Conditional Desires and Utilities – an Alternative Logical Approach to Qualitative Decision Theory

Jérôme Lang [1]

Abstract. We start from Boutilier's preference-based handling of conditional desires (*if a then ideally, b*) for modelling the preferences of an agent [2] and we propose an explicit construction of the agent's preference relation induced by a problem specification (desires + knowledge), via the definition of a class of utility functions satisfying the desires; we also define feasible worlds and preferred decisions with respect to a problem specification.

1 Introduction

Modelling preferences of rational agents has been tackled for long by decision theory (with utility-based or relational models), but generally without focussing on representational issues (*e.g.*, all possible states with their utility value are represented explicitely). On the contrary, AI languages for knowledge representation can also be used for preference modelling, to which they offer a way to represent preferences compactly and implicitly. Several steps have been done in this direction. In this paper we focus on conditional desires. Conditional desires are a way to express preferences implicitly and compactly. Boutilier [2] proposed a language for expressing qualitative preferences, where conditional desires express ideal states ("if A then ideally, B") . Other approaches to qualitative decision theory include [4], [12] (where conditional desires are interpreted very differently, via a *ceteris paribus* assumption) and [5] who give a possibility theory based view of decision theory.

What we propose in this paper is a procedure to induce the preference relation modelling the preferences of the agent directly from the specification of the problem, consisting of a set of initial conditional desires. Conditional desires $D(b|a)$ are still based on the interpretation in terms of ideal situations as in [2] (b holds in all preferred a-worlds), and we assume furthermore that when a rational agent expresses $D(b|a)$, there should a corresponding *loss of utility* resulting from its violation. As these losses of utilities will be summed up, we use numerical, additive utility functions as the basic semantical object. What we will actually get is a family of distinguished additive utility functions – a world ω will then be preferred to a world ω' iff for each distinguished utility function u, then $u(\omega) \geq u(\omega')$ – which defines a partial preordering on the set of worlds. Thus, what we get is a (qualitative) partial preference ordering on worlds, obtained from an initial set of conditional desires, via additive utility functions which are used only as intermediate objects.

Then we consider the case where the input is composed not only of conditional desires (expressing implicit preferences) but also *knowledge*. As we will see, not only it will be necessary to distinguish between desires (goals) and knowledge as in [2] but it will be also necessary to distinguish as in [15] between background factual knowledge (which tells which worlds are physically impossible) and contingent knowledge (which tells which of the physically possible worlds can be the actual state of affairs). The preference ordering is then defined on the set of feasible worlds, and then extended to the set of possible decisions. Lastly, our framework working also on many examples from the deontic logics literature, we discuss some links between conditional desires and violated obligations in deontic logics.

2 From conditional desires to preference relations

2.1 Conditional desires and utility functions

We work with a propositional language $\mathcal{L}$ generated from a finite number of propositional variables. Formulas are denoted by a, b, c etc. Tautology and contradiction are denoted respectively be $\top$ and $\bot$. W denotes the set of classical worlds associated with $\mathcal{L}$; worlds are denoted by ω, ... The set of worlds satisfying a is denoted by $Mod(a)$. $\subset$ denotes strict inclusion.

Definition 1 *A* **conditional desire** *is defined by a pair of propositional formulas a and b and is denoted $D(b|a)$. A* **desire specification** *is a finite set of conditional desires $DS = \{D(b_1|a_1), ..., D(b_n|a_n)\}$*

We could add without difficulty *absolute constraints*, *i.e.*, formulas that the agent absolutely wants to be satisfied. For the sake of brevity we omit them in this paper.

Definition 2 *A utility function is a map from W to **R**. The preorder $\geq_u$ induced by u is defined by $\omega \geq_u \omega'$ iff $u(\omega) \geq u(\omega')$. For $S \subseteq W$, the preferred worlds in S w.r.t. u are the worlds in S maximizing u, also denoted by $Max(\geq_u, S)$.*

We now define satisfaction of conditional desires by utility functions.

Definition 3 *Let $DS = \{D(b_1|a_1), ..., D(b_n|a_n)\}$ be a desire specification and u a utility function.*
$u \models DS$ *iff* $\forall i = 1 \ldots n$, $Max(\geq_u, Mod(a_i)) \subseteq Mod(b_i)$

We say that DS is consistent iff there is a u satisfying DS. Note the analogy with satisfaction and consistency in conditional logics (see [2]). Now, among all utility functions satisfying a desire specification, we distinguish a subset, obtained by assuming (1) that there is a specific loss of utility associated with the violation of each expressed conditional desire (or more precisely the violation of the material implication associated $a \rightarrow b$ associated to the conditional desire), and (2) that worlds gravitate towards preference. (1) leads to the definition of a "local" utility function $u_{b|a}$ for each conditional desire $D(b|a)$, which equals a negative constant $-\alpha$ for all worlds violating

[1] Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier, 31062 Toulouse Cedex (France). E-mail: lang@irit.fr

ECAI 96. *12th European Conference on Artificial Intelligence*
Edited by W. Wahlster
Published in 1996 by John Wiley & Sons, Ltd.

the desire (*i.e.*, satisfying $a \wedge \neg b$) and 0 for all other worlds. (2) means we assume that nothing more than these local utility functions contributes to the global utility function on W, or in other terms that this global utility function is the sum of all local utility functions.

