Utilitarian Desires

Jérôme Lang (lang@irit.fr)
Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier
31062 Toulouse Cedex (France)

Leendert van der Torre (torre@cs.vu.nl)
Vrije Universiteit
de Boelelaan 1081a
1081 HV Amsterdam, The Netherlands

Emil Weydert (weydert@mpi-sb.mpg.de)
Max-Planck-Institute for Computer Science
Im Stadtwald
D-66123 Saarbrücken, Germany

Abstract. Autonomous agents reason frequently about preferences such as desires and goals. In this paper we propose a logic of desires with a utilitarian semantics, in which we study non-monotonic reasoning about desires and preferences based on the idea that desires can be understood in terms of utility losses (penalties for violations) and utility gains (rewards for fulfillments). Our logic allows for a systematic study and classification of desires, for example by distinguishing subtly different ways to add up these utility losses and gains. We propose an explicit construction of the agent's preference relation from a set of desires together with different kinds of knowledge. A set of desires extended with knowledge induces a set of 'distinguished' utility functions by adding up the utility losses and gains of the individual desires, and these distinguished utility functions induce the preference relation.

Keywords: Qualitative decision theory, QDT, BDI logic, logic of desires, utilitarian desires

1. Introduction

Autonomous agents reason frequently about preferences such as desires and goals. For example, they compete as well as cooperate more efficiently and effectively if they reason about the desires of other agents, and they therefore build up agent profiles directly by communicating their desires or indirectly by observing each other's behavior. However, communicated desires are compact representations of the agent's preferences which lack robustness, because they are sensitive to their exact specification. Consequently they easily lead to misinterpretation of the agent's preferences. Logical frameworks contribute to this problem, because they allow for a systematic study and classification of desires by making underlying assumptions explicit. Recently several logics for desires and goals have been proposed (Doyle and Wellman, 1991; Doyle et al., 1991; Boutilier, 1994; Pearl, 1993; Tan and Pearl, 1994b; Tan and Pearl, 1994a; Lang, 1996; Bell and Huang, 1997; van der Torre, 1998) to express preferences implicitly and compactly. For example, Cohen and Levesque (Cohen and Levesque, 1990) explore principles governing the rational balance among an agent's beliefs, goals, actions and intentions, Rao and Georgeff (Rao and Georgeff, 1991) show how different rational agents can be modeled by imposing certain conditions on the persistence of an agent's beliefs, desires or intentions (the BDI model) and work in qualitative decision theory (Pearl, 1993; Boutilier, 1994; Thomason and Horty, 1996; Bacchus and Grove, 1996; Doyle and Thomason, 1999; Thomason, 2000) illustrates how planning agents are provided with goals – defined as desires together with commitments – and charged with the task of discovering (or performing) some sequence of actions to achieve those goals.

In this paper we propose a logic of desires with a utilitarian semantics, in which we study non-monotonic reasoning about desires and preferences based on the idea that desires can be understood in terms of so-called utility losses (penalties for violations) and utility gains (rewards for fulfillments). Like in for example (Boutilier, 1994), conditional desires $D(a|b)$ are interpreted in terms of ideal situations by "a holds in all preferred b-worlds". Unlike other approaches we use numerical utility functions as our basic semantic objects, because we assume that the expression of $D(a|b)$ by a rational agent implies a corresponding *loss of utility* resulting from its violation ($\neg a \wedge b$) and/or a *gain of utility* from its fulfillment ($a \wedge b$). Moreover we use *additive* utility functions as these losses and gains of utilities are summed up (and weighted). Conditional or context-sensitive desires are thus formalized with basic concepts provided by decision theory (Dean and Wellman, 1991; Doyle and Wellman, 1991; Boutilier, 1994). Much of decision theory is concerned with conditions under which the preference ordering is representable by an order-preserving, real-valued *value* or (under uncertainty) *utility function*, and with identifying regularities in preferences that justify value or utility functions with convenient structural properties (Keeney and Raiffa, 1976). In this paper the formalization of desires is mainly concerned with the value and utility functions themselves, because on first approximation they only have a utilitarian or preference-based semantics. For desires we also introduce strength and polarity parameters, such that stronger desires can override weaker desires, like more specific desires override more general desires, and gain, loss and mixed desires can be distinguished, which respectively may induce only a gain of utility, only a loss of utility or specified combination of these two.

In particular we propose in this paper different procedures to induce from a set of initial conditional desires, which we call the problem specification, the preference relation of the agent. This preference relation, a partial pre-ordering on the set of worlds, is based on the so-called 'distinguished' utility functions of a problem specification, obtained by adding utility losses and gains. A world ω is at least as preferred as a world ω' if and only if for each distinguished utility function u, we have $u(\omega) \geq u(\omega')$. Summarizing, a set of desires induces a set of distinguished utility functions by adding up the utility losses and gains of the individual desires, and these distinguished utility functions induce a (qualitative) partial preference ordering on worlds. In the most advanced procedure to induce this preference relation we consider problem specifications that are composed of conditional desires expressing implicit preferences and *knowledge* expressing world constraints. Like (van der Torre, 1994) but in contrast to Boutilier (Boutilier, 1994) we also distinguish between factual background knowledge telling us which worlds are physically impossible and contingent knowledge telling us which of the physically possible worlds can be the actual state of affairs.

Our logic allows for a systematic study and classification of desires by making underlying assumptions explicit. For example, we distinguish between subtly different ways to add up utility losses and utility gains. Suppose that in the problem specification there are two desires to go to the zoo, one with utility loss 2 and one with utility loss 1. That is, the first desire states that not going to the zoo induces a utility loss of at least 2, and the second desire states that not going to the zoo induces a utility loss of at least 1. One straightforward way to add up the two utility losses is to induce from this desire specification that not going to the zoo induces a utility loss of at least 3. Another more subtle way to combine the desires is based on the intuition that the first desire implies the second one. In this case we induce from this desire specification that not going to the zoo only induces a utility loss of at least 2. The first way to add up the utility losses is based on a stronger notion of independence of the desires in the desire specification, because in this approach desires cannot be redundant. The different ways in which gains and losses can be added thus reflect different notions of dependency and redundancy.

The paper is organized as follows. In Section 2 we define and illustrate the monotonic and the non-monotonic logic of desires, and in Section 3 we introduce the procedure to induce a preference relation from a desire specification. In Section 4 we extend the framework with different types of knowledge. In Section 5 we consider potential further extensions by discussing the extension with defaults and the use of conditional desires as heuristic approximations of preferences in decision making and planning. Finally in Section 6 we discuss related research.

2. Specification of conditional desires

In this section we introduce a non-monotonic logic of desires, and in Section 3 we show how to induce a preference relation from a desire specification. In Section 2.1 we introduce the monotonic logic, in Section 2.2 we introduce the non-monotonic extension and in Section 2.3 we illustrate the logic by several benchmark examples from reasoning with desires.

2.1. Utility models for conditional desires

We consider a propositional language $\mathcal{L}$ generated from a finite number of propositional variables and the standard propositional connectives $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$, $\top$ (tautology), and $\perp$ (contradiction). Formulas are denoted by $a, b, c, \ldots$, and W denotes the set of propositional models (or worlds) associated with $\mathcal{L}$, to which we refer by $\omega, \omega', \ldots$. We write $\omega \models a$ to say that the world ω satisfies the formula a, and we set $\mathrm{Mod}(a) = \{\omega \mid \omega \models a\}$. On top of this language, we introduce the concept of parameterized conditional desires $D_s^p(b|a)$. Put in a nutshell, $D_s^p(b|a)$ states that $a \wedge b$ is preferred to $a \wedge \neg b$ in a specific way determined by the parameters p and s.

