Utilisation d'une logique de préférences conditionnelles pour raisonner avec des normes Contrary-To-Duties

Laurence Cholvy
ONERA Centre de Toulouse
Toulouse, France
cholvy@cert.fr

Christophe Garion
ONERA Centre de Toulouse
Toulouse, France
garion@cert.fr

Résumé

Le contexte de ce travail[1] *est la modélisation en logique du raisonnement déontique. Plus précisément, nous nous intéressons à la modélisation de normes de type Contrary-To-Duties (CTD) qui sont des structures exprimant une obligation primaire et une obligation secondaire qui prend effet lorsque l'obligation primaire est violée. Notre travail est basé sur la proposition de Carmo et Jones qui ont listé un certain nombre de postulats que doit satisfaire une logique pour raisonner correctement avec les CTD. Nous proposons d'adapter une logique de préférences conditionnelles, définie dans un tout autre contexte, celui de la Décision Qualitative.*

mot-clés : *logique déontique, logique de préférences*

1 Généralités sur les logiques déontiques

Les logiques déontiques sont des logiques qui permettent de raisonner sur des notions déontiques comme l'obligation, la permission et l'interdiction. La logique déontique la plus simple est connue sous le nom de SDL (Standard Deontic Logic) [Che80] et est une version déontique de la logique modale KD où l'opérateur primitif est l'obligation O. La permission est alors définie par $Pf =_{def} \neg O \neg f$ et l'interdiction par $If =_{def} O \neg f$. Cette logique a cependant de grandes limites, notamment lorsqu'il s'agit de raisonner avec des normes conditionnelles, et aussi avec des normes Contrary-To-Duties (CTD).

Les CTD sont des structures déontiques exprimant une obligation primaire et une obligation secondaire, qui prend effet lorsque l'obligation primaire est violée. Pour illustrer la notion de CTD, considérons l'exemple suivant :

(a) Les chiens sont interdits,

[1] Une première version de ce travail a été présentée à DEON'00. Une version modifiée est actuellement soumise à publication dans une revue internationale. L'article soumis ici présente non seulement ces derniers résultats mais fait également un survol des problématiques majeures du raisonnement déontique.

(b) (mais) s'il y a un chien, alors il doit y avoir une pancarte (le signalant)

L'obligation primaire est qu'il ne doit pas y avoir de chien (en fait, il s'agit ici d'une interdiction). L'obligation secondaire est qu'il doit y avoir une pancarte. Cette obligation secondaire prend effet lorsque l'obligation primaire est violée c.a.d, lorsque, malgré l'interdiction, il y a un chien.

Considérons un langage propositionnel avec deux variables *chien* (il y a un chien) et *pancarte* (il y a une pancarte). En SDL la phrase (a) est modélisée par : $O(\neg chien)$. Il y a deux façons de modéliser la phrases (b). On peut la modéliser par $O(chien \rightarrow pancarte)$ ou bien par $chien \rightarrow Opancarte$. Mais ces deux modélisations posent problème. Selon la première modélisation, la phrase (b) devient redondante par rapport à la phrase (a) puisque en SDL on peut déduire $O(chien \rightarrow pancarte)$ à partir de $O(\neg chien)$. Selon la seconde modélisation, on peut conclure, dans le cas où il y a un chien, qu'il y a à la fois l'obligation qu'il n'y ait pas de chien et une pancarte signalant la présence du chien ! Ce problème est connu sous le terme "d'idiotie pragmatique" (pragmatic oddity).

L'étude des CTD a fait l'objet de nombreuses recherches dans le domaine des logiques déontiques et a conduit à la proposition d'autres sémantiques. Très récemment, Carmo et Jones [CJ98] ont adopté une approche "axiomatique" des CTD en listant un certain nombre de postulats que doit satisfaire une logique déontique pour raisonner correctement avec les CTD. Ils ont aussi défini une logique modale qui satisfait ces postulats. L'originalité de ce travail est qu'il propose de prendre en compte un modèle de l'agent et notamment, de modéliser ce que l'agent peut faire et ce qu'il a décidé ou non de faire, ceci afin de distinguer ce que les auteurs appellent les obligations idéales (ideal obligations) et les obligations effectives (actual obligations).

Parallèlement, de nombreux travaux ont suggéré de doter la logique déontique d'une sémantique de type mondes ordonnés, la relation entre les mondes ordonnant les mondes des plus idéaux vers les moins idéaux. Hansson, [Han69], a été l'un des premiers à le proposer. Plus récemment, Prakken et Sergot [PS97], ainsi que van der Torre and Tan [VT99] ont également proposé d'adapter une sémantique de type mondes ordonnés pour raisonner correctement avec des normes de type Contrary-To-Duties. La justification principale à cette démarche est que, d'un point de vue sémantique, on peut considérer que les obligations exprimées par des normes induisent une préférence entre les mondes possibles. Par exemple, dans le cas de normes de type CTD les deux notions d'obligations primaires et d'obligations secondaires induisent un ordre de préférence entre les mondes.

Reprenons le petit exemple précédent et considérons à nouveau un langage propositionnel dont les deux lettres propositionnelles sont : *chien* et *pancarte*. Les mondes possibles sont : $w_1 = \{chien, pancarte\}, w_2 = \{chien, \neg pancarte\}$, $w_3 = \{\neg chien, pancarte\}$, $w_4 = \{\neg chien, \neg pancarte\}$.