Definition 4 *Let $D(b|a)$ be a conditional desire. $u_{b|a}$ is a* **local utility function** *associated to $D(b|a)$ iff there exists $\alpha > 0$ such that*

$$u_{b|a}(\omega) = \begin{cases} -\alpha & \text{if } \omega \models a \wedge \neg b \\ 0 & \text{otherwise} \end{cases}$$

Definition 5 *A utility function u is* **distinguished w.r.t.** DS *iff*
(1) $u \models DS$
(2) $u = u_1 + ... + u_n$
where u_1, ..., u_n are local utility functions associated respectively to $D(b_1|a_1), ..., D(b_n|a_n)$

Note that distinguished utility functions are never positive.[2] They make worlds tend towards preference (especially, if a world ω satisfies all conditional desires then it has a maximal utility, namely 0).

Thus, at a local level $D(b|a)$ behaves as the material implication ($a \rightarrow b$) whose violation costs a negative utility $-\alpha$ where α is a positive value which characterizes the strength of the conditional desire. The "ideality" conditions expressed by the conditional desires then induce constraints on the α's. Thus the set of distinguished utility functions associated to DS may be computed as the intersection of the set of all possible sums of local utility functions and the set of all utility functions satisfying DS.

Example 1 (from [2]): $DS = \{D(\neg U|\top), D(U|R)\}$, U and R standing for *umbrella* and *raining*. Local utilities are of the form
$u_{\neg U|\top}(\omega) = -\alpha$ iff $\omega \models U$, $= 0$ otherwise;
$u_{U|R}(\omega) = -\beta$ iff $\omega \models R \wedge \neg U$, $= 0$ otherwise.
It gives $u(UR) = u(U\bar{R}) = -\alpha; u(\bar{U}R) = -\beta; u(\bar{U}\bar{R}) = 0$; now, $u \models D(\neg U|\top)$ iff $max(u(\bar{U}R), u(\bar{U}\bar{R})) > max(u(UR), u(U\bar{R}))$, which yields $max(-\beta, 0) > max(-\alpha, -\alpha)$ *i.e.*, $\alpha > 0$. And $u \models D(U|R)$ iff $u(UR) > u(\bar{U}R)$, which yields something more interesting: $-\alpha > -\beta$, *i.e.*, $\beta > \alpha$: $D(U|R)$ takes precedence over $D(\neg U|\top)$ (because it is more specific). Therefore, u is a distinguished utility function satisfying DS iff $u(UR) = u(U\bar{R}) = -\alpha$, $u(\bar{U}R) = -\beta$, $u(\bar{U}\bar{R}) = 0$, with $\beta > \alpha > 0$.

Example 2 (Reykjavik scenario)[3]:
1. *You should not tell the secret to Gorbachev*;
2. *You should not tell the secret to Reagan*;
3. *If you tell the secret to Reagan, you should tell it to Gorbachev too*;
4. *If you tell the secret to Gorbachev, you should tell it to Reagan too*
Thus, $DS = \{D(\neg g|\top), D(\neg r|\top), D(g|r), D(r|g)\}$.
$u_{\neg g|\top}(\omega) = -\alpha$ iff $\omega \models g$, 0 otherwise
$u_{\neg r|\top}(\omega) = -\beta$ iff $\omega \models r$, 0 otherwise
$u_{g|r}(\omega) = -\gamma$ iff $\omega \models r \wedge \neg g$, 0 otherwise
$u_{r|g}(\omega) = -\delta$ iff $\omega \models g \wedge \neg r$, 0 otherwise
Now, $u = u_{\neg g|\top} + u_{\neg r|\top} + u_{g|r} + u_{r|g}$:
$u(rg) = -(\alpha + \beta) \quad u(r\bar{g}) = -(\beta + \gamma)$
$u(\bar{r}g) = -(\alpha + \delta) \quad u(\bar{r}\bar{g}) = 0$
Now, the constraints on α, β, γ, δ are the following:

$$\begin{aligned} u \models D(\neg g|\top) &\Leftrightarrow min(\alpha + \beta, \alpha + \beta) > 0 \\ u \models D(\neg r|\top) &\Leftrightarrow min(\alpha + \beta, \beta + \gamma) > 0 \\ u \models D(g|r) &\Leftrightarrow \alpha + \beta < \beta + \gamma \\ u \models D(r|g) &\Leftrightarrow \alpha + \beta < \alpha + \delta \end{aligned}$$

While the first two inequalities are trivially satisfied, the last two ones give $\alpha < \gamma$, $\beta < \delta$.