DEFINITION 1. (conditional desires, desire specification).
A conditional desire over $\mathcal{L}$ is defined by a pair of formulas $a, b \in \mathcal{L}$, a strength parameter $s \in \{\geq r, > r \mid r \in [0, \infty]\}$, and a polarity parameter $p \in [0, 1]$. It is denoted by

$$D_s^p(b|a).$$

A desire specification over $\mathcal{L}$ is a finite set of conditional desires

$$DS = \{D_{s_1}^{p_1}(b_1|a_1), ..., D_{s_n}^{p_n}(b_n|a_n)\}.$$

Our basic semantic units for interpreting conditional desires and evaluating desire specifications are extended real-valued utility functions. Similarly to (Lehmann, 1996; Lehmann, 1998), but in contrast to classical utility theory (von Neumann and Morgenstern, 1944), we also include $-\infty$ and ∞ in the range of our utility functions. The discussion on the use of these extreme values is out of the scope of this paper.

DEFINITION 2. (utility function).
A utility function u is a map from W to $\mathbb{R} \cup \{-\infty, +\infty\}$. u induces a preorder $\geq_u$ defined by $\omega \geq_u \omega'$ iff $u(\omega) \geq u(\omega')$. For $S \subseteq W$, let $\max(\geq_u, S)$ be the set of those $\omega \in S$ maximizing u, i.e. of the most desirable worlds in S according to u. For convenience, we abbreviate $\max_{\omega \in \mathrm{Mod}(a)} u(\omega)$ by $u(a)$.

What are the appropriate truth conditions for conditional desires? If we ignore the parameters, a common intuition is that the conditional desire $D_s^p(b|a)$ tells us that the best possible $a \wedge b$-worlds are strictly preferred to the best possible $a \wedge \neg b$-worlds. In other words, the best a-worlds have to satisfy b, i.e. $\max(\geq_u, \mathrm{Mod}(a)) \subseteq \mathrm{Mod}(b)$. The intuitive reading of $D_s^p(b|a)$ is that "ideally, if a is satisfied, then b is satisfied as well". This evaluation rule is well-known from many conditional logics (see for instance (Boutilier, 1994; Lehmann and Magidor, 1992)). This choice is based on the optimistic assumption that all worlds are accessible to the agent by means of the performance of some action, in particular those worlds which are the most desirable in the given context; in this case, $D_s^p(b|a)$ can be roughly interpreted as "if a is true, then it is in my interest to achieve b". We will come back to this issue in Section 4.2. See also (Boutilier, 1994) for further discussion of this point and see (van der Torre and Weydert, 2000) for other ways to interpret conditional desires in this semantics.

The parameters s and p allow a more fine-grained representation of conditional desires. The strength parameter $s \in \{\geq r, > r \mid r \in [0, \infty]\}$ fixes the minimal utility gap between the best $a \wedge b$- and the best $a \wedge \neg b$-worlds. In other words, the conditional desire $D_s^p(b|a)$ tells us that the best possible $a \wedge b$-worlds are s preferred to the best possible $a \wedge \neg b$-worlds. For practical reasons explained later, linked to the applicability of minimization procedures for determining preferred utility models, our discussion mainly concentrates on $\geq r$. For r unequal to zero, the distinction between $\geq r$ and $> r$ is mainly technical, and in both cases we have that it is implied that the best a-worlds have to satisfy b. However, for r equal to zero the distinction between ≥ 0 and > 0 is fundamental, because ≥ 0 just guarantees that no $a \wedge \neg b$-world is preferred to all $a \wedge b$-worlds, whereas > 0 tells us that there is an $a \wedge b$-world which is strictly preferred to each $a \wedge \neg b$-world. In other words, the former does not even express (positive) desirability in the sense that it is *not* implied that the best a-worlds have to satisfy b. $D_{\geq 0}^p(b|a)$ is a borderline case which should not be interpreted as a desire (for b given a) but as a lack of desire (for $\neg b$ given a). In fact, if a does not have utility $-\infty$ or $+\infty$, then we have $u \models D_{\geq 0}^p(b|a)$ if and only if $u \not\models D_{>0}^p(\neg b|a)$. For example, the neutral utility function, where all the worlds have utility 0, satisfies $D_{\geq 0}(b|a)$ for all consistent a, b. The desire $D_{\geq 0}$ still adds substantial expressive power to our logic although it is (nearly) equivalent to a negated desire, because our desire specifications as defined in Definition 1 do not explicitly include negated desire statements. See also Example 1 in the Section 2.2.

The polarity parameter $p \in [0, 1]$ has a special character, because unlike s it has no impact on the standard models of $D_s^p(b|a)$. However, as we explain in Section 2.2, conditional desires not only define constraints on utility valuations, but they also carry information about which utility models should be preferred. The polarity parameter is involved in the choice of preferred models of $D_s^p(b|a)$. In other words, it does not affect the monotonic logic of desires, but it affects the non-monotonic logic of desires. Consequently, parameter p does not appear in the truth conditions in the following Definition 3.

DEFINITION 3. (satisfaction).
Let $D^p_{\rhd r}(b|a)$ with $\rhd \in \{\geq, >\}$ be a conditional desire over $\mathcal{L}$ and u a utility function over W. The satisfaction relation $\models$ is defined as follows (by convention, $\infty \geq \infty + r$ and $\infty > \infty + r$)

$$u \models D^p_{\rhd r}(b|a) \quad \textit{iff} \quad u(a \wedge b) \rhd r + u(a \wedge \neg b)$$

If DS is a desire specification, then $u \in \mathrm{Mod}(DS)$ iff $u \models DS$ iff $u \models \delta$ for every $\delta \in DS$. A monotonic entailment relation $\vdash$ for conditional desires is defined by $DS \vdash \delta$ iff $\mathrm{Mod}(DS) \subseteq \mathrm{Mod}(\delta)$. We say that DS is consistent iff $\mathrm{Mod}(DS) \neq \emptyset$.

It follows from the truth conditions that, on a purely formal level, there is a close analogy between conditional desires, dyadic obligations (see section 6.2) and default conditionals. $D^p_{>0}(b|a)$, for instance, is axiomatized by the postulates for Lehmann's rational conditional logic (Lehmann and Magidor, 1992). On the other hand, $D^p_{\rhd r}(b|a)$ fails to verify rational monotony if $r > 0$ (see e.g. (Makinson, 1994) for details). In the present paper we do not discuss the axiomatic principles and the monotonic logic of conditional desires, because we are mainly interested in their non-monotonic logic related to strategies to extract preferred utility models, as explained in the following section. We also do not discuss extreme types of desire like for example $D^p_{>\infty}(b|a)$, which requires either that all the $a \wedge \neg b$-worlds are infinitely disliked, i.e. have utility $-\infty$, or that there must be an infinitely or absolutely desirable $a \wedge b$-world.

2.2. Distinguished utility models

Most desire specifications admit countless possible utility models. In particular, they are not specific enough to determine a unique utility distribution to be used by the agent. However, the more models there are, the less specific preferences are to guide and motivate the agent's decisions. Consequently, there is a pressing need to find reasonable strategies permitting the agent to constrain the set of utility models in a reasonable way, picking up those which are the most "likely", "intuitive", or "practical". In what follows, we are therefore going to define and investigate several procedures for identifying the most plausible - what we call the distinguished - models of a given specification. To prevent confusion with the preferences encoded by the utility functions, we avoid talking about preferred models.

How should we formalize distinguishedness? Let us start with the simplest task, namely finding distinguished interpretations of individual conditional desires. We assume that different desires may be considered independent, at least as long as they do not subsume each other, and that desires of the desire specification may interact. The idea is to build the distinguished models of any desire specification from the distinguished models of its elements, using some kind of additive aggregation. Our approach follows several plausible desiderata. Because these principles are not specific enough to determine a single best model, we are going to present additional constraints, whose choice is left to the user but which may be useful in some contexts. Without loosing too much, we focus our discussion on conditional desires of strength $\geq r$.

