 1 de la CERNA page 27 : "les robots dans la défense et la sécurité" ou (ETHICAA, 2015 ; Bonnemains *et al*., 2016 ; Cointe *et al*., 2016)), par exemple, la capacité d'analyser des images afin d'identifier une cible, et de "choisir" de déclencher l'attaque de cette cible après l'avoir identifiée, sans intervention de l'agent humain H_2, alors il se peut que les règles d'éthique s'appliquent aux différents intervenants qui ont conduit à la réalisation et à la mise à disposition de M_2.

En effet, le terme "choisir" dans ce type de situation correspond à un usage courant mais peut prêter à confusion car, même si le drone n'effectue pas toujours les mêmes actions, selon qu'il est dans telle ou telle situation, ce sont les concepteurs du drone qui ont réellement choisi qu'il réalise tel type d'action plutôt que tel autre. De plus, il se peut que, si le drone utilise des fonctions complexes du type Intelligence Artificielle, les conséquences des différents choix faits par les concepteurs soient difficiles à percevoir, et que ceci donne l'illusion que c'est le drone qui fait les choix.

Sur la figure 2 nous avons représenté, sans rentrer dans les détails, ces intervenants. Les différents types de flèches permettent de distinguer les différents types d'actions. Par exemple $H_1 - H_2$ et $C - R$ représentent deux formes d'influence, $H_2 - M_1$ représente la relation directe de causalité, $H_2 - p$ représente la relation indirecte de causalité et $I - H_2$ représente un acte institutionnel.

Il y a tout d'abord l'agent C qui conçoit le système M_2. On appelle e cette activité de conception et le résultat de l'acte $C : e$ est représenté par la proposition t_1 qui signifie que les spécifications de M_2 sont entièrement définies. L'agent C peut être, par exemple, un bureau d'études.

Il y a ensuite l'agent R qui réalise M_2 conformément à ces spécifications. On appelle f cette activité et le résultat de l'acte $R : f$ est représenté par la proposition t_2 qui signifie que M_2 a été réalisé. L'agent R peut être, par exemple, une entreprise de fabrication.

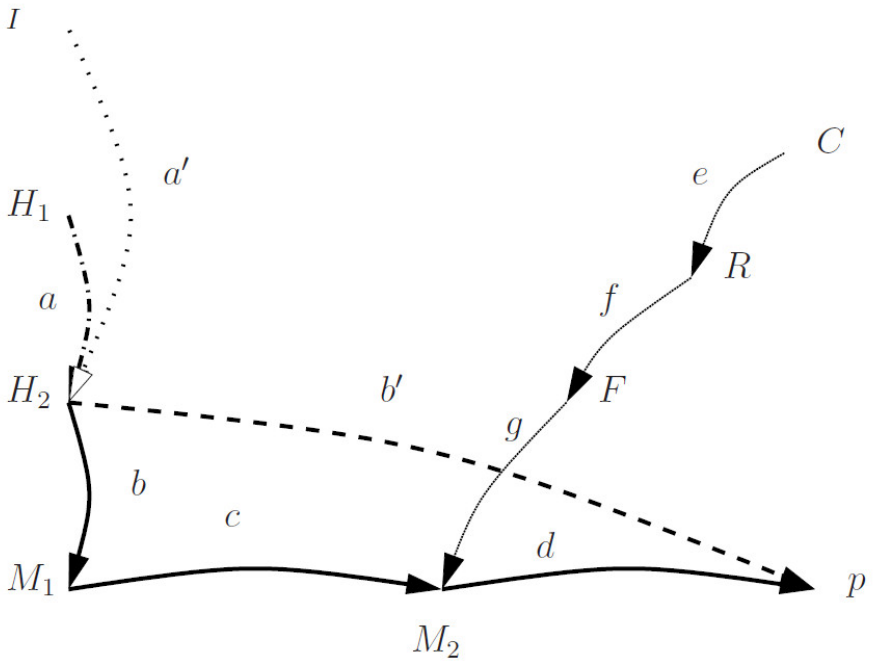

Figure 2. Actions des agents institutionnels, humains et artificiels

Il y a enfin l'agent F qui met le système M_2 à la disposition de l'agent humain H_2. On appelle g cette activité et le résultat de l'acte $F : g$ est que l'agent humain H_2 a la possibilité de déclencher effectivement l'activité de M_2. Le résultat de $F : g$ est représenté par la proposition t_3 qui signifie que H_2 peut activer M_2. L'agent F peut être une entreprise qui joue le rôle de fournisseur vis-à-vis de l'agent l'institutionnel I.

Si on se concentre sur ce sous-ensemble du système, les interactions entre les agents peuvent être décrites comme suit :

– l'acte $C : e$ **influence** l'acte $R : f$ (en effet il n'y a pas de relation de causalité car R peut choisir d'utiliser, ou non, les spécifications et de réaliser M_2),

– l'acte $R : f$ **influence** l'acte $F : g$ (il n'y a pas de relation de causalité pour la même raison que précédemment),

– l'acte $F : g$ **influence** l'acte $H_2 : b$ car $H_2 : b$ ne peut être réalisé que si M_2 a été mis à disposition de H_2 (il n'y a pas de relation de causalité pour les mêmes raisons que précédemment).

On pourrait entrer un peu plus dans les détails dans la mesure où les agents C, R et F sont en fait des agents institutionnels et qu'il y a des agents humains dont les actes

comptent comme des actes de ces agents institutionnels. Ici aussi les règles d'éthiques pourraient s'appliquer aux agents institutionnels ou aux agents humains.

Par exemple, une règle pourrait spécifier qu'un agent humain peut refuser de participer à la conception, à la fabrication ou à la vente d'un drone M_2 qui aurait la capacité de réaliser certaines actions. Une autre règle concernant les agents institutionnels pourrait être qu'un agent institutionnel ne doit pas sanctionner un agent humain, qui agit pour le compte de cet agent institutionnel, et qui refuse d'obéir à certains ordres qui le conduirait à ne pas respecter une règle d'éthique qui s'applique aux agents humains.