2.2 Preference relations

Let $DS = \{D(b_1|\top), ..., D(b_n|\top)\}$ be a consistent desire specification. From DS we draw a preference relation on worlds as follows:

Definition 6
$\omega \geq_{DS} \omega'$ iff for any distinguished u w.r.t. DS, $\omega \geq_u \omega'$.
$\omega >_{DS} \omega'$ iff $\omega \geq_{DS} \omega'$ and not ($\omega' \geq_{DS} \omega$).
$\omega \approx_{DS} \omega'$ iff $\omega \geq_{DS} \omega'$ and $\omega' \geq_{DS} \omega$.

Note that obviously, $\geq_{DS}$ is a preorder which is generally not complete. If $\omega \geq_{DS} \omega'$ we say that ω is preferred to ω' w.r.t. DS.

Example 2 (continued): since we have no constraint between $\beta + \gamma$ and $\alpha + \delta$, the worlds $r\bar{g}$ and $\bar{r}g$ are not comparable. The partial ordering $\leq_D$ is then the following (note that it is the intended ordering).

$$\begin{tikzcd} & \bar{r}\bar{g} \arrow[d] & \\ & rg \arrow[dl] \arrow[dr] & \\ r\bar{g} & & \bar{r}g \end{tikzcd}$$

Several properties of $\geq_{DS}$ are worth mentioning. The first one characterizes exactly indifferent worlds as worlds satisfying the same set material counterparts of conditional desires.

Proposition 1 *For a world ω, define*
$Sat(\omega, DS) = \{D(b_i|a_i) \in DS$ s.t. $\omega \models a_i \rightarrow b_i\}$. Then
(i)[4]if $Sat(\omega, DS) \subset Sat(\omega', DS)$ then $\omega <_{DS} \omega'$;
(ii) $\omega \approx_{DS} \omega'$ iff $Sat(\omega, DS) = Sat(\omega', DS)$.

The following properties concern particular cases of desires.

Proposition 2 (unconditional desires)
Let $DS = \{D(b_1|\top), ..., D(b_n|\top)\}$. Then DS is consistent iff $\wedge_{i=1..n} b_i$ is classically consistent, and in this case, for any two worlds ω, ω', $\omega \geq_{DS} \omega'$ iff $\{b_i, \omega \models b_i\} \subseteq \{b_i, \omega' \models b_i\}$

The result comes from the fact that there is no constraint linking the α_i's.

Proposition 3 (independent desires)
Let $DS = \{D(b_1|a_1), ..., D(b_n|a_n)\}$ such that $\forall i \neq j$, $a_i \wedge a_j$ is inconsistent. In this case for a given world ω there is at most one i such that $\omega \models a_i$: we denote this index by $i(\omega)$ when it is defined. Let us say that ω is good (w.r.t. DS) iff $i(\omega)$ is not defined or $\omega \models a_{i(\omega)} \rightarrow b_{i(\omega)}$; then, $\omega \geq_{DS} \omega'$ iff (ω' is good $\Rightarrow$ ω is good).

Since the way we use to distinguish utility functions makes use of addition, we might wonder if it is sensitive to the repetition of conditional desires. Fortunately, it is not:

Proposition 4 (redundant desires) *Assume that there are two conditional desires $D(b|a)$ and $D(b'|a')$ in DS such that a is logically equivalent to a' and $a \wedge b$ is logically equivalent to $a' \wedge b'$. Let $DS^* = DS \setminus \{D(b'|a')\}$. Then DS and DS^* generate exactly the same sets of distinguished utility functions and as a consequence, $\geq_{DS}$ is identical to $\geq_{DS^*}$.*

[2] Thus we could speak of *penalty functions* instead and work with positive quantities.
[3] This example is borrowed from the literature on deontic logics. In [15], two interpretations of the Reykjavik scenario are given; here we refer to the *mixed defeasibility and violability interpretation.*

[4] The converse of (i) does obviously not hold (see Examples 1 and 2).

Thus, $\{D(\neg g|\top), D(\neg r|\top), D(g|r), D(r|g), D(r|g)\}$ induces the same preference relation as $\{D(\neg g|\top), D(\neg r|\top), D(g|r), D(r|g)\}$.