Consider the conditional desire $D^p_{\geq r}(b|a)$. How may we characterize its distinguished models? The basic idea is that a conditional desire does more than just formulating a constraint on utility functions. It also suggests a specific way to distribute penalties (negative utilities) and rewards (positive utilities) over worlds. $D^p_{\geq r}(b|a)$ proposes that a penalty should be attached to the $a \wedge \neg b$-worlds and a reward to the $a \wedge b$-worlds, whereas the $\neg a$-worlds should stay unaffected. However, it is important to point out that in the context of a desire specification, the exact values of these penalties and rewards may well depend on other desires. More about this later on.

First, it seems appropriate to ask for fairness in the sense that the process of attaching penalties or rewards should be unbiased. That is, worlds within the same desire context should receive equal treatment, which recalls Laplace's indifference principle for probabilities (see e.g. (Neapolitan, 1990)). These considerations motivate our first two desiderata for distinguished models of a single conditional desire.

- *1. Local uniformity.* The agent should be indifferent with respect to any two $a \wedge b$-worlds. Similarly for $a \wedge \neg b$-worlds and $\neg a$-worlds. In other words, the utility values should be constant within these three propositions.

- *2. Neutrality.* The agent should be neutral with regard to $\neg a$-worlds, which receive utility 0, and neither strictly prefer $\neg a$-worlds to $a \wedge b$-worlds, nor $a \wedge \neg b$-worlds to $\neg a$-worlds. That is, $u(a \wedge \neg b) \leq u(\neg a) = 0 \leq u(a \wedge b)$.

Utility functions which verify these two requirements about common utility values take the following form.

DEFINITION 4. (local utility functions - ignoring polarity).
Let $D^p_{\rhd r}(b|a)$ be a conditional desire. Then u is called a local utility function associated with $D^p_{\rhd r}(b|a)$ ignoring polarity iff there are $\alpha, \beta \geq 0$ such that

$$u(w) = \begin{array}{lll} -\alpha & \textit{if} & w \models a \wedge \neg b \\ 0 & \textit{if} & w \models \neg a \\ \beta & \textit{if} & w \models a \wedge b \end{array}$$

α is called the penalty or utility loss, and β the reward or utility gain for u. If $\alpha = \beta = 0$, u is called the null-function or void.

The local utility functions, or subclasses thereof, will become the building blocks of our distinguished utility models. What remains to be done is to extend the above notion to take into account the polarity parameter.

Informally speaking, the polarity parameter p expresses the prima facie proportion between the gain of utility for satisfaction, and the loss of utility for violation of a conditional desire $D^p_{\rhd r}(b|a)$. Accordingly, we may distinguish three types of desires: gain, loss, and mixed desires.

- *Gain desires*, which have polarity $p = 0$, are such that - by default - their satisfaction $(a \wedge b)$ induces a reward, while their non-satisfaction ($a \wedge \neg b$ or $\neg a$) does not contribute to the overall utility function of the agent. For instance, "*if I have fish for dinner then I prefer to drink white Bourgogne wine with it*" can clearly be thought of as a purely positive desire.

- *Loss desires*, which have polarity $p = 1$, are such that - by default - their violation $(a \wedge \neg b)$ induces a penalty, while their non-violation ($a \wedge b$ or $\neg a$) leaves the overall utility function unchanged. For instance, "*if it rains then I prefer to have an umbrella*" may be seen as a purely negative desire.

- *Mixed desires* form a continuous realm between gain and loss desires. By default, they induce both a reward if the desire is satisfied ($a \wedge b$) and a penalty if it is violated ($a \wedge \neg b$), whereas the context complement $\neg a$ stays unaffected. Consider the desire "*if I eat potatoes then I prefer them to be cooked*". It seems natural to most (hungry, but not starving) human agents that eating a cooked potato is better than nothing and that eating a raw potato is worse than nothing.

The polarity parameter p is meant to induce a default requirement for the gain-loss proportion to be observed by the preferred utility models of individual conditional desires. It does so by restricting the relative values of the penalties and rewards, which gives us our third desideratum.

- *3. Gain-loss proportion.* The distinguished models of a single conditional desire should satisfy $p = \frac{\alpha}{\alpha+\beta} \in [0,1]$ if $\alpha + \beta > 0$ (α, β as above).

Note that we only require that $p = \frac{-u(a \wedge \neg b)}{-u(a \wedge \neg b) + u(a \wedge b)}$ for single desires, not for distinguished models of arbitrary desire specifications including $D^p_{\rhd r}(b|a)$. The reason is that we have to take into account the possible interaction with other conditional desires. Furthermore, it is important to keep in mind that two desires differing only with regard to their polarity still share the same monotonic semantics, i.e. they have the same utility models. Only their preferred interpretations are different. We can now state the full definition of local utility functions.

DEFINITION 5. (local utility functions).
Let $D^p_{\rhd r}(b|a)$ be a conditional desire. Then u is called a local utility function associated with $D^p_{\rhd r}(b|a)$ iff u is a local utility function associated with $D^p_{\rhd r}(b|a)$ ignoring polarity, and the corresponding penalty α and reward β satisfy $\frac{\alpha}{\alpha+\beta} = p$ if $\alpha + \beta > 0$.

The first three desiderata characterize for each conditional desire $D^p_{\rhd r}(b|a)$ a class of elementary utility functions, the local utility functions associated with $D^p_{\rhd r}(b|a)$, which do not necessarily satisfy $D^p_{\rhd r}(b|a)$. From these functions we are going to choose the distinguished utility models of $D^p_{\rhd r}(b|a)$ by adding the standard truth-condition for conditional desires as our fourth desideratum.

- *4. Semantic constraint.* Penalty and reward should satisfy

$$-\alpha + r = u(a \wedge \neg b) + r \rhd u(a \wedge b) = \beta \quad (\rhd \in \{\geq, >\})$$

  That is, we must have $r \rhd \alpha + \beta$.

These four desiderata describe the weakest notion of distinguishedness for individual conditional desires.

DEFINITION 6. (distinguished utility model - single desire).
Let $D^p_{\rhd r}(b|a)$ *be a conditional desire. Then* u *is called a distinguished utility model of* $D^p_{\rhd r}(b|a)$ *iff* u *is a local utility function associated with* $D^p_{\rhd r}(b|a)$ *and* u *satisfies* $D^p_{\rhd r}(b|a)$.

It is easy to see that each conditional desire $D^p_{\rhd r}(b|a)$ admits a whole family of distinguished models. For instance, if u is a distinguished model, then all its multiples ηu, for $\eta \geq 1$, are distinguished as well. Because p is fixed, each distinguished model is already characterized by its penalty respectively reward. Local utility functions may be seen as partial distinguished models.

For conditional desires of the form $D^p_{\geq r}(b|a)$, there exists a single distinguished model minimizing α and β, i.e. the absolute values of the utilities. It encodes the most neutral stance - staying as close to 0 as possible - compatible with the given utility constraints, and implements the principle of avoiding unmotivated excitement. This seems to be a very natural way to strengthen distinguishedness and to guarantee uniqueness for desires of strength $\geq r$. It therefore becomes our - optional - fifth desideratum.

- *5. Minimality (for single desires).* Penalty and reward should be minimized, i.e. $\alpha + \beta = r$, where possible (for $\geq r$).

Putting all the requirements together, the intuitively distinguished model of $D^p_{\geq r}(b|a)$ then is fixed as follows.

DEFINITION 7. (minimal distinguished utility model - single desire).
Let $D^p_{\geq r}(b|a)$ *be a conditional desire. Then* u *is called a minimal distinguished utility model of* $D^p_{\geq r}(b|a)$ *iff* u *is the distinguished model of* $D^p_{\geq r}(b|a)$ *with minimal* $\alpha = -u(a \wedge \neg b)$ *and* $\beta = u(a \wedge b)$, *i.e.* $\alpha + \beta = r$.