3. Formalisation

Dans ce qui suit nous allons voir comment les principales notions que nous avons mises en évidence sur ces exemples peuvent être formalisées en logique modale. On peut trouver une bonne introduction aux logiques modales dans (Chellas, 1988). D'autre part il y a dans la littérature de nombreuses définitions de la causalité dont on peut trouver une bonne synthèse dans (Hilpinen, 1997). Ici nous avons adopté une définition qui est proche de celle retenue par Hilpinen dans (Hilpinen, 1997).

3.1. Causalité

3.1.1. Langage

Pour définir le langage formel que nous allons utiliser nous adoptons les notations suivantes.

– AGT : ensemble d'agents; on notera les agents : i, j, k, ...,

– $ACTION$: ensemble de types d'actions; on notera les types d'actions : a, b, c, ...,

– ACT : ensemble d'actes; ce sont des couples de la forme : $i : a$, où i est un agent dans AGT et a un type d'action dans $ACTION$,

– $Br_{ACT,ACT'}(\phi)$: opérateur qui signifie que la réalisation de l'acte ACT', qui fait partie de la réalisation de l'ensemble d'actes ACT, est la cause du fait que ϕ est vraie, où ϕ est une formule du langage L défini plus bas. La définition de cet opérateur fait référence à l'ensemble d'actes ACT car le plus souvent un agent n'est pas seul à agir et, selon que ACT' est réalisé simultanément avec tels ou tels actes, il se peut qu'il ne puisse pas causer ϕ ou bien que ce ne soit pas lui qui soit la cause de ϕ[5].

Le langage L est défini de la façon suivante :

5. La notation de l'opérateur Br est une abréviation de l'expression : "*to Bring it about that*", que l'on traduit par : "*faire en sorte que*".

$$\phi ::= p \mid \neg\phi \mid \phi \vee \phi \mid Br_{ACT,ACT'}(\phi)$$

où p est une formule atomique du calcul des propositions.

Le langage est étendu de façon habituelle avec les opérateurs : $\phi \rightarrow \phi$ et $\phi \wedge \phi$. Voici quelques exemples de formules de L avec leurs significations intuitives :

– $Br_{\{i:a,j:b\},j:b}(p)$: l'agent j fait en sorte que p soit vraie en réalisant une action de type b alors qu'est réalisé simultanément l'acte $i : a$, mais c'est $j : b$ qui est la cause du fait que p soit vraie. Ici, on a $ACT = \{i : a, j : b\}$ et $ACT' = j : b$ [6],

– $Br_{\{i:a,j:b\},j:b}(p \vee q)$: l'agent j fait en sorte que p ou q soit vraie en réalisant une action de type b alors qu'est réalisé simultanément l'acte $i : a$,

– $Br_{\{i:a,j:b\},j:b}(Br_{\{j:b,k:c\},k:c}(\neg p))$: l'agent j, en réalisant une action de type b, fait en sorte que l'agent k, en réalisant une action de type c, fasse en sorte que p soit fausse.

3.1.2. Sémantique

Le sens précis des formules du langage L est défini dans le cadre formel des modèles de Kripke.

Ces modèles sont définis par un ensemble de mondes possibles W. Chaque monde possible représente un état du monde dans sa globalité qui est défini par la valeur de vérité des formules atomiques de L. Ces valeurs de vérités sont définies, d'une part par une fonction V qui assigne à chaque formule atomique propositionnelle l'ensemble des mondes dans lesquels cette formule est vraie et, d'autre part, par un ensemble R de relations ternaires définies sur des triplets de mondes de $W \times W \times W$. À chaque couple $< ACT, ACT' >$ correspond une relation de R que l'on note $R_{ACT,ACT'}$. Ce sont ces relations qui permettent de définir le sens qu'on donne aux opérateurs modaux et les valeurs de vérité des formules qui contiennent ces opérateurs modaux. On les appelle "relations d'accessibilité".

Formellement un modèle M est un triplet $< W, V, R >$.

La signification intuitive du fait que le triplet de mondes $< w, w'_1, w''_1 >$ (voir figure 3) appartient à $R_{ACT,ACT'}$ est que la réalisation de l'ensemble d'actes ACT a commencé en w et s'est terminée en w'_1, et qu'en w''_1 ont été réalisés tous les actes de ACT sauf ACT', et que la seule différence entre w'_1 et w''_1 est qu'en w''_1 ACT' n'a pas été réalisé. On dit que w''_1 est un monde contrefactuel de w'_1.

A w peuvent correspondre d'autres mondes tels que w'_1 qui correspondent à d'autres instances de réalisation du type d'actes ACT', par exemple w'_2. Le monde contrefac-

6. Pour simplifier les notations lorsque ACT' est un ensemble qui contient un seul élément, par exemple $\{j : b\}$, on notera cet ensemble comme le seul élément qu'il contient, par exemple $j : b$.

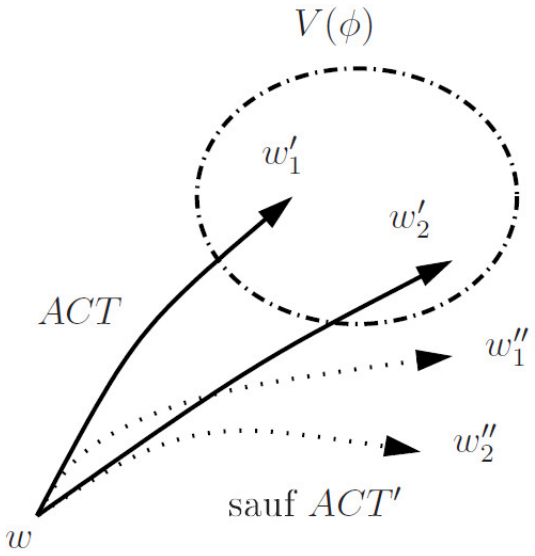

Figure 3. La réalisation de l'acte ACT' cause ϕ

tuel w''_2 de w'_2 n'est pas le même que le monde contrefactuel w''_1 de w'_1, car w'_1 et w'_2 correspondent à différentes instances de réalisation des actes de ACT.