Proposition 5 (specificity) *Let $DS = \{D(b|a), D(c|a')\}$ with $\vdash a' \rightarrow a$, $\not\vdash a \rightarrow a'$, and $b \wedge c$ inconsistent. Then the more specific desire takes precedence over the less specific one,* i.e., *the $\geq_{DS}$-maximal worlds satisfying $a \wedge a'$ (*i.e., *a') satisfy c (and $\neg b$).*

Proof: attach α to $D(b|a)$ and β to $D(c|a')$; distinguished utilities have the form $u(\omega) = 0$ for $\omega \models \neg a' \wedge (a \rightarrow b)$; $u(\omega) = -\alpha$ for $\omega \models a' \wedge \neg b \wedge c$; $u(\omega) = -\beta$ for $\omega \models a' \wedge b \wedge \neg c$; $u(\omega) = -(\alpha + \beta)$ for $\omega \models a' \wedge \neg b \wedge \neg c$. Now, $u \models D(b|a)$ iff $\alpha > 0$ and $u \models D(c|a')$ iff $min\{\alpha\} < min\{\beta, \alpha + \beta\}$, *i.e.*, iff $\beta > \alpha$. Hence the result.

Proposition 6 (absence of drowning effect) [5]
*Let $DS = \{D(b|a), D(c|a'), D(e|a)\}$ with $\vdash a' \rightarrow a$, $\not\vdash a \rightarrow a'$, and $b \wedge c$ inconsistent. Then the $\geq_{DS}$-maximal worlds satisfying $a \wedge a'$ (*i.e., *a') satisfy not only c but also e (the precedence of $D(c|a')$ over $D(b|a)$ does not inhibit $D(e|a)$) : there is no "drowning effect" [1].*

2.3 Taking account of background knowledge

In the previous sections we assumed that only desires were specified, but no pieces of knowledge. Knowledge restricts the set of possible worlds to a subset of W. However it is important here to distinguish between two kinds of "knowledge": "background knowledge" expressing that some worlds are physically impossible, and "contingent knowledge" (facts) expressing that some worlds which are physically possible are not the actual state of affairs (see examples in Section 3). Such a distinction has been made by van Der Torre [15] for the treatment of violated obligations in deontic logic. This distinction is important when computing the preference relation on the set of worlds (contingent knowledge should not be taken into account at this point, but only background knowledge). Computing the preference relation is a kind of "compilation", *i.e.* it is the same whatever the contingent knowledge is. Then, contingent knowledge will be taken into account by looking at the preferred worlds among those which will be feasible from the actual situation (for this we need a notion of decisions and feasibility that we introduce in Section 3).

Definition 7 (background knowledge, physically possible worlds) *Let $DS = \{D(b_1|a_1), ..., D(b_n|a_n)\}$, and BK a set of formulas (the background knowledge base). $\langle DS, BK\rangle$ is called a KD-background specification (KDBS). Worlds satisfying BK are called physically possible worlds.*

Definition 8 *$u \models \langle DS, BK\rangle$ iff*
(i) $\forall D(b|a) \in DS,\ Max(\geq_u, Mod(BK \wedge a)) \subseteq Mod(b)$;
(ii) $\forall \omega \models \neg BK, u(\omega)$ is undefined.

Definition 9 *$u_{b|a}$ is a local utility function associated to $D(b|a)$ w.r.t. BK iff there exists $\alpha > 0$ such that*

$$u_{b|a}(\omega) = \begin{cases} -\alpha & \text{if } \omega \models BK \wedge a \wedge \neg b \\ 0 & \text{if } \omega \models BK \wedge (a \rightarrow b) \\ \text{undefined} & \text{if } \omega \models \neg BK \end{cases}$$

A utility function u is distinguished w.r.t. $\langle DS, BK\rangle$ iff $u = u_1 + ... + u_n$ where u_1, ..., u_n are local utility functions associated respectively to $D(b_1|a_1), ..., D(b_n|a_n)$.

[5] As an illustration, see Example 4 in Section 3.1.

Definition 10 (preference relation on physically possible worlds) *Let $\langle DS, BK\rangle$ be a KDBS and ω, ω' two physically possible worlds. The definition of $\geq_{\langle DS,BK\rangle}$ is the same as for $\geq_{DS}$ (replacing "for any distinguished utility function w.r.t DS" by "for any distinguished utility function w.r.t $\langle DS, BK\rangle$".*

All results of Section 2.2 are easily generalisable to the case where background knowledge is taken into account.

3 Feasible worlds and preferred decisions

3.1 Definitions

So far we have defined two very different kinds of notions :
- desirability: some worlds are more desirable than others by the agent. These notions concern the agent and disconnected from "reality".
- physical possibility : some worlds are physically possible, some are not. This notion concerns the outside world.