In particular, the minimal distinguished model of $D^p_{\geq 0}(b|a)$ is the null function. It is clear that this approach does not work for conditional desires of the form $D^p_{>r}(b|a)$, where minimality cannot be achieved without violating the semantic constraint. It is therefore preferable to model desires with expressions of the form $D^p_{\geq r}(b|a)$. The following example illustrates minimality for single desires.

EXAMPLE 1. $DS_1 = \{D^{0.9}_{\geq 0}(p|\top)\}$ *and* $DS_2 = \{D^{0.9}_{\geq 1}(p|\top)\}$
The unique minimal distinguished utility function of DS_1 *is the null function, and the unique minimal distinguished utility function* u *of*

DS_2 is given by $u(\omega) = -0.9$ iff $\omega \models -p$, $u(\omega) = 0.1$ iff $\omega \models p$, and 0 otherwise. This shows the unusual character of the desires with strength ≥ 0: their minimal distinguished utility function is identical to the minimal distinguished utility function of the empty desire specification.

Our next task is now to generalize these notions of distinguishedness to (finite) sets of conditional desires. The basic intuition is that conditional desires, at least if they do not subsume each other, may be considered independent. This suggests a strategy where we build the global distinguished utility models by adding the rewards and penalties of the individual desires, but weighted in a way which reflects the interactions between the different desires. In other words, we add some suitable associated local utility functions. They constitute the elementary building blocks for constructing preferred utility models of the whole desire specification. Following this philosophy, only those models of a desire specification DS will be called distinguished which are representable as a sum of local utility functions associated with conditional desires in DS.

Observe that we do not require these local utility functions to be distinguished models of the specific desires. For instance, if one of two desires δ_1, δ_2 is logically redundant, e.g. if δ_1 entails δ_2, we may well ignore the penalty and reward corresponding to δ_2, i.e. accept a certain redundancy. This amounts to associate the null-function to δ_2, even if it fails to satisfy the desire (see also Example 5). In Definition 10 we are also going to discuss stronger notions of distinguishedness where every desire counts.

DEFINITION 8. (distinguished utility models).
Let $DS = \{D^{p_1}_{s_1}(b_1|a_1), ..., D^{p_n}_{s_n}(b_n|a_n)\}$ be a desire specification. A utility function u is called a distinguished utility model of DS iff there are (possibly void) local utility functions u_i associated with $D^{p_i}_{s_i}(b_i|a_i)$ such that

1. $u \models DS$
2. $u = u_1 + ... + u_n$.

$\mathrm{Mod}_d(DS)$ *is the set of all distinguished models of DS, and $DS \mathrel{|\!\sim}_d \phi$ iff for all $u \models DS$ we have $u \models \phi$.*

The following two examples illustrate the distinction between models and distinguished models, i.e. between monotonic and non-monotonic entailment. Both examples show that in the context of distinguished models, conditional desires stay valid if we strengthen the condition by some types of information which is considered irrelevant, and there are no further desires around. This doesn't hold for arbitrary utility models.

EXAMPLE 2. $DS = \{D^1_{\geq 1}(a|\top)\}$.
$\mathrm{Mod}(DS) = \{u \mid u(\neg a) + 1 \leq u(a)\}$ *is the set of all its utility models. We have strengthening of the condition* $\top$ *to* b *only if* b *is implied by* a*, i.e.:*

$$DS \models D^1_{\geq 1}(a|b) \text{ iff } a \vdash b$$

$\mathrm{Mod}_d(DS) = \{u \mid u(w) = -\alpha \leq -1 \text{ iff } w \models \neg a,\ u(w) = 0 \text{ iff } w \models a\}$ *is the set of distinguished utility models. We have strengthening of the condition* $\top$ *to* b *in all cases except when* $\neg b$ *is implied by* a*, i.e.:*

$$DS \mathrel{|\!\!\sim}_d D^1_{\geq 1}(a|b) \text{ iff } a \not\vdash \neg b$$

Strengthening is thus more easy to achieve, because we derive strengthening for all irrelevant b*.*

EXAMPLE 3. $DS = \{D^1_{\geq 1}(a|\top), D^1_{\geq 1}(b|\top)\}$.

$$\begin{array}{ll} \mathrm{Mod}(DS) = \{u \mid u(\neg a) + 1 \leq u(a), u(\neg b) + 1 \leq u(b)\} & \\ \mathrm{Mod}_d(DS) = \{u \mid u(w) = & -\alpha \leq -1 \text{ iff } w \models b \wedge \neg a \\ & -\alpha' \leq -1 \text{ iff } w \models a \wedge \neg b, \\ & -\alpha - \alpha' \text{ iff } w \models \neg a \wedge \neg b, \\ & 0 \text{ otherwise } \}. \end{array}$$

We have strengthening of the condition $\top$ *to* $\neg b$ *only in the non-monotonic case:*

$$DS \not\models D^1_{\geq 1}(a|\neg b)$$

$$DS \mathrel{|\!\!\sim}_d D^1_{\geq 1}(a|\neg b) \text{ iff } a \not\vdash b \text{ and } b \not\vdash a$$

As for individual conditional desires, the resulting general distinguishedness concept can now be strengthened or fine-tuned by restricting the choice of local utility functions in some reasonable ways. We start by defining minimal distinguished models of arbitrary desire specifications. Here the idea (minimality principle) is that each local utility function should contribute only as much as necessary to let the whole additive construction validate all the conditional desires. This amounts to prefer those distinguished utility models which minimize penalties and rewards. As before, we restrict ourselves to conditional desires of strength $\geq r$.

DEFINITION 9. (minimal distinguished utility models).
Let $DS = \{D^{p_1}_{\geq r_1}(b_1|a_1), ..., D^{p_n}_{\geq r_n}(b_n|a_n)\}$ *be a desire specification and* u_i *be a local utility function associated with* $D^{p_i}_{\geq r_i}(b_i|a_i)$ *with penalty* α_i *and reward* β_i*. A distinguished utility model* $u = u_1 + ... + u_n$ *of* DS *is called minimal iff there is no distinguished utility model* $u' = u'_1 + ... + u'_n$ *of* DS *with weights* (α'_i, β'_i) *such that for all* i, $\alpha'_i \leq \alpha_i$, $\beta'_i \leq \beta_i$, *and* $\alpha'_j < \alpha_j$ *or* $\beta'_j < \beta_j$ *for some* j. $\mathrm{Mod}_{md}(DS)$ *and* $|\!\sim_{md}$ *are defined in the obvious way.*

This approach does not work for desires of the form $D^p_{>r}(b|a)$, because then minimal distinguished utility functions do not exist. The present account may be compared to Tan and Pearl's gravitation towards indifference (Tan and Pearl, 1994b; Tan and Pearl, 1994a), but also to non-polarized minimality accounts in default reasoning (Goldszmidt et al., 1993; Weydert, 1998).

Because we get a smaller number of preferred models, or even a single one, this approach allows us to nonmonotonically infer more conditional desires or preferences. Nevertheless, the following example shows us that the minimal distinguished utility models are not necessarily unique (see also Example 8).

EXAMPLE 4. $DS = \{D^1_{\geq 1}(a|\top), D^1_{\geq 1}(b|\top), D^1_{\geq 1}(a \wedge b|\top)\}$
Let $u_a, u_b, u_{a\wedge b}$ *be the minimal local utility models of the desires in* DS*. Obviously, we have* $\alpha = 1$ *and* $\beta = 0$*, i.e. penalties and no rewards. We proceed by parallel minimization of the penalties of all the desires. But making one smaller, while supporting the semantic constraints, may force us to increase another one. This causes a trade-off between the local utility functions representing the different desires, resulting in multiple minimal distinguished utility models. More concretely, each sum* $\eta u_a + \eta u_b + (1-\eta) u_{a\wedge b}$*, for* $\eta \in [0.1]$*, is a minimal distinguished model of* DS*.*

A completely different strategy for strengthening distinguishedness, which is applicable to arbitrary conditional desires ($> r$ and $\geq r$), is to take each conditional desire seriously, avoiding subsumption and redundancy. This may be achieved by restricting the set of local utility functions used for building preferred utility models of DS. We distinguish two alternatives.