Une interprétation possible des phrases (a) et (b) que nous pouvons donner en terme d'ordre sur ces mondes, est la suivante[2] :

D'après (a), on conclut que les mondes préférés sont ceux où il n'y a pas de chien, c.a.d., w_3 et w_4 sont potentiellement les mondes préférés. Donc, on a $w_3 < w_1$, $w_3 < w_2$, $w_4 < w_1$ et $w_4 < w_2$. Quant aux mondes w_1 et w_2 (ceux où il y a un chien) ils ne sont pas les plus préférés. Cependant, d'après (b), le monde w_1 est préféré au monde w_2 (car dans w_1, il y a une pancarte). Donc $w_1 < w_2$. On a donc deux ordres possibles : $w_3 < w_4 < w_1 < w_2$ et $w_4 < w_3 < w_1 < w_2$

Or, dans le contexte de la Théorie de la décision qualitative, [Bou94b] [Bou94a], Boutilier a défini une logique, appelée $CO*$ qui permet de raisonner sur des préférences conditionnelles pour calculer les buts de l'agent. En particulier, les deux ordres précédents correspondent à deux modèles de $CO*$ de deux formules qui expriment (a) et (b) en terme de préférences conditionnelles. De plus, dans son travail Boutilier a montré lui aussi l'importance de prendre en compte un modèle de l'agent pour caractériser ses buts à partir des préférences conditionnelles.

Il nous a semblé intéressant de rapprocher le travail de Boutilier et celui de Carmo et Jones, dans le but d'utiliser la logique $CO*$ pour raisonner avec des CTD, d'autant que la sémantique de $CO*$ est basée sur la logique de Hansson et que Boutilier lui-même a proposé une lecture déontique des opérateurs de sa logique. Cette mise en relation constitue la contribution principale de notre travail.

Cet article est organisé de la façon suivante. La section 2 résume le travail de Carmo et Jones sur les CTD en insistant sur la définition du modèle d'agent et sur son impact sur la définition de deux types d'obligations. La logique $CO*$ est présentée dans la section 3. Nous insisterons particulièrement sur le modèle d'agent suggéré par Boutilier et sur son impact sur la définition des buts. La section 4 montre l'extension que nous suggérons au travail de Boutilier afin de satisfaire les postulats définis pas Carmo et Jones pour raisonner correctement avec des CTD. Un exemple benchmark est examiné dans la section 5. Enfin, nous terminerons dans la section 6 par une discussion critique sur ce travail.

2 Proposition de Carmo et Jones relativement aux CTD

La proposition de Carmo et Jones est double : tout d'abord, ils ont listé huit postulats que doit satisfaire une logique pour raisonner correctement avec des CTD. Ensuite, ils ont défini une logique qui satisfait ces postulats.

2.1 Les postulats définis par Carmo et Jones

Pour proposer les postulats, Carmo et Jones se basent sur une structure de phrases, qui incluent un CTD et dont l'exemple benchmark est le paradoxe de

[2]On notera $\leq$ une relation de préférence entre les mondes. Et on écrira $w \leq w'$ ssi w est préféré à w'. $<$ est la relation stricte induite.

Chisholm[3] :

(a) les chiens sont interdits
(b) s'il y a un chien, alors il doit y avoir une pancarte
(c) s'il n'y a pas de chien, alors il ne doit pas y avoir de pancarte
(d) il y a un chien.

La modélisation en SDL de ces quatre phrases conduit, selon les modélisations que l'on fait des normes conditionnelles (cf section 1), soit à des redondances entre les phrases, soit à des inconsistances. Des tentatives ont été faites pour utiliser des logiques temporelles pour résoudre le paradoxe de Chisholm. Mais, tout comme Prakken et Sergot [PS96], Carmo et Jones rejettent cette solution car les exemples de CTD ne sont pas tous des exemples où le temps intervient. Parallèlement, certains ont proposé d'utiliser des logiques non-monotones pour résoudre le problème posé par ce paradoxe. Mais cette approche a été critiquée du fait que la règle (c) ne doit pas être considérée comme une exception à la règle (a) mais comme décrivant un cas de violation de l'obligation primaire.

Un autre problème grandement discuté concerne la représentation de (b) et de (c). Selon Prakken et Sergot [PS96], (b) et (c) doivent être modélisées de façon différente car (c), contrairement à (b) est la règle qui définit l'obligation secondaire et qui donc caractérise le CTD. Carmo et Jones rejettent ce point de vue car il a pour conséquence que la représentation des ces phrases dépend des mises à jour : si l'on veut introduire un nouveau CTD (par exemple, rajouter une obligation ternaire du genre : (mais) s'il n'y a pas de pancarte alors, il doit y avoir un grillage) alors il faut modifier la modélisation déjà faite.

De plus, il faut noter, comme le soulignent Carmo et Jones, que les exemples de CTD font apparaitre deux notions d'obligations, qu'ils appellent *obligations idéales* (*ideal obligations)* et *obligations effectives* (*actual obligations*). Ainsi, dans l'exemple précédent, on doit être capable de dériver que l'obligation idéale est qu'il n'y ait pas de chien. Mais que, sous des circonstances particulières (violation de l'obligation précédente) il faut pouvoir dériver l'obligation effective de mettre une pancarte.

Enfin, le problème d'idiotie pragmatique, déjà mentionné dans la section 1, doit être évité.

Tout ceci conduit Carmo et Jones à l'ensemble de postulats suivants :

(i) L'ensemble des formules (a), (b), (c) et (d) doit être consistant ;
(ii) Les formules (a), (b), (c) et (d) doivent être logiquement indépendantes.
(iii) La logique doit s'appliquer sur les exemples de CTD sans référence au temps ni aux actions ;
(iv) Les phrases (b) et (c) doivent avoir des structures similaires ;
(v) On doit être capable de dériver des *obligations idéales* ;
(vi) On doit être capable de dériver des *obligations effectives* ;
(vii) On doit être capable de représenter le fait qu'une obligation a été violée ;

[3] L'exemple que nous donnons ici est une reformulation de la version initiale du paradoxe de Chisholm qui est : X doit aller aider ses voisins ; si X va aider ses voisins alors il doit leur dire qu'il y va ; si X ne va pas aider ses voisins, alors il ne doit pas leur dire qu'il y va ; X ne va pas aider ses voisins.