Interestingly, it is generally not the case that all physically possible worlds are *feasible from the actual situation*. Namely, we introduce now a third notion, namely feasibility (ability): from the initial situation, there are some decisions the agent can take; each possible decision enables to reach a new world which may be different from the initial world (however the resulting world must be physically possible). Thus, we turn the problem into a decision problem: given a set of desires inducing a preference relation and factual knowledge specifying which worlds are feasible and which ones are not, the agent should attempt to achieve the best feasible situation. This is where contingent knowledge will be taken into account.

Our action model is rather simple. As in [2] we assume that our atomic propositions are partitioned into two classes: controllable atoms ($\mathcal{C}$), whose truth value is fixed by the agent, and uncontrollable atoms ($\bar{\mathcal{C}}$), whose truth value is fixed by nature. A truth assignment ω (resp. d) of all uncontrollable (resp. controllable) atoms is called a $\bar{\mathcal{C}}$-world (resp. a $\mathcal{C}$-world, or a *decision*). The set of all $\mathcal{C}$-worlds (decisions) is denoted by $\mathcal{D}$. Intuitively, an initial situation consists in a $\bar{\mathcal{C}}$-world ω (that the agent may not be know precisely since he may have an incomplete knowledge of the world) and a *initial partial decision* d_0, which is a truth assignment of a subset of the controllable atoms, representing the choices already made by the agent (see Examples 6, 7). When not specified, this initial partial decision is $\emptyset$ (no controllable is assigned a value). Now, the decisions that the agent may take (consist in assigning a truth value to the controllable atoms *which have not been assigned a truth value yet*. More formally, the set of possible decisions w.r.t. the initial partial decision d_0 is defined by $\mathcal{D}(d_0) = \{d \in \mathcal{D} | d \supseteq d_0\}$ where $d \supseteq d_0$ means that all controllable atoms already assigned in d_0 are assigned the same value in d. We further assume that facts involve only uncontrollable variables; the models of a set of facts F is thus a set of $\bar{\mathcal{C}}$-worlds. Since the agent may have an incomplete knowledge, he may not know the initial $\bar{\mathcal{C}}$-world with precision (*i.e.* $BK \wedge F$ may have several models), but he is always aware of the initial partial decision (d_0) he has already made – hence d_0 is in the specification of the problem. Lastly, we say that the effect of the possible decision $d \supseteq d_0$ on the $\bar{\mathcal{C}}$-world ω is the concatenation of assignments in ω and in d, and is denoted by (ω, d); it is called a *feasible world*:

Definition 11 (decisions, feasible worlds) *Let ω be a $\bar{\mathcal{C}}$-world (corresponding to the initial situation) and d_0 an initial partial decision. The set of feasible worlds w.r.t. BK, ω and d_0 is defined by $FW(BK, \omega, d_0) = \{d(\omega) | d \in \mathcal{D}(d_0)$ and $d(\omega) \models BK\}$*

The problem is that in the general case the agent does not know the actual initial world ω with precision, but instead he knows a set of possible initial worlds, one of which is the actual initial world (incomplete knowledge). These possible initial worlds are those satisfying both background knowledge and contingent knowledge (facts).

Definition 12 (contingent knowledge) *Let $\langle DS, BK\rangle$ be a KDBS, and F be a set of facts (contingent knowledge). (F, BK) is said to be $\bar{\mathcal{C}}$-complete iff there is only one $\bar{\mathcal{C}}$-world satisfying $BF \wedge F$; in this case this world will be denoted by ω_F.*

Intuitively, (F, BK) being $\bar{\mathcal{C}}$-complete means that the agent has a complete knowledge of the initial world regarding the uncontrollable atoms and thus that the set of feasible worlds is known exactly, and is equal to $FW(BK, \omega, d_0)$. On the contrary, if (F, BK) is not $\bar{\mathcal{C}}$-complete, the agent cannot always tell whether a world is feasible or not. What we want to do is to extend the preference relation induced on W by $\langle DS, BK\rangle$ to a preference relation on the set of possible decisions, taking account of contingent knowledge. As the case of $\bar{\mathcal{C}}$-complete initial knowledge is simpler than the general case, we first make this assumption before studying the general case.

3.2 Case of complete initial knowledge

$\mathcal{L}$ is a propositional language with $\mathcal{C}$ set of controllable atoms and $\bar{\mathcal{C}}$ set of uncontrollable atoms; DS is a set of conditional desires; BK is the background knowledge; $\geq_{\langle DS,BK\rangle}$ is the preference relation induced by $\langle DS, BK\rangle$; F is a set of facts *such that (F, BK) is $\bar{\mathcal{C}}$-complete*; if specified, d_0 is the initial partial decision (otherwise $d_0 = \emptyset$). ω_F is the only $\bar{\mathcal{C}}$-world satisfying $Mod(BK \wedge F)$.