- Only those local utility functions associated with $\delta \in DS$ which are different from the null-function (except for desires of strength ≥ 0).
- Only those local utility functions associated with $\delta \in DS$ which are also distinguished utility models of δ.

Otherwise, we proceed as before, and arrive at the following definitions.

DEFINITION 10. (strict/superstrict distinguished utility models).
Let $DS = \{D^{p_1}_{s_1}(b_1|a_1), ..., D^{p_n}_{s_n}(b_n|a_n)\}$ be a desire specification. Let the u_i be (possibly void) local utility functions associated with $D^{p_i}_{s_i}(b_i|a_i)$.

A utility function u is called a strictly distinguished utility model of DS iff $u \models DS$ and, for $s_i \neq$ " $\geq 0''$, there are non-void u_i with $u = u_1 + ... + u_n$.

A utility function u is called a superstrictly distinguished utility model of DS iff $u \models DS$ and there are distinguished utility models u_i of $D^{p_i}_{s_i}(b_i|a_i)$ with $u = u_1 + ... + u_n$.

$\mathrm{Mod}_{sd}(DS), \mathrm{Mod}_{ssd}(DS)$, $\mid\!\sim_{sd}$, *and* $\mid\!\sim_{ssd}$ *are defined as their counterparts.*

Distinguishedness is our weakest notion of preference over utility models. Strict distinguishedness reflects the proposal in (Lang, 1996). Superstrict distinguishedness is an even more consequent implementation of that idea. Whereas strict and superstrict distinguishedness support stronger conclusions - by restricting the set of preferred models - they block redundant desires, to different degrees. This is shown by the following examples which illustrate the relations between different types of distinguished models. The first one clarifies the difference between distinguished and strictly distinguished models.

EXAMPLE 5. $DS = \{D^1_{\geq 1}(a|\top), D^1_{\geq 1}(a \wedge b|\top)\}$.
The distinguished utility models of $D^1_{\geq 1}(a \wedge b|\top)$ are also distinguished models of DS. Therefore, through distinguishedness, we cannot nonmonotonically infer from DS that $a \wedge \neg b$-worlds are automatically preferred to $\neg a$-worlds, because the worlds in $\neg a \vee \neg b$ may have uniform utility in some distinguished model. That is, $DS \not\mid\!\sim_d D^1_{>0}(a|\neg(a \wedge b))$.

On the other hand, the strictly distinguished models of $D^1_{\geq 1}(a \wedge b|\top)$ are not strictly distinguished models of DS, because in each such preferred model of DS, the $\neg a$-worlds get less utility than the $a \wedge \neg b$-worlds. In fact, by strictness, we are not allowed to add the void local utility function for $D^1_{\geq 1}(a|\top)$. Consequently, $DS \mid\!\sim_{sd} D^1_{>0}(a|\neg(a \wedge b))$.

The second example points at the difference between distinguished and superstrictly distinguished models.

EXAMPLE 6. $DS = \{D^1_{\geq 1}(a|\top), D^1_{\geq 2}(a|\top)\}$.
The strictly distinguished utility models of $D^1_{\geq 2}(a|\top)$ are exactly the strictly distinguished models of DS. Therefore, using strict distinguishedness, we cannot infer that the gap between a-worlds and $\neg a$-worlds is

at least 3. For instance, u with $u(w) = -2$ over $\neg a$ and $u(w) = 0$ over a is a strictly distinguished model of DS. That is $DS \not\mid\!\sim_{sd} D^1_{\geq 3}(a|\top)$. This conclusion is however supported in the context of superstrict distinguishedness, i.e. $DS \mid\!\sim_{ssd} D^1_{\geq 3}(a|\top)$, where not just local utility functions, but local distinguished utility models are summed up.

It is up to the user to decide whether in his application, these additional assumptions make sense. Our favourite choice criterion is distinguishedness, respectively minimal distinguishedness. It follows from the definition of strict distinguishedness that it cannot be naively strengthened to minimal strict distinguishedness. Minimal superstrict distinguishedness, on the other hand, doesn't cause any problems, at least if the loss of redundancy is accepted.

2.3. Some benchmark examples

The simple examples in the previous section have a technical flavour and illustrated amongst others subtle distinctions between distinguished, strictly distinguished and superstrictly distinguished utility functions. However, for many practical purposes, these notions do not make a difference. In this section, we consider several benchmark examples of reasoning about desires which are analogous for all types of distinguished utility models. The first two examples illustrate a notion of 'overriding' (Example 7-8) and the latter two illustrate the parameters (Example 9-10).

When the strengths of a desire specification are unknown, there are two ways to proceed. First and most obviously, we may assume strength > 0 for each desire of the desire specification, though as we discussed this has the drawback that minimal distinguished utility functions are not defined. Alternatively, we may stipulate the same strength for all the desires. In such cases we use by convention strength ≥ 1 for all desires, though any value different from 0 and ∞ would be acceptable. In technical terms, all these are structurally indistinguishable or isomorphically exchangable. In the following examples we concentrate on desire specifications with desires of strength ≥ 1, where minimal distinguished utility functions are known to exist. On the other hand, we pay particular attention to the desires of strength > 0 which are supported by the different kinds of distinguished utility models, i.e. which are non-monotonically entailed by the desire specification, because as explained in the following section we are often primarily interested in the resulting preferences.

In the examples in this paper, the default values for strength and polarity are respectively > 0 and 1 – thus we write $D_s(a|b)$ for $D^1_s(a|b)$,

$D^p(a|b)$ for $D^p_{>0}(a|b)$, and $D(a|b)$ for $D^1_{>0}(a|b)$. Let $\subset$ denote strict inclusion. The first example illustrates how more specific desires override more general conflicting desires.

EXAMPLE 7. (umbrella). *(from (*Boutilier, 1994*))*

1. *I prefer not to carry an umbrella;*

2. *If it rains, then I prefer to carry an umbrella.*

They are intuitively loss desires, thus $DS = \{D_{\geq 1}(\neg m|\top), D_{\geq 1}(m|r)\}$, *where* m *and* r *stand respectively for* umbrella *and* raining.

$u_{\neg m\|\top}(\omega)$	$-\alpha_1$ *iff* $\omega \models m$	0 *otherwise*
$u_{m\|r}(\omega)$	$-\alpha_2$ *iff* $\omega \models r \wedge \neg m$	0 *otherwise*

The (strict, superstrict) distinguished utility functions have the form $u = u_{\neg m|\top} + u_{m|r}$:

ω		$u(\omega)$	ω		$u(\omega)$
m	r	$-\alpha_1$	$\neg m$	r	$-\alpha_2$
m	$\neg r$	$-\alpha_1$	$\neg m$	$\neg r$	0

The constraints on α_1 *and* α_2 *are the following:*

$u \models D_{\geq 1}(\neg m\|\top)$	$\Leftrightarrow \max(-\alpha_2, 0) \geq 1 + \max(-\alpha_1, -\alpha_1)$	$\Leftrightarrow \alpha_1 \geq 1$
$u \models D_{\geq 1}(m\|r)$	$\Leftrightarrow -\alpha_1 \geq 1 - \alpha_2$	$\Leftrightarrow \alpha_2 \geq 1 + \alpha_1$

The unique minimal (strict, superstrict) distinguished utility function u *satisfying* DS *is given by* $u(m \wedge r) = u(m \wedge \neg r) = -1$, $u(\neg m \wedge r) = -2$, $u(\neg m \wedge \neg r) = 0$. $D_{\geq 1}(m|r)$ *takes precedence over* $D_{\geq 1}(\neg m|\top)$*, because it is more specific.*

The following example is borrowed from the literature on deontic logics (Belzer, 1986) and illustrates a more complex form of overriding based on specificity. In (van der Torre, 1994), two interpretations of the Reykjavik scenario are given; here we refer to the *mixed defeasibility and violability interpretation.*

EXAMPLE 8. (Reykjavik scenario).