(viii) Le problème d'idiotie pragmatique doit être évité ;

2.2 La logique de Carmo et Jones

La logique proposée par Carmo et Jones est basée sur un opérateur modal dyadique O qui permet d'exprimer les normes conditionnelles et deux opérateurs modaux monadiques O_i et O_a pour exprimer respectivement les obligations idéales et les obligations effectives. La question principale est : quels sont les contextes qui permettent de dériver ces deux types d'obligations ? Pour répondre à cette question, Carmo et Jones proposent un modèle de l'agent qui distingue "la nécessité dépendante de l'agent" et "la nécessité indépendante de l'agent"

L'opérateur de nécessité dépendante de l'agent est dénoté ici $\Box_a$[4] et son dual, $\Diamond_a$ (i.e. $\Diamond_a\phi \equiv \neg\Box_a\neg\phi$). Intuitivement, $\Box_a\phi$ exprime que la proposition ϕ est fixée, dans une certaine situation, en fonction de ce que l'agent a décidé de faire (ou de ne pas faire).

Dans l'exemple, le fait que l'agent ait un chien et décide de ne pas s'en débarrasser est formalisé par : $\Box_a chien$. Par ailleurs, $\Diamond_a\neg chien$ exprime le fait que l'agent n'a pas décidé de ne pas se débarrasser de son chien (autrement dit, il accepte l'idée de s'en débarrasser).

Cet opérateur est utilisé pour dériver les obligations effectives grâce à l'axiome suivant : $O(B|A) \wedge \Box_a A \wedge \Diamond_a B \wedge \Diamond_a \neg B \rightarrow O_a B$.

Dans le contexte de l'exemple, une instance de cet axiome est : si l'agent a un chien et qu'il a décidé de le garder, qu'il n'a pas décidé de mettre une pancarte ni de ne pas mettre une pancarte, alors l'agent a l'obligation effective de mettre une pancarte. Notons que cet axiome de permet pas de dériver l'obligation effective pour l'agent, de mettre une pancarte si, par contre, il accepte l'idée de se débarrasser de son chien.

L'opérateur de nécessité indépendante de l'agent est quant à lui dénoté ici $\Box_i$ et son dual $\Diamond_i$. Intuitivement, $\Box_i\phi$ exprime que la proposition ϕ n'est pas fixée par une décision de l'agent mais est fixé, quoi que l'agent décide de faire. Par exemple, si la présence du chien est imposé à l'agent (le chien a élu domicile dans le jardin de l'agent et quoi que celui fasse, le chien ne part pas), on écrira $\Box_i chien$.

La nécessité indépendante de l'agent est utilisée pour dériver les obligations idéales, par l'axiome suivant : $O(B|A) \wedge \Box_i A \wedge \Diamond_i B \wedge \Diamond_i \neg B \rightarrow O_i B$.

Par exemple, si présence du chien est imposée à l'agent et si l'agent peut mettre ou non une pancarte, alors idéalement, il doit mettre une pancarte.

Enfin, Carmo et Jones définissent la notion de violation par : $Viol(A) =_{def} O_i A \wedge \neg A$, c.a.d. A est violée si A est idéalement obligatoire et si A est fausse.

[4] Nous avons changé les notations car Carmo et Jones utilisent des symboles que Boutilier a déjà utilisé dans son travail pour représenter des choses différentes

3 CO^*, une logique de préférences conditionnelles

Dans cette section, nous présentons la logique CO^* développée par Boutilier pour calculer les buts d'un agent à partir d'une ensemble de préférences conditionnelles, exprimées qualitativement. Nous rappelons sa sémantique, l'axiomatique étant présentée dans [Bou94a].

3.1 Sémantique de CO^*

Boutilier définit un langage propositionnel bimodal sur un ensemble de variables propositionnelles ($PROP$) avec les connecteurs habituels et deux opérateurs modaux, $\Box$ et $\overleftarrow{\Box}$. La sémantique de CO^* est fondée sur des modèles de la forme $M = \langle W, \leq, \phi \rangle$, où W est un ensemble de mondes possibles (interprétations du langage propositionnel), ϕ est une fonction de valuation qui associe toute variable de $PROP$ à un ensemble de mondes dans laquelle elle est vraie, et $\leq$ est un préordre total[5] sur W. $v \leq w$ signifie que v est un monde au moins autant préféré que w. Une contrainte est imposée sur les modèles de CO^*. Nous la rappellerons après la définition de la satisfaction d'une formule.

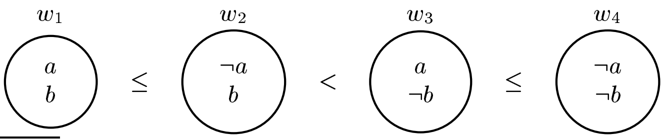
Soit $M = \langle W, \leq, \phi \rangle$ un modèle de CO^*. La satisfaction d'une formule dans M est définie par :

Définition 1 $M \models_w \alpha$ *ssi* $w \in \phi(\alpha)$, *pour tout variable propositionnelle* α.
$M \models_w \neg\alpha$ *ssi* $M \not\models_w \alpha$, *pour toute formule* α.
$M \models_w (\alpha_1 \wedge \alpha_2)$ *ssi* $M \models_w \alpha_1$ *et* $M \models_w \alpha_2$, *si* α_1 *et* α_2 *sont des formules.*
$M \models_w \Box\alpha$ *ssi pour tout* v *tel que* $v \leq w$, $M \models_v \alpha$.
$M \models_w \overleftarrow{\Box}\,\alpha$ *ssi pour tout* v *tel que* $w < v$, $M \models_v \alpha$.
$M \models \alpha$ *ssi pour tout* $w \in W$, $M \models_w \alpha$, *pour tout formule* α.