La valeur de vérité de chaque formule ϕ de L est définie dans chaque monde w de chaque modèle M par des conditions appelées "conditions de satisfaisabilité". Le fait que ϕ est vraie pour le couple M, w est noté :

$$M, w \models \phi$$

Si p est une formule atomique de L, on a : $M, w \models p$ si, et seulement si, w appartient à $V(p)$

Pour les négations et disjonctions, on a :

– $M, w \models \neg\phi$ si, et seulement si, il est faux que $M, w \models \phi$
– $M, w \models \phi \vee \psi$ si, et seulement si, $M, w \models \phi$ ou $M, w \models \psi$

Intuitivement, une formule de la forme $Br_{ACT,ACT'}(\phi)$ est vraie en w si, et seulement si, il suffit de réaliser ACT' dans le contexte de ACT pour que ϕ soit vraie, et qu'il est nécessaire de réaliser ACT' pour que ϕ soit vraie. Cette deuxième condition

peut se reformuler ainsi : si ACT' n'avait pas été réalisé, "toutes choses égales par ailleurs" [7], ϕ aurait été fausse.

Dans le formalisme que nous avons défini ces conditions s'expriment de la façon suivante [8] :

$M, w \models Br_{ACT,ACT'}(\phi)$ si, et seulement si, :

1. pour tous mondes w'et w'' tels que $R_{ACT,ACT'}(w, w', w'')$ on a $M, w' \models \phi$ et,
2. pour tous mondes u'et u'' tels que $R_{ACT,ACT'}(w, u', u'')$ on a $M, u'' \models \neg\phi$

L'intérêt de cette définition précise est d'éviter des erreurs de raisonnement qui paraissent intuitives. En particulier, il peut paraître intuitif que de $Br_{ACT,ACT'}(\phi)$ on puisse déduire $Br_{ACT,ACT'}(\phi \vee \psi)$, en donnant comme argument que si dans tous les mondes tels que w' on a ϕ qui est vraie, alors dans tous ces mondes $\phi \vee \psi$ est vraie. Cependant, du fait que ϕ est fausse dans tous les mondes contrefactuels w'' on ne peut pas déduire que $\phi \vee \psi$ est fausse dans tous ces mondes contrefactuels. Un cas extrême est celui où ψ est $\neg\phi$ car dans ce cas $\phi \vee \psi$ est une tautologie. Cet exemple montre qu'il est parfois tentant de confondre la relation de causalité avec la relation de conséquence logique.

On peut illustrer le paragraphe précédent en prenant comme exemple le fait que l'agent H_2 a fait en sorte que le bouton soit appuyé. On ne peut pas en déduire que H_2 a fait en sorte que le bouton soit appuyé ou que la terre tourne, car si H_2 n'avait pas réalisé une action de type b, le bouton n'aurait pas été enfoncé mais cela n'aurait pas empêché la terre de tourner.

D'une façon plus générale, il n'est pas vrai que si ϕ implique logiquement ψ, alors $Br_{ACT,ACT'}(\phi)$ implique $Br_{ACT,ACT'}(\psi)$.

En particulier il n'est pas vrai que de $Br_{ACT,ACT'}(\phi \wedge \psi)$ on puisse déduire $Br_{ACT,ACT'}(\phi)$. Ceci peut paraître paradoxal quand on raisonne sur l'éthique. Supposons, par exemple, que ACT' soit un acte réalisé par un drone qui a pour effet qu'un pont est détruit (proposition p) et aussi qu'un hôpital est détruit (proposition q).

Si une règle d'éthique dit qu'il est interdit de détruire un hôpital, comme on ne peut pas déduire, d'après ce qui précède, que l'acte ACT' est la cause du fait que l'hôpital est détruit, on pourrait en déduire que l'interdiction a été respectée. Mais ce paradoxe disparaît si on accepte que, d'une façon générale, s'il est interdit de faire en sorte que p soit vraie, alors il est aussi interdit de faire en sorte que n'importe quelle

7. La définition précise de cette expression, appelée "contrefactuelle", (ou condition "*ceteris paribus*") pose de grosses difficultés (voir (Lewis, 1973)). En première approximation on peut dire que les mondes contrefactuels sont des mondes imaginaires où l'action n'a pas été réalisée et qui sont le plus semblable possible au monde dans lequel on est après avoir réalisé l'action.

8. Cette définition présente de façon moins formelle celle donnée dans (Demolombe, 2012) pour le cas où ACT' est une ensemble d'actes.

proposition qui implique logiquement p, telle que $p \wedge q$, soit vraie. Donc, il est aussi interdit de faire en sorte qu'un pont et un hôpital soient détruits.

Cet exemple montre la distinction qu'il faut faire entre le **raisonnement sur les interdictions** [9] (ici si ϕ implique logiquement ψ et ψ est interdit, alors ϕ est interdit), le **raisonnement sur les conséquences logiques** (ici ϕ implique logiquement ψ), et le **raisonnement sur la causalité** (ici si ACT' est la cause de ϕ on ne peut pas en déduire que ACT' est la cause de ψ).

La question de définir une éthique quand c'est l'action simultanée de plusieurs agents qui est la cause d'un certain effet est plus compliquée à analyser. C'est la raison pour laquelle nous avons étendu la définition de l'opérateur $Br_{ACT,ACT'}(\phi)$ dans le cas où ACT' désigne un ensemble d'actes, et non plus un seul acte.