Definition 13 (preferred feasible worlds and decisions)
Let d, d' be two possible decisions with respect to d_0. We say that $d \sqsupseteq_{\langle DS,BK,F\rangle} d'$ iff $(\omega_F, d) \geq_{\langle DS,BK\rangle} (\omega_F, d')$. A decision d is a preferred decision iff d is $\geq_{\langle DS,BK\rangle}$-maximal in $\mathcal{D}$; in this case (ω, d) will be called a preferred feasible world.

Equivalently, ω is a preferred feasible world iff ω is $\geq_{\langle DS,BK\rangle}$-maximal in $FW(BK, \omega_F, d_0)$. Preferred decisions represent optimal choices for the agent. There may be physically possible worlds above a preferred feasible world ω but these worlds are not feasible from the actual situation (although there are physically possible in other circumstances). Note that since $\geq_{\langle DS,BK\rangle}$ is generally not complete, $\sqsupseteq_{\langle DS,BK,F\rangle}$ is generally not complete either. Thus there may be several incomparable preferred decisions.

Example 3 (from deontic logics):
1. *If you are invited to a dinner you should not eat with your fingers*;
2. *If you are served asparagus you should eat with your fingers*;
3. *Being served asparagus means that you are invited to a dinner*;
4. *You are currently being served asparagus.*

The controllable atom is f (*eat with your fingers*) while d and a are uncontrollable.[6] 1. and 2. give conditional desires, while 3. is background knowledge ($\bar{d}a$ is physically impossible) and 4. is contingent knowledge (relevant to the current situation only). We check that the truth value of all uncontrollable atoms is known ($\bar{\mathcal{C}}$-completeness). Thus,

$DS = \{D(\neg f|d), D(f|a)\}$; $BK = \{a \rightarrow d\}$; $F = \{a\}$

Let α and β be associated respectively with $D(\neg f|d)$ and $D(f|d \wedge a)$.

[6] one may wonder why d is considered uncontrollable. Here, once you have started participating to the dinner, the truth of d cannot be changed, hence the uncontrollability.

The distinguished utility functions have the following form:
$u_{DS,BK}(daf) = -\alpha$ $\quad u_{DS,BK}(da\bar{f}) = -\beta$
$u_{DS,BK}(d\bar{a}f) = -\alpha$ $\quad u_{DS,BK}(d\bar{a}\bar{f}) = 0$
$u_{DS,BK}(\bar{d}\bar{a}\times) = 0$ $\quad u_{DS,BK}(\bar{d}a\times)$ is undefined.

The constraints expressed by the conditional desires give $\beta > \alpha > 0$ ($D(f|a)$ being more specific than $D(\neg f|d)$). Hence, we get the following preference relation on the set of physically possible worlds:

$$[d\bar{a}\bar{f}, \bar{d}\bar{a}f, \bar{d}\bar{a}\bar{f}] > [daf, d\bar{a}f] > [da\bar{f}]$$

Now, since $F = \{a\}$, the set of feasible worlds is $\{daf, da\bar{f}\}$. Since $daf >_{DS,BK} da\bar{f}$, the only preferred feasible world is daf. Hence the only preferred decision is $f :=$ true, *i.e.* the agent should eat with her fingers. Note that at in another situation where the agent is at a dinner where she is served another dish than asparagus $F = \{d, \neg a\}$, the preferred feasible world would then be $d\bar{a}\bar{f}$ and the preferred decision would be $f :=$ false. This example is an instance of what Horty [9] calls *overridden rules*.

Example 4: same as Example 3 before plus the conditional desire $D(n|d)$ (n means *you eat with a napkin* – controllable).
$DS = \{D(\neg f|d), D(f|a), D(n|d)\}$; $BK = \{a \rightarrow d\}$; $F = \{a\}$
The preferred feasible world is $dafn$, *i.e.* the preferred decision is (f := true, n := true). There is no "drowning effect" [1].

Example 5: p uncontrollable, a and b controllable.
$DS = \{D(a|\top), D(b|\top), D(\neg a \wedge \neg b|p)\}$; $BK = \emptyset$; $F = \{p\}$.
Since $F = \{p\}$, the set of feasible worlds is $\{pab, pa\bar{b}, p\bar{a}b, p\bar{a}\bar{b}\}$. Now, it can be checked that the preferred feasible world is $p\bar{a}\bar{b}$, *i.e.*, $D(\neg a \wedge \neg b|p)$ overwhelms the set of two desires $\{D(a|\top), D(b|\top)\}$. The preferred decision is thus (a := false, b := false).