Ainsi, $\Box\alpha$ est vraie dans un monde w ssi α est vraie dans tous les mondes au moins autant préférés que w et $\overleftarrow{\Box}\,\alpha$ est vraie dans un monde w ssi α est vraie dans tous les mondes moins préférés que w. Les deux opérateurs $\Diamond$ et $\overleftarrow{\Diamond}$ sont définis par $\Diamond\alpha \equiv_{def} \neg\Box\neg\alpha$ et $\overleftarrow{\Diamond}\,\alpha \equiv_{def} \neg\,\overleftarrow{\Box}\,\neg\alpha$. Boutilier définit de plus $\overleftrightarrow{\Box}\,\alpha \equiv_{def} \Box\alpha \wedge \overleftarrow{\Box}\,\alpha$ et $\overleftrightarrow{\Diamond}\,\alpha \equiv_{def} \Diamond\alpha \vee \overleftarrow{\Diamond}\,\alpha$.

La contrainte qui est imposée sur tout modèle M est la suivante : pour toute formule satisfaisable φ de $PROP$, il existe au moins un monde w tel que $M \models_w \varphi$.

Exemple : Considérons un modèle M composé de 4 mondes w_1, w_2, w_3 et w_4 ordonnés comme suit :



[5]i.e. $\leq$ est une relation binaire réflexive, transitive et connectée

On a alors $M \models_{w_2} \Box b$, car tous les mondes au moins autant préférés que w_2 satisfont b.

Définition 2 *On dit que M satisfait une formule α ssi $M \models \alpha$. Soit E un ensemble de formules et α une formule de CO^*. On dit que α est dérivée de E ssi tout modèle qui satisfait E satisfait α. On le note $E \models \alpha$.*

3.2 Préférences conditionnelles

Pour pouvoir exprimer des préférences conditionnelles, Boutilier définit un opérateur $I(-|-)$ de la façon suivante :

Définition 3 $I(b|a) \equiv_{def} \overleftrightarrow{\Box} \neg a \vee \overleftrightarrow{\Diamond} (a \wedge \Box(a \rightarrow b))$ *et* $I(a) \equiv_{def} I(a|\top)$

On peut interpréter$I(b|a)$ par “si a, l'agent doit s'assurer que b”.

Définition 4 *La préférence relative entre deux propositions est définie par* $a \leq_P b \equiv_{def} \overleftrightarrow{\Box} (b \rightarrow \Diamond a)$

$a \leq_P b$ signifie donc que a est au moins autant préféré que b.

Pour pouvoir déterminer ses buts, un agent doit avoir une certaine connaissance du monde réel. Pour représenter cela, Boutilier introduit KB, un ensemble fini et consistant de formules propositionnelles qui représente la connaissance qu'a l'agent du monde. KB est appelé une base de connaissances. En prenant KB et un CO^*-modèle, les situations idéales sont caractérisées par les mondes les plus préférés qui satisfont toutes les conséquences de KB. Cette notion est modélisée formellement par :

Définition 5 *Soit E un ensemble de préférences conditionnelles. Soit KB une base de connaissance. Un but idéal dérivé de E est une formule propositionnelle α telle que $E \models I(\alpha|Cl(KB))$, où $Cl(KB) = \{\alpha \in PROP : \models KB \rightarrow \alpha\}$*

Exemple : Soit $PROP$ le langage propositionnel défini par les deux variables : r (l'agent traverse la rue) et v (l'agent est à vélo). Considérons les deux conditionnelles $I(r)$ et $I(\neg r|v)$ qui expriment que l'agent préfére traverser la rue, mais que si l'agent est à vélo, il préfére ne pas la traverser.

Les mondes possibles sont $w_1 = \{r, \neg v\}$, $w_2 = \{\neg r, v\}$, $w_3 = \{r, v\}$ et $w_4 = \{\neg r, \neg v\}$.

Soit M un modèle de CO^* satisfaisant $I(r)$ et $I(\neg r|v)$. Comme $M \models I(r)$, w_1 et w_3 peuvent être les mondes les plus préférés. Mais comme $M \models I(\neg r|v)$, w_3 ne peut pas être un des mondes les plus préférés. w_1 est donc le monde le plus préféré. De plus, on ne peut pas avoir $w_3 \leq w_2$ d'après $I(\neg r|v)$, donc $w_2 < w_3$. Les modèles possibles sont donc :

$$\begin{array}{lccccccc} M_1: & w_1 & < & w_2 & < & w_3 & < & w_4 \\ M_2: & w_1 & < & w_2 & < & w_4 & < & w_3 \\ M_3: & w_1 & < & w_4 & < & w_2 & < & w_3 \end{array}$$

Supposons que $KB_1 = \{\neg v\}$ (i.e. l'agent n'est pas à vélo). Alors, $Cl(KB) = \{\neg v\}$, donc les buts idéaux sont les formules α telles que $I(\alpha|\neg v)$ soit satisfaite dans tous les modèles précédents. Le but idéal est donc r. Comme l'agent n'est pas à vélo, l'agent a pour but de traverser la rue.

Si on suppose maintenant que $KB_2 = \{v\}$, alors on peut déduire que $\neg r$ est un but idéal. L'agent a pour but de ne pas traverser la rue.