De façon informelle on peut dire dans ce cas que ACT' est la cause de ϕ si, et seulement si, les 3 conditions suivantes sont satisfaites :

1. Il suffit que tous les actes de ACT soient réalisés pour que ϕ soit vraie,
2. Si l'un des actes de ACT' n'est pas réalisé, alors ϕ peut être fausse,
3. Pour tout acte $i : a$ qui est dans ACT, mais pas dans ACT', il suffit que tous les actes de ACT, sauf $i : a$, soient réalisés pour que ϕ soit vraie.

La conséquence de la condition 1) est que les actes qui sont dans ACT, mais pas dans ACT', n'empêchent pas que l'on ait ϕ vraie.

La conséquence de la condition 2) est que la réalisation de tous les actes de ACT' est nécessaire pour que ϕ soit vraie.

La conséquence de la condition 3) est que les actes qui sont dans ACT mais pas dans ACT' ne sont pas nécessaires pour que ϕ soit vraie.

Enfin, l'ensemble de ces trois conditions a pour conséquence que ce sont les actes qui sont dans ACT', et eux seuls, qui sont la cause du fait que ϕ est vraie [10].

Plus formellement (voir une définition plus précise dans (Demolombe, 2012)), dans le cas où ACT' désigne un ensemble d'actes, on a $M, w \models Br_{ACT,ACT'}(\phi)$ si, et seulement si, :

1. pour tous mondes w'et w'' tels que $R_{ACT,ACT'}(w, w', w'')$ on a $M, w' \models \phi$,
2. pour tout acte $i : a$ dans ACT', pour tous mondes u'et u'' tels que $R_{ACT,ACT'-i:a}(w, u', u'')$ on a $M, u'' \models \neg\phi$,
3. pour tout acte $j : b$ qui est dans ACT mais pas dans ACT', pour tous mondes v'et v'' tels que $R_{ACT-j:b,ACT'}(w, v', v'')$ on a $M,' v \models \phi$.

---

9. On peut, en général, représenter à la fois les obligations et interdictions avec la même modalité : l'obligation. En effet, l'interdiction de faire en sorte que ϕ peut s'exprimer comme l'obligation de ne pas faire en sorte que ϕ. Par ailleurs la permission peut être implicite, car on accepte généralement que tout ce qui n'est pas interdit est permis.

10. Voir dans (Demolombe, 2012) la preuve du Théorème 1.

On pourrait prendre comme exemple de la réalisation simultanée de plusieurs actes qui sont la cause d'un certain effet le cas où, dans l'analyse de la section précédente, il faut, pour déclencher l'action de l'agent logiciel M_1, que deux agents humains tels que H_2 appuient simultanément sur deux boutons. En désignant par $H_2 : b$ et $H'_2 : b'$ ces deux actes, on aurait en utilisant les notations précédentes : $ACT' = \{H_2 : b, H'_2 : b'\}$.

Il est intéressant de considérer aussi, du point de vue de la responsabilité et de l'éthique, le cas où ces agents agissent simultanément, mais un seul de ces actes suffit pour obtenir l'effet que l'action de M_1 soit déclenchée. Dans ce cas il faut modifier la condition 2) qui n'est pas satisfaite car sinon on devrait conclure que l'action d'aucun des deux agents n'est la cause du fait que M_1 est déclenchée. Intuitivement, si l'un n'avait pas agi, l'effet aurait été obtenu grâce à l'action de l'autre (voir la définition 6 dans (Demolombe, 2012)).

3.2. Influence

Nous avons défini l'influence comme une forme affaiblie de la causalité dans le sens où un acte influence le fait que ϕ soit vraie si ϕ est vraie pour certaines instances de réalisation de cet acte, mais pas nécessairement pour toutes.

On note $Infl_{ACT,ACT'}(\phi)$ le fait que l'acte ACT', réalisé dans le contexte de ACT, influence le fait que ϕ soit vraie.

Pour simplifier on suppose ici que ACT' dénote un seul acte. Les conditions de satisfaisabilité de cet opérateur s'expriment formellement de la façon suivante :

$M, w \models Infl_{ACT,ACT'}(\phi)$ si, et seulement si, il existe deux mondes w'et w'' tels que $R_{ACT,ACT'}(w, w', w'')$ et :

1. $M, w' \models \phi$ et,
2. $M, w'' \models \neg\phi$.

La condition 2) exprime que, pour l'instance de réalisation représentée par $< w, w', w'' >$, si ACT' n'avait pas été réalisé, toutes choses égales par ailleurs, on n'aurait pas eu ϕ. Donc, pour cette instance là, ACT' est la cause de ϕ.

Pour les instances où ϕ est faux en w' on ne peut pas conclure que la réalisation de l'acte est la cause du fait que ϕ est faux car il est possible que dans ces cas ϕ soit faux aussi en w''.

Il est intéressant de pouvoir comparer la force de l'influence de deux types d'actes dans le même contexte.

Intuitivement nous dirons que ACT'_1 influence plus le fait que l'on ait ϕ que ACT'_2 si, et seulement si, pour tous les mondes qui sont des instances de réalisation de ACT'_2 qui causent que ϕ est vraie, ces mondes sont aussi des instances de réalisation de ACT'_1 qui causent que ϕ est vraie, et il existe des instances de réalisation de ACT'_1