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.*

$DS = \{D_{\geq 1}(\neg g|\top), D_{\geq 1}(\neg r|\top), D_{\geq 1}(g|r), D_{\geq 1}(r|g)\}$, *because intuitively they are loss desires. Local utility functions, distinguished utility models and constraints are as follows.*

$u_{\neg g\|\top}(\omega)$	$-\alpha_1$ *iff* $\omega \models g$	0 *otherwise*
$u_{\neg r\|\top}(\omega)$	$-\alpha_2$ *iff* $\omega \models r$	0 *otherwise*
$u_{g\|r}(\omega)$	$-\alpha_3$ *iff* $\omega \models r \wedge \neg g$	0 *otherwise*
$u_{r\|g}(\omega)$	$-\alpha_4$ *iff* $\omega \models g \wedge \neg r$	0 *otherwise*

ω		$u(\omega)$	ω		$u(\omega)$
g	r	$-\alpha_1 - \alpha_2$	$\neg g$	r	$-\alpha_2 - \alpha_3$
g	$\neg r$	$-\alpha_1 - \alpha_4$	$\neg g$	$\neg r$	0

$u \models D_{\geq 1}(\neg g\|\top)$	$\Leftrightarrow$	$0 \geq 1 + \max(-\alpha_1 - \alpha_2, -\alpha_1 - \alpha_4)$	$\Leftrightarrow$	$\alpha_1 + \alpha_2 \geq 1$
				$\alpha_1 + \alpha_4 \geq 1$
$u \models D_{\geq 1}(\neg r\|\top)$	$\Leftrightarrow$	$0 \geq 1 + \max(-\alpha_1 - \alpha_2, -\alpha_2 - \alpha_3)$	$\Leftrightarrow$	$\alpha_1 + \alpha_2 \geq 1$
				$\alpha_2 + \alpha_3 \geq 1$
$u \models D_{\geq 1}(g\|r)$	$\Leftrightarrow$	$-\alpha_1 - \alpha_2 \geq 1 - \alpha_2 - \alpha_3$	$\Leftrightarrow$	$\alpha_3 \geq 1 + \alpha_1$
$u \models D_{\geq 1}(r\|g)$	$\Leftrightarrow$	$-\alpha_1 - \alpha_2 \geq 1 - \alpha_1 - \alpha_4$	$\Leftrightarrow$	$\alpha_4 \geq 1 + \alpha_2$

The unique minimal (strict) distinguished utility function satisfying DS *is* $u(g \wedge r) = -1$, $u(g \wedge \neg r) = u(\neg g \wedge r) = -2$, $u(\neg g \wedge \neg r) = 0$. *It can be established by different values of the variables, such as* $\alpha_1 = \alpha_2 = 0.5$ *and* $\alpha_3 = \alpha_4 = 1.5$, *or* $\alpha_1 = 0.1$, $\alpha_2 = 0.9$, $\alpha_3 = 1.1$, *and* $\alpha_4 = 1.9$. *It is a strict model, because* $0 < \alpha_i$ *for all* i, *and it is not a superstrict model, because we have* $\alpha_1, \alpha_2 \not\geq 1$. *In the distinguished utility model, the desire* $D_{\geq 1}(g|r)$ *takes precedence over* $D_{\geq 1}(\neg g|\top)$, *and* $D_{\geq 1}(r|g)$ *over* $D_{\geq 1}(\neg r|\top)$, *because they are more specific.*

The following example taken from (Tan and Pearl, 1994a) is an extension of Example 3, with different strengths, and illustrates the strength parameter. Stronger desires override weaker desires in case of conflict.

EXAMPLE 9. (Healthy and wealthy). *Consider the desires*

1. *I desire to be healthy;*

2. *I desire to be wealthy.*

$DS = \{D_{\geq 2}(h|\top), D_{\geq 1}(w|\top)\}$, *because the first desire is more important or stronger than the second one. Note that for extracting preferences it does not make a difference whether we formalize these unconditional desires as loss or gain desires, because it only results in adding a constant value to each world. In other words it affects the absolute utility values but not the relative ones.*

$u_{h\|\top}(\omega)$	$-\alpha_1$ *iff* $\omega \models \neg h$	0 *otherwise*
$u_{w\|\top}(\omega)$	$-\alpha_2$ *iff* $\omega \models \neg w$	0 *otherwise*

ω	$u(\omega)$	ω	$u(\omega)$
$h \quad w$	0	$\neg h \quad w$	$-\alpha_1$
$h \quad \neg w$	$-\alpha_2$	$\neg h \quad \neg w$	$-\alpha_1 - \alpha_2$

$u \models D_{\geq 2}(h\|\top)$	$\Leftrightarrow \max(0, -\alpha_2) \geq 2 + \max(-\alpha_1, -\alpha_1 - \alpha_2)$	$\Leftrightarrow \alpha_1 \geq 2$
$u \models D_{\geq 1}(w\|\top)$	$\Leftrightarrow \max(0, -\alpha_1) \geq 1 + \max(-\alpha_2, -\alpha_1 - \alpha_2)$	$\Leftrightarrow \alpha_2 \geq 1$

The unique minimal (strict, superstrict) distinguished utility function u *satisfying* DS *is given by* $u(h \wedge w) = 0$, $u(h \wedge \neg w) = -1$, $u(\neg h \wedge w) = -2$, *and* $u(\neg h \wedge \neg w) = -3$. *Thus we have the intended conclusion that healthy and not wealthy is preferred to unhealthy and wealthy.*

The following example adapted from (Tan and Pearl, 1994b; Bacchus and Grove, 1996) illustrates the impact of the polarity parameter.

EXAMPLE 10. (Bananas). *Consider the desire*

1. *If you are alive, then you desire bananas.*

Let $DS_1 = \{D^1_{\geq 1}(b|a)\}$ *(loss) and* $DS_2 = \{D^0_{\geq 1}(b|a)\}$ *(gain). The minimal (strict, superstrict) distinguished utility models are as follows.*

	ω	$u(\omega)$	ω	$u(\omega)$	ω	$u(\omega)$
DS_1	$a \quad b$	0	$a \quad \neg b$	-1	$\neg a \quad \times$	0
DS_2	$a \quad b$	1	$a \quad \neg b$	0	$\neg a \quad \times$	0

We can non-monotonically derive the desire 'to be dead if you cannot eat bananas' from DS_1 *but not from* DS_2, *because only in the former the utility of 'being dead'* $\neg a$ *is equal to the utility of 'being alive and having bananas'* $a \wedge b$, *and consequently dying is the only way to escape the violation of the desire.*

$$DS_1 \mid\sim_{md} D(\neg a|\neg(a \wedge b)) \text{ and } DS_2 \not\mid\sim_{md} D(\neg a|\neg(a \wedge b))$$

*If we extend the desire to the following more intuitive set, not discussed in (*Tan and Pearl, 1994b*;* Bacchus and Grove, 1996*), then we get similar answers.*

1. *If you are alive, then you desire bananas.*

2. *You desire to be alive.*

$DS_1 = \{D^1_{\geq 1}(b|a), D^1_{\geq 1}(a|\top)\}$ *and* $DS_2 = \{D^0_{\geq 1}(b|a), D^1_{\geq 1}(a|\top)\}$*, because intuitively the second desire is a loss desire.*

	ω	$u(\omega)$	ω	$u(\omega)$	ω	$u(\omega)$
DS_1	$a\ b$	0	$a\ \neg b$	-1	$\neg a\ \times$	-1
DS_2	$a\ b$	1	$a\ \neg b$	0	$\neg a\ \times$	-1

$DS_1 \not\mid\!\sim_{md} D(a|\neg(a \wedge b))$ *and* $DS_2 \mid\!\sim_{md} D(a|\neg(a \wedge b))$

After the introduction of preference orders on worlds and background knowledge we illustrate in Example 13 more complex reasoning with different polarities.