Example 6 (Forrester paradox).
1. *You should not kill*;
2. *If you kill, you should do it gently*.
$DS = \{D(\neg k|\top), D(g|k)\}$ (g and k controllable).
$BK = \{g \rightarrow k\}$; $F = \emptyset$.
Let α and β be associated respectively with $D(\neg k|\top)$ and $D(g|k)$. We get $u_{\langle DS,BK\rangle}(\bar{k}\bar{g}) = 0$; $u_{\langle DS,BK\rangle}(kg) = -\alpha$; $u_{\langle DS,BK\rangle}(k\bar{g}) = -(\alpha + \beta)$. When $d_0 = \emptyset$, the preferred feasible world is $\bar{k}\bar{g}$. Now, assume that the agent has already decided to kill, ($d_0 = \{k\}$), the preferred feasible world becomes kg.

Example 7 (Reykjavik scenario, continued): DS as in Example 1; $BK = F = \emptyset$; g, r controllable. The preference ordering is shown in Example 2. When $d_0 = \emptyset$, the preferred decision is r := false, g := false. Now, if the secret has already been told to Reagan ($d_0 = \{r\}$), then the preferred decision becomes (r := true, g := true).

3.3 The general case (incomplete initial knowledge)

The problem consists now in extending $\sqsupseteq_{\langle DS,BK,F\rangle}$ to the case where the knowledge about the initial truth values of some uncontrollable is not known. Several possible definitions are possible; we propose here two of them:

Definition 14 (preferred decisions, Wald criterion)
Let d, d' be two decisions and DS, BK, F as before.
$d \sqsupseteq^{Wald}_{\langle DS,BK,F\rangle} d'$
iff $\forall\omega \models BK \wedge F, \exists\omega' \models BK \wedge F$ s.t. $d(\omega) \geq_{DS,BK} d'(\omega')$

Definition 15 (preferred decisions, Pareto criterion)
$d \sqsupseteq^{Pareto}_{\langle DS,BK,F\rangle} d'$ *iff* $\forall\omega \models BK \wedge F, d(\omega) \geq_{DS,BK} d'(\omega)$

Intuitively, $d \sqsupseteq^{Wald}_{\langle DS,BK,F\rangle} d'$ if the worst possible outcomes with d are not worse than the worst possible outcomes with d'; this pessimistic definition coincides with Wald's criterion of decision theory when the preference relation is a complete preordering. Boutilier [2] uses this criterion with a complete preference relation. The Pareto criterion has less discriminating power since $d \sqsupseteq^{Pareto}_{\langle DS,BK,F\rangle} d'$ iff, whatever the initial world, d will not yield a worse outcome than d', which might be rarely satisfied.

Proposition 7 $d \sqsupseteq^{Pareto}_{\langle DS,BK,F\rangle} d' \Rightarrow d \sqsupseteq^{Wald}_{\langle DS,BK,F\rangle} d'$

Proposition 8 *When the* (F, BK) *is* $\bar{\mathcal{C}}$*-complete,* $\sqsupseteq^{Wald}_{\langle DS,BK,F\rangle}$ *and* $\sqsupseteq^{Pareto}_{\langle DS,BK,F\rangle}$ *both collapse to* $\sqsupseteq_{\langle DS,BK,F\rangle}$ *defined in Section 3.2.*

4 Discussion

Related approaches belong to two different domains, which we think of being close to each other, namely, *logics for qualitative decision theory* (QDT) and *deontic logics*. Both distinguish between what *should be* the case (obligations / desires) from *is actually* the case (facts). In particular, dyadic deontic logics handle conditional obligations $O(b|a)$ ("if a is true then b should be true), which are veru similar in essence to conditional desires ("if a then I ideally would like b to be true"), especially when these deontic logics are based on a preference-based semantics (see [10] for a survey) : "$O(b|a)$ holds iff b holds in all preferred a-worlds [7]". Dyadic deontic logics were developed to handle so-called "contrary-to-duty (CTD) obligations" (see Examples 2,3,4,6,7). Defeasible obligations are handled in a nonmonotonic framework by Horty [9] who translates conditional obligations into default rules; Horty's approach was improved by van der Torre [15] (from whom we borrowed our distinction between background and contingent knowledge). Alternative approaches to the handling of CTD obligations were proposed by Prakken and Sergot [11], by Cholvy and Cuppens [3] who make use of *roles* ranked by a priority ordering assumed to be given – note that our approach could clearly be a basis for ordering roles automatically, taking account of specificity – and by Tan and van der Torre [13] who apply diagnosis techniques to finding a minimal set of violated obligations, similarly to our violations of desires but without taking account of specificity.