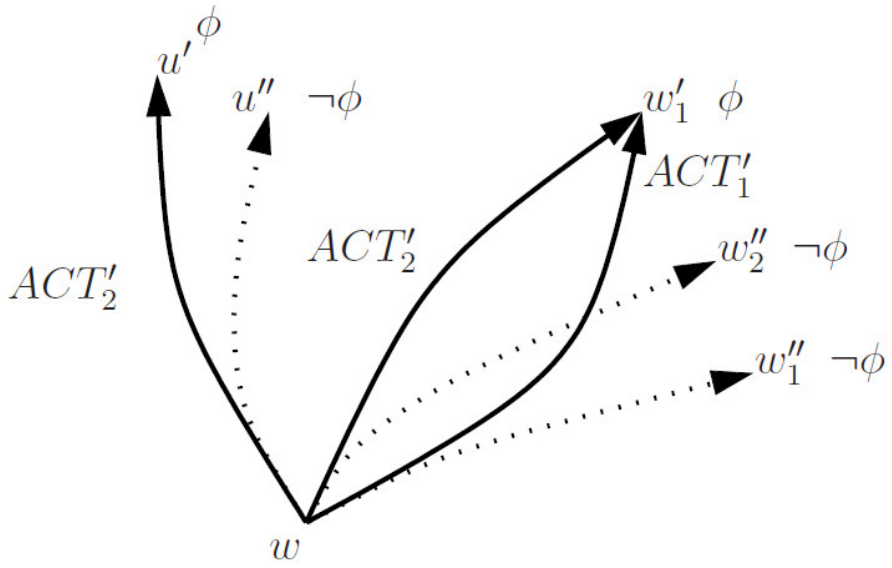

Figure 4. L'influence de ACT_1' sur ϕ est plus forte que celle de ACT_2'

qui causent que ϕ est vraie telles qu'il n'existe pas d'instance de réalisation de ACT_2' qui cause que ϕ est vraie dans le même monde.

Le fait que, dans le monde w, l'acte ACT_1' a plus d'influence sur ϕ que l'acte ACT_2' est noté :

$$M, w \models Infl_{ACT,ACT_1'}(\phi) > Infl_{ACT,ACT_2'}(\phi)$$

On note $Br_{ACT,ACT'}(w, w', \phi)$ le fait que pour l'instance de réalisation $< w, w' >$ de ACT', ACT' cause ϕ. Formellement on a $Br_{ACT,ACT'}(w, w', \phi)$ si, et seulement si, il existe w'' tel que $R_{ACT,ACT'}(w, w', w'')$ et $M, w' \models \phi$ et $M, w'' \models \neg\phi$

Formellement on a (voir figure 4):

$M, w \models Infl_{ACT,ACT_1'}(\phi) > M, w \models Infl_{ACT,ACT_2'}(\phi)$ si, et seulement si, pour tous les mondes w_1' tels que $Br_{ACT,ACT_2'}(w, w_1', \phi)$ on a $Br_{ACT,ACT_1'}(w, w_1', \phi)$ et il existe un monde u' tel que :

1. on a $Br_{ACT,ACT_1'}(w, u', \phi)$,
2. on n'a pas $Br_{ACT,ACT_2'}(w, u', \phi)$.

La figure 5 montre comment l'ensemble des mondes où des instances de réalisation de ACT_2' qui causent ϕ (rectangle étiqueté par ACT_2') est inclus dans l'ensemble des

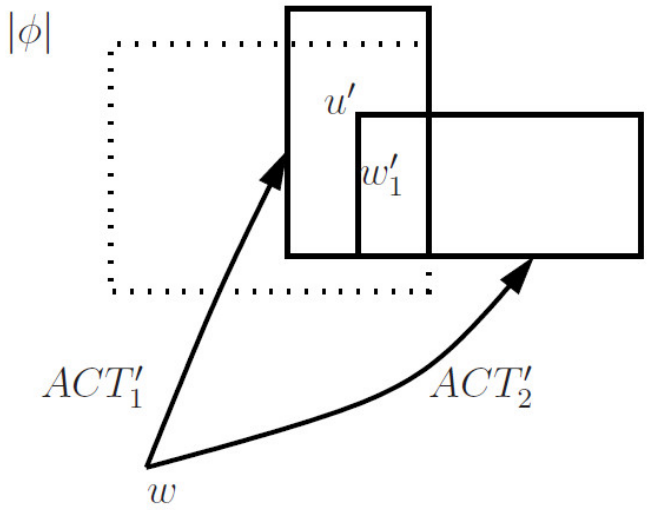

Figure 5. L'influence de ACT'_1 sur ϕ est plus forte que celle de ACT'_2

mondes où des instances de réalisation de ACT'_1 causent ϕ (rectangle étiqueté par ACT'_1).

3.3. Action d'un agent institutionnel

Nous avons vu que certains actes réalisés par un agent humain comptent comme des actes réalisés par un agent institutionnel (voir (Carmo, Pacheco, 2001 ; J. Gelati, Sartor, 2002 ; Santos, Pacheco, 2003 ; Grossi, 2007)). La formalisation de la relation entre ces deux types d'actes s'exprime par l'opérateur "compte comme" (*"counts as"*, en anglais; voir (Jones, Sergot, 1996)). Cette formalisation est complexe car elle ne peut pas être formalisée dans une logique modale normale et on doit l'exprimer dans une logique modale classique (voir Chellas (Chellas, 1988)) [11]. C'est pour cette raison que nous n'allons présenter ici que les schémas d'axiomes qui sont les plus significatifs et que nous ne présenterons pas la sémantique.

L'opérateur "compte comme" est noté : $\Rightarrow_S$, et la formule $\phi \Rightarrow_S \psi$ peut se lire : le fait que ϕ soit vraie compte comme le fait que ψ soit vraie dans le cadre de l'institution S.

11. La différence la plus importante entre une logique modale classique et une logique modale normale est que dans les interprétations des logiques modales classiques l'ensemble des formules qui sont vraies dans un monde est explicitement défini par une fonction. Par exemple, cette fonction peut spécifier que dans un monde ϕ et ψ sont vraies alors que $\phi \wedge \psi$ n'est pas vraie.

Le schéma d'axiome le plus intuitif que l'on peut accepter est que si ϕ' est logiquement équivalent à ϕ, et ψ' est logiquement équivalent à ψ, alors on obtient une formule équivalente en remplaçant dans $\phi \Rightarrow_S \psi$, ϕ par ϕ' et ψ par ψ'.