In the technical examples in Section 2.1 and 2.2 we use the logic of desires to derive desires from a set of desires. However, in the benchmark examples in this section we used the logic in a different way: we used the logic to derive utility functions from a set of desires. In other words, we are primarily interested in the semantics of a set of desires. In the following section we define our second step, the derivation of preference relations from a set of utility functions. This is not defined in the logic anymore, but it is an additional definition.

3. Preference relations

In this section we show how to induce from a set of initial conditional desires the preference relation of the agent. The problem of the logic of desires defined in the previous section for decision making is that a desire specification induces a *set* of distinguished utility functions, whereas for decision making we would prefer a single utility function. Also for other applications like negotiation or cooperation it is inconvenient to deal with a set of utility functions. The question thus rises how we can summarize the information of this set of utility functions into a single structure, which we call the agent's preference relation. In this section we define this preference relation as a partial pre-ordering on the set of worlds. Thus, a set of desires induces a set of distinguished utility functions by adding up the utility losses and gains of the individual desires, and these distinguished utility functions induce a partial preference ordering on worlds. From a qualitative description we induce a quantitative description, from which we induce again a qualitative description. In Section 3.1 we define the preference ordering and in Section 3.2 we prove several properties of our construction.

3.1. Definitions

It seems straightforward to draw a preference relation on worlds from the distinguished utility functions of a desire specification. A world ω is at least as preferred as a world ω' if and only if for each distinguished utility function u, we have $u(\omega) \geq u(\omega')$. Note that obviously, $\geq_{DS}$ is a pre-ordering which in general is not complete.

However, for strict preference there are two options. Either we can say that ω is strictly preferred to ω' if w is at least as preferred as ω' but not vice versa, or we can state that a world ω is strictly preferred to a world ω' if and only if for each distinguished utility function u, we have $u(\omega) > u(\omega')$. Obviously the latter implies the former, but we illustrate in Example 11 that the former does not imply the latter and thus that the two are not equivalent. In this paper we use the first definition.

The preference relation is defined for our four types of distinguished utility functions. We write $\omega \geq^x_{DS} \omega'$ for 'ω is preferred to ω' w.r.t. DS for criterion x'. Our default choice for distinguishedness is the standard one, i.e., $x = d$. Accordingly, we abbreviate $\geq^d_{DS}$ (respectively $>^d_{DS}$, $\approx^d_{DS}$) by $\geq_{DS}$ (respectively $>_{DS}$, $\approx_{DS}$).

DEFINITION 11. *Let $DS = \{D^{p_1}_{s_1}(b_1|a_1), ..., D^{p_n}_{s_n}(b_n|a_n)\}$ be a consistent desire specification and $x \in \{d, md, sd, ssd\}$ be a distinguishedness criterion.*
$\omega \geq^x_{DS} \omega'$ iff for all x-distinguished u with respect to DS, we have $\omega \geq_u \omega'$.
$\omega >^x_{DS} \omega'$ iff $\omega \geq^x_{DS} \omega'$ and not ($\omega' \geq^x_{DS} \omega$).
$\omega \approx^x_{DS} \omega'$ iff $\omega \geq^x_{DS} \omega'$ and $\omega' \geq^x_{DS} \omega$.

The following example illustrates that the induced preference relation may be different for the different types of distinguished utility functions, and it also illustrates the two types of strict preference between worlds.

EXAMPLE 11. (Redundancy, continued). *The preference ordering induced by $DS = \{D(a|\top), D(a \wedge b|\top)\}$ is the following, if the criterion x is not md:*

$$[a, b]$$

$$\begin{array}{c} \downarrow \\ [a, \neg b] \\ \downarrow \\ [\neg a, b] \approx [\neg a, \neg b] \end{array}$$

When the criterion is md the preference ordering is as follows:

$$\begin{array}{c} [a, b] \\ \downarrow \\ [a, \neg b] \approx [\neg a, b] \approx [\neg a, \neg b] \end{array}$$

We thus have $[a, \neg b] >^d_{DS} [\neg a, b]$ *when we take all distinguished utility functions into account, but* $[a, \neg b] \approx^{md}_{DS} [\neg a, b]$ *if only the minimal distinguished utility functions are taken into account. That is, we loose a strict preference by reducing the set of utility functions taken into account. This is a direct consequence of our indirect definition of* $>_{DS}$ *in terms of* $\geq_{DS}$*. If we use the alternative definition that a world is strictly preferred to another world if it is strictly preferred in all the distinguished utility functions, then even for* x *is* d *we have that* $a \wedge \neg b$ *is not preferred to* $\neg a \wedge b$*.*

The following example illustrates similar results for the Reykjavik example.

EXAMPLE 12. (Reykjavik, continued). *Since we have no constraint between* $\alpha_2 + \alpha_3$ *and* $\alpha_1 + \alpha_4$*, the worlds* $[r, \neg g]$ *and* $[\neg r, g]$ *are not comparable for standard, strict or superstrict distinguishedness, and the preference ordering induced by* DS *is in all three cases the following:*

$$\begin{array}{ccc} & [\neg g, \neg r] & \\ & \downarrow & \\ & [g, r] & \\ \swarrow & & \searrow \\ [g, \neg r] & & [\neg g, r] \end{array}$$

However, the preference ordering for $x = md$ *is as follows, given that* Mod^{md}_{DS} *is the unique utility function given by* $u([\neg g, \neg r]) = 0$*,* $u([g, r]) = -1$*, and* $u([\neg g, r]) = u([g, \neg r]) = -2$*.*

$$\begin{array}{c} [\neg g, \neg r] \\ \downarrow \\ [g, r] \\ \downarrow \\ [g, \neg r] \approx [\neg g, r] \end{array}$$

3.2. Properties

In this section we show several properties of $\geq_{DS}$ worth mentioning. For all these results, we focus first on *standard distinguishedness* ($x = d$), for which proofs are made in detail, and we then discuss the other types of distinguishedness.

PROPOSITION 1. (monotonicity with respect to set inclusion).
Let $DS = \{D^{p_1}_{s_1}(b_1|a_1), \ldots, D^{p_n}_{s_n}(b_n|a_n)\}$ where strengths are positive (i.e., $s_i \neq [\geq 0]$, for every i). We abbreviate $D^{p_i}_{s_i}(b_i|a_i)$ by D_i. For a world ω, define $NonViol(\omega, DS) = \{D_i \in DS \text{ s.t. } \omega \models a_i \rightarrow b_i\}$ and $Sat(\omega, DS) = \{D_i \in DS \text{ s.t. } \omega \models a_i \wedge b_i\}$. Then

(i)if DS is a set of loss *desires,* i.e., *$p_i = 1$ for all i, then*
(a) $NonViol(\omega, DS) \subset NonViol(\omega', DS)$ implies $\omega <_{DS} \omega'$.
(b) $\omega \approx_{DS} \omega'$ iff $NonViol(\omega, DS) = NonViol(\omega', DS)$.

(ii)if DS is a set of gain *desires,* i.e., *$p_i = 0$ for all i, then*
(a) $Sat(\omega, DS) \subset Sat(\omega', DS)$ implies $\omega <_{DS} \omega'$.
(b) $\omega \approx_{DS} \omega'$ iff $Sat(\omega, DS) = Sat(\omega', DS)$.

Proof:

(i)(a) Assume $NonViol(\omega, DS)$ is strictly contained in $NonViol(\omega', DS)$ and $D^1_{s_i}(b_i|a_i) \in NonViol(\omega', DS) \setminus NonViol(\omega, DS)$. Let u be a distinguished utility function for DS, then $u = \sum_j u_j$ with $u_i(\omega) = 0$ if $\omega \models a_i \wedge b_i$, $u_i(\omega) = -\alpha_i$ if $\omega \models a_i \wedge \neg b_i$. Then $u(\omega') - u(\omega) \geq \alpha_i \geq 0$. Since this is true for any distinguished u, we have $\omega \leq_{DS} \omega'$ holds; now, it is always possible to choose a non-null value for α, which prevents $\omega' \leq_{DS} \omega$ from being true, hence we have $\omega <_{DS} \omega'$.