3.3 Propositions contrôlables, influençables et non-influençables

D'après la définition précédente, les buts idéaux sont dérivés à partir de $Cl(KB)$. Mais, comme Boutilier le note dans [Bou94b], cette définition n'est plus correcte si l'agent peut changer la valeur de vérité de certains éléments de KB. Si l'agent est à vélo (comme dans KB_2) mais qu'il peut choisir de le laisser, on ne peut pas dériver ses buts idéaux à partir de v. Boutilier suggère donc, pour calculer les but idéaux, d'utiliser une restriction de $Cl(KB)$ aux formules dont l'agent ne peut pas changer la valeur de vérité. De plus, même si l'agent peut émettre des préférences sur des propositions sur lesquels il n'a aucun contrôle (il peut préférer qu'il fasse beau par exemple), ces propositions ne peuvent pas raisonnablement être appelées buts pour l'agent. Pour raffiner la notion de but, Boutilier introduit un modèle simple des capacités de l'agent[6]. Il partitionne l'ensemble des variables propositionnelles en deux classes disjointes C et $\overline{C}$ qui représentent respectivement les variables contrôlables par l'agent (dont il peut changer la valeur de vérité) et les variables incontrôlables par l'agent (dont il ne peut pas changer la valeur de vérité). Ainsi la variable représentant le fait "l'agent est à vélo" peut être considérée comme contrôlable, mais la variable représentant le fait "il fait beau" est incontrôlable par l'agent. Boutilier étend ensuite ces notions aux propositions :

Définition 6 *Soit P un ensemble de variables propositionnelles. $V(P)$ est l'ensemble des distributions de valeurs de vérité de cet ensemble. Si P et Q sont deux ensembles disjoints et que $v \in V(P)$ et $w \in V(Q)$, alors $v;w \in V(P \cup Q)$ représente les distributions correspondant aux variables de P et de Q. L'élément neutre pour ; est la distribution vide.*

Définition 7 *Une proposition α est* contrôlable *ssi, pour tout $u \in V(\overline{C})$, il existe $v \in V(C)$ et $w \in V(C)$ tels que $v;u \models \alpha$ et $w;u \models \neg\alpha$*

Une proposition α est influençable *ssi il existe $u \in V(\overline{C})$, il existe $v \in V(C)$ et w tels que $v;u \models \alpha$ et $w;u \models \neg\alpha$*

α est non-influençable *ssi elle n'est pas influençable.*

Définition 8 *Les croyances non influençables d'un agent sont définies par $UI(KB) = \{\alpha \in Cl(KB) : \alpha$ est non-influençable$\}$.*

Ainsi, si a est contrôlable et b est incontrôlables, $a \wedge b$ est influençable, mais pas contrôlable. En effet, l'agent ne peut pas changer la valeur de vérité de $a \wedge b$ si b est faux, mais par contre, si b est vrai, alors l'agent peut rendre $a \wedge b$ vraie ou fausse suivant la valeur qu'il donne à a.

[6] *agent's ability model* en anglais

Dans un premier temps, Boutilier considère que $UI(KB)$ est complet, i.e. que la valeur de vérité de tous les atomes incontrôlables est connue[7]. Sous cette hypothèse, il définit la notion de *CK-but*.

Définition 9 *Soient E un ensemble de préférences conditionnelles et KB une base de connaissances telle que $UI(KB)$ est complet. Une proposition α est un CK-but dérivé de E ssi $E \models I(\alpha|UI(KB))$ et α est contrôlable.*

Exemple : Considérons à nouveau $E = \{I(r), I(\neg r|v\}$. Supposons que l'agent ne soit pas à vélo, car celui-ci a été volé (dans ce cas, on a $\neg v \in KB$ et $v \in \overline{C}$).

Supposons que l'agent puisse encore traverser la rue (r est contrôlable). Dans ce cas, d'après la définition précédente, r est un CK-but dérivé de E : l'agent doit traverser la rue.

Supposons maintenant qu'il y ait des travaux sur la route et qu'il soit donc impossible de la traverser. Dans ce cas, r est incontrôlable. Toujours d'après la définition précédente, r n'est pas un CK-but dérivé de E, car l'agent ne peut pas traverser la rue.

4 Adaptation de $CO*$

Nous montrons ici comment adapter le formalisme de Boutilier de façon à satisfaire les postulats de Carmo et Jones, si, comme on le verra, on accepte de les modifier légèrement.

4.1 Modéliser l'exemple avec CO^*

Nous proposons de modéliser l'exemple benchmark des CTD (cf section 2) par les formules de $CO*$ suivantes :

(a) $I(\neg chien)$

(b) $I(\neg pancarte|\neg chien)$

(c) $I(pancarte|chien)$

(e) $(\neg chien \wedge pancarte) \leq_P (chien \wedge pancarte)$

(d) $chien \in KB$

On remarquera que les trois premières phrases de l'exemple sont modélisées par quatre formules. De ce fait, littéralement, on ne peut pas dire que cette modélisation satisfait les deux premiers postulats de Carmo et Jones. Toutefois, si on accepte de les reformuler en :

(i) *L'ensemble des formules qui modélisent (a), (b), (c) et (d) doit être consistant*

(ii) *Les formules qui modélisent (a), (b), (c) et (d) doivent être logiquement indépendantes*

On peut vérifier que (a), (b) (c) et (e) sont consistantes et qu'aucune n'est conséquence logique d'autres. Cependant, il y a d'une certaine façon une

[7] dans une seconde partie, il étudie le cas où $UI(KB)$ n'est pas complet. Nous n'étudierons pas ce cas ici

dépendance entre ces formules puisque (e) vient des phrases du langage naturel (a) et (c). Nous commenterons cela dans la section 6. Toutefois, cette modélisation respecte les postulats (iii) et (iv).