On utilisera aussi la modalité D_S. La formule $D_S\phi$ peut se lire : dans le cadre de l'institution S la proposition ϕ est vraie; on pourrait aussi la lire comme : dans le contexte de l'institution S, ϕ est reconnue comme vraie. Cette modalité est une modalité normale dans laquelle on a, en particulier, le schéma :

$$(K) \models D_S(\phi \rightarrow \psi) \rightarrow (D_S\phi \rightarrow D_S\psi)$$

Ces deux modalités sont liées par les deux schémas d'axiomes (D) et (C) ci-dessous :

$$\text{(D)} \models (\phi \Rightarrow_S \psi) \rightarrow D_S(\phi \rightarrow \psi)$$

$$\text{(C)} \models (\phi \Rightarrow_S \psi) \rightarrow (\phi \rightarrow D_S\psi)$$

Nous allons voir, en faisant référence à l'exemple présenté dans la section 2, quelles sont les conséquences qui peuvent être dérivées à partir de ces deux schémas d'axiomes. Les hypothèses sont les suivantes :

1. l'agent H_2 a fait en sorte que p (où p peut signifier, par exemple : tel hôpital a été détruit),

2. l'agent H_2 a fait en sorte que p, compte comme l'agent H_1 a fait en sorte que p,

3. l'agent H_1 a fait en sorte que p, compte comme : l'agent institutionnel I a fait en sorte que p.

L'hypothèse (2) peut être justifiée par le fait que c'est H_1 qui a donné l'ordre à H_2 de faire en sorte que p. L'hypothèse (3) peut être justifiée par le fait que H_1 a agi en tant que représentant de I. On utilise les notations :

- ϕ : l'agent H_2 a fait en sorte que p,
- ψ : l'agent H_1 a fait en sorte que p,
- θ : l'agent institutionnel I a fait en sorte que p.

Les hypothèses s'expriment alors formellement de la façon suivante :

1. ϕ,
2. $\phi \Rightarrow_S \psi$,
3. $\psi \Rightarrow_S \theta$.

De (2) et (C) on déduit : (4) $\phi \rightarrow D_S\psi$

De (3) et (D) on déduit : (5) $D_S(\psi \rightarrow \theta)$

De (5) et (K) on déduit : (6) $D_S\psi \rightarrow D_S\theta$

De (4) et (6) on déduit : (7) $\phi \rightarrow D_S\theta$

De (1) et (7) on déduit : (8) $D_S\theta$

La signification intuitive de (8) est que dans le contexte de l'institution S il est reconnu que l'agent institutionnel I a fait en sorte que p (l'hôpital est détruit).

En utilisant la formalisation que nous avons vue plus haut, le fait que l'agent logiciel M_1 a fait en sorte que p est exprimé par la formule :
$K(p) \stackrel{\text{def}}{=} Br_{ACT_2,M_1:c}(Br_{ACT_1,M_2:d}(p))$,
où ACT_1 désigne l'ensemble de tous les actes qui sont réalisés en même temps et qui contient $M_2 : d$ (plus loin on utilisera des notations analogues; par exemple : pour ACT_2 et $M_1 : c$). $K(p)$ peut se lire intuitivement : M_1 a fait en sorte que M_2 ait fait en sorte que p.

Le fait que l'agent humain H_2 a fait en sorte que p est exprimé par :

$$\phi \stackrel{\text{def}}{=} Br_{ACT_3,H_2:b}(K(p))$$

Le fait que l'agent humain H_1 a influencé H_2 pour qu'il fasse en sorte que p est exprimé par :

$$Infl_{ACT_4,H_1:a}(Br_{ACT_3,H_2:b}(K(p)))$$

Cependant, dans les cas où H_2 a fait en sorte que p, on peut, du point de vue de la responsabilité et de l'éthique, accepter que dans ces cas l'influence de H_1 a causé (et pas simplement influencé) le fait que H_2 a fait en sorte que p. Ceci permet de justifier que ϕ compte pour l'institution S comme si on avait :
$\psi \stackrel{\text{def}}{=} Br_{ACT_4,H_1:a}(Br_{ACT_3,H_2:b}(K(p)))$.

En effet, on peut interpréter l'opérateur d'influence de deux façons différentes. Si on se place avant que l'action qui influence ait commencé, il n'est pas certain que l'effet de l'action soit obtenu. Par contre, si on se place après que cette action ait eu lieu et que l'effet a été obtenu, on peut dire que c'est à cause de cette instance de réalisation de l'action qu'il a été obtenu. Or, dans ce qui précède on se place après que l'action ait été réalisée, et c'est ce qui peut justifier que, pour l'institution, tout se passe comme si l'action réalisée avait garanti d'obtenir l'effet.

On peut noter cependant que si on se plaçait du point de vue d'une loi qui interdirait que H_1 fasse en sorte que l'on ait p, et qui imposerait une pénalité dans le cas où elle a été violée, la responsabilité, et la pénalité, pourrait être moins forte dans le cas où H_1 a exercé une influence que dans le cas où il a été la cause de l'effet.

De plus, on peut noter que, du point de vue de l'institution S, le fait que ϕ compte comme ψ pourrait dépendre de la force de l'influence de l'action réalisée par H_1. Par exemple, si H_1 avait un grade moins élevé qu'un autre agent H_1' qui réalise la même

action, il se pourrait que l'influence de H_1' soit considérée par l'institution comme une cause, mais pas l'influence de H_1. Ceci montre l'intérêt de pouvoir donner un sens précis au fait que la réalisation d'un acte exerce une influence plus forte que la réalisation d'un autre acte comme nous l'avons fait dans la section 3.2.

Enfin, du point de vue de l'institution S, le fait que l'agent H_1 a fait en sorte que H_2 a fait en sorte que p compte comme si c'était l'agent institutionnel I qui avait fait en sorte que H_2 a fait en sorte que p. On a alors :

$$\theta \stackrel{\text{def}}{=} Br_{ACT_4,I:a'}(Br_{ACT,H_2:b}(K(p)))$$

4. Relations avec d'autres travaux

La formalisation de la causalité qui a été présentée est inspirée de plusieurs travaux antérieurs. Le premier est l'analyse de la causalité faite par von Wright dans (von Wright, 1963) où il met en évidence le fait que ce concept fait référence à trois situations : celle dans laquelle on est avant que l'action commence, celle dans laquelle on est quand l'action se termine et celle dans laquelle on aurait été si l'action n'avait pas eu lieu, toutes choses égales par ailleurs.