It is worth noticing that our construction of a preference relation gives the intended results on the examples of CDT-obligations taken from the deontic logics literature; thus, our work could be further developed in a deontic perspective. Now, we argue that deontic logics and logics for QDT have different perspectives and are thus more complementary than concurrent:

- first, one may argue that desires and obligations have a different interpretation from a philosophical point of view : in deontic logics, obligations are generally considered exogeneous (they are imposed by a legal or moral code) while desires in logics for QDT are endogeneous (directly specified by the agent). But this is not the main point, and we follow Boutilier [2] and Tan & van der Torre [14] in arguing that this difference should not necessarily lead to a need of different treatments for obligations and desires.

- second, and more importantly, the main purpose of a deontic logic is mainly deriving new obligations (and permissions) from an initial specification, while QDT focuses on the search for optimal acts / decisions. Thus, deontic logics may be viewed "upstream" and QDT "downstream", since the former provide representation of ideal states, or of a whole preference relation between states, and the latter use this preference relation ("goalness") [7] to find the best possible actions.

[7] An alternative logical approach to "goalness" has been proposed in [17].

Note that some recent approaches have started to incorporate acts into a deontic logic, either by integrating it with a dynamic logic, or by introducing a temporal component (see [10] for a survey). Our approach has the advantage of simplicity since we avoid the multiplication of modal operators, using instead a notion of controllability. This distinction between controllable and uncontrollable variables was further developed in a constraint satisfaction framework [6].

Our approach is complementary to Boutilier's [2] : while he focuses only on the definition optimal actions from a given a preference relation, we also focus on the practical generation of this preference relation (moreover Boutilier does not distinguish between background and contingent knowledge). Interestingly, our methodology contains two phases (generate the preference relation from a set of desires, and then find the optimal feasible worlds, and thus the optimal decision) which is in accordance with Tan & van der Torre's argumentation [14] about the two-phase treatment of violated obligations.

ACKNOWLEDGEMENTS

I would like to thank Craig Boutilier for fruitful discussions and Leendert van der Torre for helpful comments on a draft version of this paper. Much of this work was done while I was visiting the University of British Columbia.

REFERENCES

[1] S. Benferhat, C. Cayrol, D. Dubois, J. Lang and H. Prade, Inconsistency management and prioritized syntax-based entailment, *Proc. of IJCAI'93*, 640-645.

[2] C. Boutilier, Towards a logic for qualitative decision theory, *Proc. of KR'94*, 75-86.

[3] L. Cholvy and F. Cuppens, Solving normative conflicts by merging roles, *Proc. of AI and Law*, Washington, May 1995.

[4] J. Doyle and M. Wellman, Preferential semantics for goals, *Proc. of AAAI'91*, 698-703.

[5] D. Dubois and H. Prade, Possibility theory as a basis for qualitative decision theory, *Proc. of IJCAI'95*.

[6] H. Fargier, J. Lang and T. Schiex, Mixed constraint satisfaction: a framework for decision problems under incomplete knowledge, *Proc. of AAAI'96*.

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

[8] S.O. Hansson, Preference-based deontic logic, *Journ. of Philosophical Logic* 19:75-93, 1990.

[9] J.F. Horty, Moral dilemnas and nonmonotonic logic, *Journal of Philosophical Logic* 23, 35-65, 1994.

[10] D. Makinson, Five faces of minimality, *Studia Logica*, 52:339-379, 1993.

[11] H. Prakken and M. Sergot, Contrary-to-duty imperatives, defeasibility and violability, in *Proc. 2nd Workshop on Deontic Logic in Computer Science (Δeon'94).*

[12] S.-W. Tan and J. Pearl, Specification and evaluation of preferences for planning under uncertainty, *Proc. KR'94*.

[13] Y.H. Tan and L. van der Torre, Representing deontic reasoning in a diagnostic framework, Tech. Report RP 94.08.05, Erasmus University Research Institute in Decision and Information Systems, Rotterdam.

[14] Y.H. Tan and L. van der Torre, How to combine ordering and minimizing in a deontic logic based on preference, *Proc. 3rd Workshop on Deontic Logic in Computer Science (Δeon'96).*, 216-232.

[15] L. van der Torre, Violated obligations in a defeasible deontic logic, *Proc. ECAI'94*, 371-375.

[16] L. van der Torre and Y.H. Tan, Cancelling and overshadowing: two types of defeasibility in defeasible deontic logic, *Proc. IJCAI'95*.

[17] J. Wainer, Yet another semantics for goals and goal priorities, *Proc. ECAI'94*, 269-273.