4.2 Modèle de l'agent

Le but de cette section est de relier les notions de contrôlabilité et influençabilité introduites par Boutilier et les notions de nécessité introduites par Carmo et Jones. Il faut préciser que cette mise en relation sera faite sous l'hypothèse que la connaissance du monde est complète, c.a.d pour toute proposition P, on a $\models KB \rightarrow P$ ou $\models KB \rightarrow \neg P$. Tout d'abord, remarquons que si une proposition F est contrôlable (resp, influençable, non-influençable) alors $\neg F$ est contrôlable (resp, influençable, non-influençable). Dans ce qui suit, nous examinons les différents statuts d'une formule propositionnelle F.

1. F est non influençable
   (a) Supposons que $\models KB \rightarrow F$ (pour simplifier, on dira que F est vraie dans le monde réel, ou plus simplement vraie). Alors, puisque F est non-influençable, quoi que fasse l'agent, F restera vraie. Cela signifie que F est nécessairement vraie, indépendamment de l'agent. Donc,

   Proposition 1 $\{F : \models KB \rightarrow F$ *et* F *non-influençable*$\} \equiv_{def} \{F : \Box_i F$[8]$\}$

   (b) Supposons maintenant que $\not\models KB \rightarrow F$. Alors, puisque KB est complète, on a $\models KB \rightarrow \neg F$. F est non-influençable, donc $\neg F$ l'est aussi. Quoi que fasse l'agent, F restera fausse.

   Proposition 2 $\{F : \not\models KB \rightarrow F$ *et* F *est non-influençable*$\} \equiv_{def} \{F : \Box_i \neg F\}$

2. F est influençable. Dans ce cas, F peut être contrôlable ou non.
   (a) F est contrôlable
      i. Supposons que $\models KB \rightarrow F$.
      Puisque F est contrôlable, l'agent peut décider de maintenir F vraie ou peut décider de changer sa valeur de vérité. Pour représenter cela, on introduit deux nouvelles notions dans le modèle d'agent de Boutilier :
         A. On dit que F, vraie, est *contrôlable-et-fixée* si l'agent décide de maintenir F vraie.

         Proposition 3 $\{F : \models KB \rightarrow F$ *et* F *contrôlable-et-fixée*$\} \equiv_{def}$

         $\{F : (F \wedge \Box_a F)\}$

[8] On pourrait ausi écrire $(F \wedge \Box_i F)$, mais ceci est équivalent car l'opérateur $\Box_i$ est de type KT.

B. On dit que F, vraie, est *contrôlable-et-non-fixée* si l'agent peut décider de changer la valeur de vérité de F.

Proposition 4 $\{F : \models KB \rightarrow F$ *et* F *est contrôlable-et-non-fixée*$\} \equiv_{def} \{F : F \wedge \Diamond_a \neg F\}$

ii. Dans le cas où F ne peut pas être déduite de KB, on a $\models KB \rightarrow \neg F$ (car KB est complète). F est contrôlable, donc $\neg F$ aussi. On peut traiter les mêmes cas (A) et (B) selon que $\neg F$ (ou, de façon équivalente F) est contrôlable-et-fixée ou contrôlable-et-non-fixée.

(b) Si F n'est pas contrôlable (mais influençable)

i. Supposons que $\models KB \rightarrow F$. F est vraie. Puisque F est influençable, il y a deux cas : l'agent peut ou non changer la valeur de vérité de F.

Ainsi, par exemple, supposons que A soit une variable contrôlable et que B soit une variable non contrôlable. Supposons que $KB = \{A, \neg B\}$. La formule $A \wedge B$ est fausse dans KB et restera fausse quoi que fasse l'agent (c.a.d quelle que soit la valeur de vérité qu'il donne à A). On dit que $A \wedge B$ est non-influençable dans KB. A l'inverse, la formule $A \vee B$ est vraie dans KB mais l'agent peut changer sa valeur de vérité en changeant celle de A. Plus précisément :

A. Si le monde réel est tel que l'agent ne peut pas changer la valeur de vérité de F, alors F restera vraie quoi que fasse l'agent. C'est comme si F était vraie et non-influençable. Dans ce cas, *on dira que* F *est non-influençable dans* KB.

B. Si le monde réel est tel que l'agent peut changer la vérité de F, deux cas se posent : soit l'agent peut décider de maintenir F vraie (cas 2.a.i.A) et F est contrôlable-et-fixée, soit l'agent peut changer sa valeur de vérité (cas 2.a.i.B) est F est contrôlable-et-non-fixée.

ii. Enfin, supposons que $\not\models KB \rightarrow F$.

Alors, puisque KB est complète, on a $\models KB \rightarrow \neg F$, donc F est fausse dans le monde réel. On peut lister les mêmes cas et déduire que F peut être non-influençable dans KB, contrôlable-et-fixée ou contrôlable-et-non-fixée.

Résumé. Cette comparaison nous permet de conclure que :

• Les propositions *vraies et non-influençables* et les propositons *vraies et non-influençables dans* KB correspondent aux propositions nécessairement vraies indépendamment de l'agent.

• Les propositions *vraies et contrôlables-et-fixées* correspondent aux propositions vraies et nécessairement vraies dépendamment de l'agent.

• Les propositions *vraies et contrôlables-et-non-fixées* correspondent aux propositions vraies et possiblement vraies dépendamment de l'agent.

4.3 Définitions

On étend ici les définitions de Boutilier pour prendre en compte le nouveau modèle des capacités de l'agent afin de définir les obligations idéales, effectives et les violations.