Nous avons repris de la formalisation par Segerberg dans (Segerberg, 2002) la notion de type d'acte formé d'un couple : <agent, type d'action>. De plus Segerberg fait explicitement la distinction entre type et instance d'action. L'avantage de cette formalisation par rapport à celle de von Wright est d'expliciter le type d'action qui est la cause de l'effet. L'avantage d'expliciter le nom de l'action que l'on considère est de pouvoir distinguer différents types d'actions qui produisent le même effet, c'est-à-dire de pouvoir indiquer comment l'effet a été obtenu (par exemple, si cela pouvait présenter un intérêt, cela permettrait de distinguer l'action d'appuyer sur un bouton avec la main droite ou avec la main gauche).

Des travaux de (Hilpinen, 1997) nous avons repris l'idée qu'à chaque instance de réalisation d'un type d'action correspond un monde contrefactuel qui est particulier à cette instance. C'est la raison pour laquelle il faut introduire une relation à trois arguments qui relie le monde initial, le monde après réalisation d'une instance d'action, et le monde contrefactuel qui correspond à cette instance particulière de réalisation.

Il y a de nombreuses autres formalisations de la causalité qu'il serait trop long de détailler ici (voir (Porn, 1977), (Hilpinen, 1997)).

En particulier dans (Horty, Belnap, 1995 ; Horty, 2001) Horty et Belnap ont présenté l'opérateur STIT (abréviation pour "*to See To It That*"). Cet opérateur permet de représenter la notion de choix que peut adopter un agent face à plusieurs alternatives, mais il peut conduire à des conclusions non intuitives si on l'interprète comme un opérateur qui représente la causalité. Considérons, par exemple, le cas où un agent a le choix entre appuyer sur un bouton qui déclenche le tir d'un drone ou ne pas appuyer sur ce bouton. Selon la définition du STIT, si l'agent choisit de ne pas appuyer sur le bouton on peut conclure que c'est son choix qui est la cause du fait que le tir du drone

n'a pas été déclenché. Cependant, la causalité a pour but de relier une action réalisée par un agent avec l'effet obtenu par cette action, et donc on ne peux pas dire que le fait de ne pas avoir réalisé une action est la cause du fait que le drone n'a pas tiré.

Pour la formalisation de l'influence nous nous sommes inspirés de Hilpinen qui dans (Hilpinen, 1997) considère plusieurs définitions de la causalité, dont une dans laquelle l'effet d'une action n'est pas garanti. Mais, à notre connaissance, il n'y a pas dans la littérature de définition d'une relation d'ordre qui permette d'exprimer qu'un type d'action est plus influent qu'un autre type d'action comme nous l'avons présenté plus haut.

Dans (Lorini, Sartor, 1994 ; 2016) l'influence est définie comme une action qui modifie (en les augmentant ou en les réduisant) le choix des types d'actions qui sont offerts à un agent. C'est une notion différente de l'influence. Elle porte sur les choix que peut faire un agent, alors que dans la définition que nous avons présentée l'influence concerne le fait que l'effet d'un type d'action est plus ou moins garanti.

La formalisation de l'opérateur *compte comme* a été reprise de Jones et Sergot (Jones, Sergot, 1996) qui donnent plus de détails sur l'axiomatique et présentent la sémantique correspondante (voir (Jones, Sergot, 1996 ; D. Grossi, Dignum, 2008)).

5. Conclusion

Les systèmes complexes auxquels nous nous intéressons font intervenir plusieurs types d'agents : les agents institutionnels, humains et artificiels, que l'on peut diviser en agents logiciels ou matériels. Nous avons vu que leurs interactions sont de natures différentes.

Les agents institutionnels sont de nature abstraite et ils n'agissent que par l'intermédiaire d'agents humains qui agissent en leurs noms. Les agents humains peuvent agir physiquement sur des agents logiciels ou matériels, dans une relation de causalité, ou sur des agents humains dans une relation d'influence, par exemple en donnant des ordres. Les agents artificiels interagissent entre eux, ou avec le monde extérieur, dans une relation de causalité.

Ces interactions sont particulièrement complexes quand certains agents artificiels utilisent des techniques d'intelligence artificielle, tel que c'est le cas pour certains systèmes d'armes qui font intervenir des drones. Dans ce cadre les effets des actions des agents artificiels peuvent être extrêmement graves. C'est la raison pour laquelle les règles d'éthique peuvent aussi concerner les agents institutionnels qui sont impliqués dans la conception, la réalisation et la fourniture des agents artificiels.

Nous avons proposé des définitions claires de ces interactions dans le formalisme des logiques modales. Nous avons ainsi précisé les définitions de la causalité, de l'influence et du fait que l'action d'un agent compte comme l'action d'un autre agent.

Nous avons montré que ces définitions permettent d'éviter des erreurs de raisonnement. Le raisonnement sur la causalité permet de raisonner sur les effets de la réalisa-

tion d'un acte. Le raisonnement classique permet de dériver les conséquences logiques des faits qui sont vrais dans le monde où on est quand les effets ont été obtenus. Le raisonnement sur les liens entre les actes réalisés par des agents et les actes réalisés par des agents institutionnels est défini par une logique particulière.

Par exemple, la réalisation de l'acte A peut avoir pour effet que p est vraie. Dans le monde où p est vraie on peut en déduire par la logique classique que $p \vee q$ est vraie, mais, contrairement à l'intuition, on ne peut pas en déduire que $p \vee q$ est un effet de A. Par ailleurs la réalisation de A peut compter comme la réalisation par un agent institutionnel de l'acte A', mais la réalisation de A' n'est pas une conséquence logique de la réalisation de A.