Définition 10 *$UI_1(KB)$ est l'ensemble des propositions qui sont vraies et non-influençables ou non-influençables dans KB. $UI_2(KB)$ est l'ensemble des propositions qui sont vraies et contrôlables-et-fixées ou vraies et non-influençables ou non-influençables dans KB.*

Définition 11 *Soit E un ensemble de préférences conditionnelles. Les obligations idéales dérivables de E sont définies par : $O_i\varphi \equiv_{def} \{O_i\varphi : E \models I(\varphi|UI_1(KB))$ et φ est contrôlable$\}$*[9]

Définition 12 *Soit E un ensemble de préférences conditionnelles. Les obligations effectives dérivables de E sont définies par : $O_a\varphi \equiv_{def} \{O_a\varphi : E \models I(\varphi|UI_2(KB))$ et φ est contrôlable-et-non-fixée$\}$*

Définition 13 *Les violations sont définies par : viol $\varphi \equiv_{def} O_i\varphi$ et $(\models KB \rightarrow \neg\varphi)$*

5 Etude de l'exemple

Les trois premières phrases de l'exemple sont modélisées par les formules : $I(\neg chien)$, $I(\neg pancarte|\neg chien)$, $I(pancarte|chien)$ et $(\neg chien \wedge pancarte) \leq_P (chien \wedge pancarte)$. Ci-après, nous examinons quelques cas intéressants.

5.1 Premier cas : $KB = \{chien, pancarte\}$

1. *chien* est contrôlable-et-fixée et *pancarte* est contrôlable-et-non-fixée. L'agent a un chien et a décidé de le garder. Il a mis une pancarte mais il accepte l'idée de l'enlever. On peut déduire $O_i\neg chien$ et $O_i\neg pancarte$ (c.a.d., idéalement, il ne doit y avoir ni chien ni pancarte). Ces deux obligations sont violées (puisqu'il y a un chien et une pancarte). Cependant, l'agent a une obligation effective : $O_a pancarte$. En d'autres termes, puisque l'agent a décidé de garder son chien, il a maintenant l'obligation de ne pas enlever la pancarte.
2. *chien* est contrôlable-et-non-fixée et *pancarte* est incontrôlable. L'agent a un chien mais il accepte l'idée de s'en débarrasser. Par contre, il y a une pancarte et l'agent ne peut pas l'enlever. La seule obligation idéale est $O_i\neg chien$ (idéalement il ne doit pas y avoir de chien) et elle est violée. Notons qu'il n'y a pas l'obligation idéale $O_i\neg pancarte$ car ici, la présence

[9] On rappelle que : $E \models \alpha$ ssi tout modèle M qui satisfait E satisfait aussi α

de la pancarte est imposée à l'agent. Toutefois, il y a une obligation effective : $O_a \neg chien$. Ceci signifie que, puisque l'agent accepte l'idée de se débarrasser de son chien, il a l'obligation effective de le faire.

5.2 Deuxième cas : $KB = \{chien, \neg pancarte\}$

1. *chien* est incontrôlable et *pancarte* est contrôlable-et-non-fixée. La présence du chien est imposée à l'agent. L'agent n'a pas mis de pancarte, mais il accepte l'idée d'en mettre une. La seule obligation idéale est $O_i pancarte$ (car *chien* est incontrôlable). Cette obligation est violée, car il n'y pas de pancarte dans la situation courante. Comme l'agent peut changer d'avis et mettre une pancarte, il y a une obligation effective : $O_a pancarte$, c.a.d., l'agent a l'obligation effective de mettre une pancarte.
2. *chien* est contrôlable-et-fixée et *pancarte* est incontrôlable. Cette fois ci, l'absence de la pancarte est imposée à l'agent (par exemple, l'agent n'a aucun moyen de mettre une pancarte car il n'y en a plus de disponible). L'agent a un chien et a décidé de ne pas s'en débarrasser. Dans ce cas, la seule obligation idéale est $O_i \neg chien$ et elle est violée, car il y a un chien. Il n'y a pas d'obligation effective.
3. *chien* est contrôlable-et-non-fixée et *pancarte* est incontrôlable. Comme précédemment, l'absence de la pancarte est imposée à l'agent, mais cette fois ci, l'agent peut décider de se débarrasser de son chien. Dans ce cas, la seule obligation idéale est $O_i \neg chien$, et elle est violée, car il y a un chien. Mais l'agent peut encore se débarrasser du chien, donc il a une obligation effective qui est $O_a \neg chien$: l'agent doit se débarrasser du chien.

5.3 Troisième cas : $KB = \{\neg chien, pancarte\}$

1. *chien* est incontrôlable et *pancarte* est contrôlable-et-fixée. Il n'y a pas de chien, et l'agent ne peut rien y faire, mais il y a une pancarte et l'agent a décidé de la garder. Dans ce cas, il y a une obligation idéale, $O_i \neg pancarte$ et elle est violée, car il y a une pancarte. Il n'y a pas d'obligation effective.
2. *chien* et *pancarte* sont contrôlable-et-non-fixées. Les obligations idéales sont $O_i \neg chien$ et $O_i \neg pancarte$ et elles sont violées toutes les deux. Les deux obligations effectives sont $O_a \neg chien$ et $O_a \neg pancarte$. L'agent ne doit pas acheter de chien et doit enlever la pancarte.

5.4 Quatrième cas : $KB = \{\neg chien, \neg pancarte\}$

Dans ce cas, suivant la contrôlabilité des deux variables *chien* et *pancarte*, on obtiendra $O_i \neg chien$, $O_i \neg pancarte$, $O_a \neg chien$ et $O_a \neg pancarte$. Dans tous les cas, puiqu'on est dans la situation idéale, aucune obligation idéale ne sera violée et l'agent sera tenu de rester dans cette situation.

6 Discussion

L'exemple benchmark précédent a été entièrement examiné et nous avons retrouvé les mêmes résultats que Carmo et Jones. Ceci tend à montrer l'intérêt du formalisme présenté ici. Même si certaines hypothèses de travail, comme l'hypothèse de complétude de KB, devraient être élargies. Il faut noter que dans une première présentation de ce travail [CG00], nous avions proposé une autre modélisation du benchmark qui conduisait à des résultats différents et que nous avons donc remise en cause par la suite. Cette hésitation illustre la difficulté du problème, par ailleurs général, de la modélisation.

Ce qui peut être sujet à critique dans la présente modélisation du benchmark, est que d'un point de vue syntaxique, les trois premières phrases de l'exemple sont modélisées par quatre formules, et que plus précisément, la quatrième formule dépende de deux de ces trois formules. Ceci sous-entend que les obligations primaires et secondaires d'un CTD ne peuvent pas être définies indépendamment. En d'autres termes, la première obligation d'une CTD n'a le statut d'obligation primaire que parce qu'une obligation secondaire est exprimée. Réciproquement, une obligation n'est secondaire que parce qu'elle prend effet lors de la violation d'une obligation primaire. Ceci peut paraître étrange, mais une remarque similaire peut être faite en ce qui concerne les normes révisables : on peut considérer qu'une obligation est révisable seulement parce qu'on envisage une exception. Ainsi, par exemple, dans la phrase *il ne doit pas y avoir de barrière sauf si la maison est près de la falaise*, on considère que la première obligation est révisable uniquement parce qu'il y a la deuxième partie de la phrase (qui exprime un cas d'exception). On peut remarquer, à ce sujet, que la modélisation donnée ici, implique que l'ajout ou la suppression d'une norme peut nécessiter une modification de la modélisation courante. Ce problème de dépendance du contexte a été soulevé par Carmo et Jones au sujet du travail de Prakken et Sergot et donc pourrait être à nouveau souligné ici.

Toutefois, l'approche préconisée dans cet article, qui consiste à adapter une logique des préférences conditionnelles pour raisonner avec des CTD, offre l'avantage de permettre de raisonner correctement aussi sur des normes avec exception (ce que ne fait pas la logique de Carmo et Jones). Ceci n'est pas surprenant puisque la logique $CO*$ a été définie pour cela. Considérons par exemple, la norme avec exception suivante : *il ne doit pas y avoir de barrière sauf si la maison est près de la falaise*. Du fait de l'ambiguité du langage naturel, nous pensons que deux interprétations peuvent être données à cette phrase. La première est : $I(\neg fence), \neg I(\neg fence|seaside)$. C'est à dire, généralement, il ne doit pas y avoir de barrière, mais si la maison est près de la falaise, alors il peut (au sens il est permis) y avoir une barrière. La deuxième interprétation est : $I(\neg fence), I(fence|seaside)$. C'est à dire, généralement, il ne doit pas y avoir de barrière, mais si la maison est près de la falaise, alors il doit y avoir une barrière. Evidemment, ces deux modélisations correspondent à des $CO*$-modèles différents, et les conclusions que l'on peut en tirer sont différentes.

Pour terminer, on notera que van der Torre et Tan [VT99] ont récemment suivi la même approche que celle qui est décrite ici, en définissant une logique dont la sémantique est encore basée sur la logique de Hansson. La grande différence est qu'à ce jour, ils ne prennent pas en compte de modèle des capacités de l'agent. La comparaison entre leur formalisme et celui qui a été présenté ici reste à faire.

Remerciements. Ce travail a été financé par l'ONERA.

Références

[Che80] B. F. Chellas. Modal logic, an introduction. Cambride University Press Publishers. 1980.

[Bou94a] C. Boutilier. Conditional logics of normality : a modal approach. *Artificial Intelligence*, 68 :87–154, 1994.

[Bou94b] C. Boutilier. Toward a logic for qualitative decision theory. In *Principles of Knowledge representation and Reasoning (KR'94)*. J. Doyle, E. Sandewall and P. Torasso Editors, 1994.

[CG00] L. Cholvy and Ch. Garion. An attempt to adapt a logic of conditional preferences for reasoning with Contrary-To-Duties. In *proceedings of DEON'00*. Toulouse, january 2000.

[CJ98] J. Carmo and A. Jones. Deontic logic and contrary-to-duties. In D. Gabbay, editor, *Handbook of Philosophical Logic (Revised Edition)*. Kluwer Academic Publishers (In Press).

[Han69] B. Hansson. An analysis of some deontic logics. *Noûs*, 3 :373–398, 1969.

[JP85] A. Jones and I Pörn. Ideality, sub-ideality and deontic conditionals. *Synthese*, (65) :275–290, 1985.

[PS94] H. Prakken and M. Sergot. Contrary-to-duty imperatives, defeasibility and violability. In *Proceedings of DEON'94, Oslo, Norway*, 1994.

[PS96] H. Prakken and M. Sergot. Contrary-to-duty obligations. *Studia Logica*, (57), 1996.

[PS97] H. Prakken and M. Sergot. Dyadic deontic logic and contrary-to-duty obligations. In D. Nute, editor, *Defeasible Deontic Logic*. Synthese Library, 1997.

[VT99] L. van der Torre and Y. Tan. Contrary-To-Duty Reasoning with Preference-based Dyadic Obligations. *Annals of Mathematics and Artificial Intelligence*. 27, 1-4, 1999.