L'étude que nous avons présentée pourrait être prolongée dans plusieurs directions.

La première concerne les règles d'éthique qui, dans les applications qui nous intéressent, peuvent être nombreuses, et qu'il serait utile d'exprimer en logique déontique pour raisonner sur les obligations, permissions et interdictions afin, par exemple, d'en analyser toutes les conséquences, ou même de vérifier qu'elles ne contiennent pas de contradictions au sens de la logique (Garion *et al*., 2009). Ces règles de raisonnement sont différentes de celles que nous avons mentionnées précédemment et les travaux présentés dans (Makinson, 1998 ; Carmo, Jones, 2002 ; Demolombe, Jones, 2002 ; Broersen *et al*., 2003) montrent que leur définition pose des problèmes difficiles.

Une autre direction serait d'analyser une autre définition de l'influence. Nous avons défini l'influence comme un affaiblissement de la causalité. Une autre définition exprime qu'un agent peut influencer un deuxième agent en réalisant une action qui a pour effet de modifier les choix offerts au deuxième agent (comme dans (Lorini, Sartor, 1994 ; 2016)). Un exemple simple est celui où le premier agent réalise un acte qui permet au second de déclencher une arme dangereuse.

Remerciements

Nous tenons à remercier les relecteurs de cet article dont les commentaires ont permis de nombreuses améliorations sur la forme et sur le contenu. Nous remercions également Laurence Cholvy pour ses commentaires constructifs.

Bibliographie

Bonnemains V., Claire S., Tessier C. (2016). How ethical frameworks answer to ethical dilemmas: Towards a formal model. In *1st workshop on ethics in the design of intelligent agents.*

Broersen J., Dastani M., van der Torre L. (2003). BDIO-CTL: obligations and the specification of agent behaviour. In *International joint conference on artificial intelligence.*

Carmo J., Jones A. (2002). Deontic logic and contrary-to-duties. In D. Gabbay (Ed.), *Handbook of philosphical logic (volume 8).* Reidel.

Carmo J., Pacheco O. (2001). Deontic and action logics for organized collective agency, modeled through institutionalized agents and roles. *Fundamenta Informaticae*, vol. 48, p. 129–163.

Chellas B. F. (1988). *Modal logic: An introduction*. Cambridge University Press.

Cointe N., Bonnet G., Boissier O. (2016). Ethical judgment of agents' behaviors in multi-agent systems. In *International conference on autonomous agents & multiagent systems.*

Cuppens F. (2015). Roles and deontic logic. In *28th annual conference on legal knowledge and information systems.*

Demolombe R. (2012). Causality in the context of multiple agents. In T. Agotnes, J. Broersen, D. Elgesem (Eds.), *Deontic logic in computer science (LNAI volume 7393).* Springer Verlag.

Demolombe R., Jones A. (2002). Actions and normative positions. a modal-logical approach. In D. Jacquette (Ed.), *Companion to philosophical logic.* Blackwell.

Demolombe R., Louis V. (2006). Norms, institutional power and roles: toward a logical framework. In F. Esposito, Z. W. Ras, D. Malerba, , G. Semeraro (Eds.), *Foundations of intelligent systems (LNAI 4203).* Springer Verlag.

D. Grossi J.-J. C. M., Dignum F. (2008). The many faces of counts-as: A formal analysis of constitutive rules. *Journal of Applied Logic*, vol. 6.

ETHICAA. (2015). Dealing with ethical conflicts in autonomous agents and multi-agent systems. In *1st international workshop on artificial intelligence and ethics at the 29th AAAI conference on artificial intelligence.*

Garion C., Roussel S., Cholvy L. (2009). A modal logic for reasoning on consistency and completeness of regulations. In *Normative multi-agent systems.*

Grossi D. (2007). *Desigining invisible hancuffs. formal investigations in institutions and organizations for multi-agent systems*. Thèse de doctorat non publiée, Utrecht University.

Hilpinen R. (1997). On action and agency. In E. Ejerhed, S. Lindstrom (Eds.), *Logic, action and cognition: Essays in philosophical logic.* Kluwer.

Horty J. (2001). *Agency and deontic logic*. Oxford University Press.

Horty J., Belnap N. (1995). The deliberative STIT: A study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, vol. 24, p. 583–644.

J. Gelati A. R., G. Governatori, Sartor G. (2002). Declarative power, representation, and mandate: A formal anaysis. *Frontieres in Artificial Intelligence and Applications*, vol. 89.

Jones A. J., Sergot M. (1996). A formal characterisation of institutionalised power. *Journal of the Interest Group in Pure and Applied Logics*, vol. 4, n° 3.

Lewis D. (1973). *Counterfactuals*. Harvard University Press.

Lorini E., Sartor G. (1994). Influence and responsibility: A logical analysis. In *2nd international workshop on deontic logic in computer science.*

Lorini E., Sartor G. (2016). A STIT logic for reasoning about social influence. *Studia Logica*, vol. 104, n° 4.

Makinson D. (1998). On a fundamental problem of deontic logic. In P. McNamara, H. Prakken (Eds.), *Norms, logic and information systems.* IOS Press.

Pacheco O., Santos F. (2004). Delegation in a role-based organization. In A. Lomuscio, D. Nute (Eds.), *Deontic logic in computer science (LNCS 3065).* Springer.

Porn I. (1977). Action theory and social science. some formal models. *Synthese Library*, vol. 120.

Santos F., Pacheco O. (2003). Specifying and reasoning with institutional agents. In *9th international conference on artificial intelligence and law.*

Segerberg K. (2002). Outline of a logic of action. In F. Wolter, H. Wansing, W. de Rijke, M. Zakharyaschev (Eds.), *Advances in modal logic (volume 3).* World Scientific Publishing Co.

von Wright G. H. (1963). *Norm and action*. Routledge and Kegan.