(i)(b) If $NonViol(\omega, DS) = NonViol(\omega', DS)$, the by construction any distinguished utility function for DS assigns the same utility to ω and ω', hence $\omega \approx_{DS} \omega'$ holds. Conversely, assume $NonViol(\omega, DS) \neq NonViol(\omega', DS)$ and let $NonViol(\omega, DS) \setminus NonViol(\omega', DS) = \{D_i, i \in I\}$, $NonViol(\omega', DS) \setminus NonViol(\omega, DS) = \{D_i, i \in J\}$ (at least of both sets I and J is not empty and furthermore $I \cap J = \emptyset$). Let u be a distinguished utility function for DS. $u(\omega) - u(\omega') = \sum_{i \in I} \alpha_i - \sum_{i \in J} \alpha_i$. It is always possible to choose the α_i's such that the above quantity is different from 0, which leads to $\omega \not\approx_{DS} \omega'$.

(ii) similar to (i). ♣

The converse of (i) does not hold (see Examples 7 and 8), and the converse of (ii) does not hold either. It can be checked easily that this result still holds with $x = sd$ and $x = ssd$. It does not hold for minimal distinguishedness, as it is shown by the following counterexample: let $DS = \{D^0_{\geq 1}(a|\top), D^0_{\geq 1}(b|\top)\}$. The minimal distinguished utility model u assigns $+1$ to both $[a, \neg b]$ and $[\neg a, b]$, thus $[a, \neg b] \approx^{md}_{DS} [\neg a, b]$ while $Sat([a, \neg b]) \neq Sat([\neg a, b])$. Finally, for the alternative stronger definition that a world is strictly preferred to another world if it is strictly preferred in all the distinguished utility functions, the proposition does not hold in general, but it holds for $x = sd$ and $x = ssd$.

PROPOSITION 2. (unconditional desires).
$DS = \{D^{p_1}_{s_1}(b_1|\top), ..., D^{p_n}_{s_n}(b_n|\top)\}$ *of desires 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 $\{b_i \mid \omega \models b_i\} \supseteq \{b_i \mid \omega' \models b_i\}$

Proof:

1. Assume $\wedge_{i=1..n} b_i$ is consistent, then there is a $\omega^* \models \wedge_{i=1..n} b_i$. Let u be any utility function defined by $u(\omega^*) > max_{i=1..n} s_i$ and $\forall \omega \neq \omega^*$, $u(\omega) = 0$. This utility function satisfies the constraints $u(b_i) \geq u(\neg b_i) \rhd s_i$ for all i, thus it satisfies DS, which proves the consistency of the latter.

2. Assume DS is consistent, then there is a utility function u satisfying the constraints $u(b_i) \geq u(\neg b_i) \rhd s_i$ for all i. Let ω be a world maximizing u, then necessarily $\omega \models b_i$ for all i, therefore $\wedge_{i=1..n}$ is consistent.

3. Assume DS is consistent and let $\omega^* \models \wedge_{i=1..n} b_i$. We can assume that all the b_i's are non-tautological (otherwise the corresponding desire can be removed). The local utility functions attached to the D_i's have the form $u_i(\omega) = \beta_i$ if $\omega \models b_i$ and $u(\omega) = -\alpha_i$ if $\omega \models \neg b_i$, therefore the distinguished utility functions for DS have the form

$$(*)\ u(\omega) = \sum_{i, \omega \models b_i} \beta_i - \sum_{i, \omega \models \neg b_i} \alpha_i = \sum_{i=1..n} \beta_i - \sum_{i, \omega \models \neg b_i} (\alpha_i + \beta_i)$$

with $p_i = \frac{\alpha_i}{\alpha_i + \beta_i}$ for all i. Let us consider the subset U of these distinguished utility functions satisfying furthermore $\alpha_i + \beta_i \rhd s_i$ for all i. Then, for any $u \in U$, the constraints $u(b_i) \rhd u(\neg b_i) + s_i$ are satisfied for all i, because for every i, the constraint $u(b_i) \rhd u(\neg b_i) + s_i$ is equivalent to $u(\omega^*) \geq u(\omega_i) + s_i$ with ω_i being the best world satisfying $\neg b_i$ (the existence of such a world is guaranteed by the fact that b_i is nontautological) and (*) implies that $u(\omega^*) - u(\omega_i) \geq$

$\alpha_i + \beta_i \rhd s_i$. Now, $\omega \geq_{DS} \omega'$ implies that $\forall u \in U$, $u(\omega) \geq u(\omega')$ which from (*) holds *only if* $\{b_i, \omega \models b_i\} \subseteq \{b_i, \omega' \models b_i\}$. The converse is a consequence of Proposition 1.♣

It can be checked easily that this result still holds with $x = sd$ and $x = ssd$. It does not hold for minimal distinguishedness, as it is shown by the same counterexample as above.

PROPOSITION 3. (independent desires).
Let $DS = \{D^{p_1}_{s_1}(b_1|a_1), ..., D^{p_n}_{s_n}(b_n|a_n)\}$ consistent 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 (with respect to DS) when $i(\omega)$ is defined and $\omega \models b_{i(\omega)}$, bad when $i(\omega)$ is defined and $\omega \models \neg b_{i(\omega)}$ and neutral when $i(\omega)$ is not defined. Then, $\omega \geq_{DS} \omega'$ holds iff one of these conditions holds:
1. ω good or neutral, and ω' neutral or bad.
2. $i(\omega) = i(\omega')$, ω and ω' both good.
3. $i(\omega) = i(\omega')$, ω and ω' both bad.
4. $i(\omega) \neq i(\omega')$, ω and ω' both good, $p_{i(\omega)} < 1$, $p_{i(\omega')} = 1$.
5. $i(\omega) \neq i(\omega')$, ω and ω' both bad, $p_{i(\omega)} = 0$, $p_{i(\omega')} > 0$.

Proof: Note first that the consistency of DS is equivalent to the consistency of $a_i \wedge b_i$ for each i. In this case, for each i we can take any (α_i, β_i) such that $p_i = \frac{\alpha_i}{\alpha_i + \beta_i}$ and the constraints will be obviously satisfied. Thus, the distinguished utility functions are all functions of the form

$$\begin{cases} u(\omega) = \beta_{i(\omega)} & if\ \omega\ is\ good \\ u(\omega) = 0 & if\ \omega\ is\ neutral \\ u(\omega) = -\alpha_{i(\omega)} & if\ \omega\ is\ bad \end{cases}$$

from which the result follows easily.♣

It can be checked easily that this result still holds with $x = sd$ and $x = ssd$. It does not hold for minimal distinguishedness, as it is shown by the following counterexample: let $DS = \{D^0_{\geq 1}(a|c), D^0_{\geq 1}(b|\neg c)\}$. The minimal distinguished utility model u assigns $+1$ to both $[a, b, c]$ and $[a, b, \neg c]$, thus $[a, b, \neg c] \approx^{md}_{DS} [a, b, c]$ while none of the five conditions is satisfied: indeed, both worlds are good, their indices are different, and point 4 is not verified by since both polarities are 0.

When introducing superstrict distinguishedness, we discussed sensitivity to repetitions of conditional desires. This point is important, since the way we use to distinguish utility functions makes use of additions. As expected, the sensitivity to repetition varies with the distinguishedness criterion. Namely:

PROPOSITION 4. (redundant desires). *Assume that there are two conditional desires of same polarity,* $D_i = D_s^p(b|a)$ *and* $D_j = D_{s'}^p(b'|a')$ *in* DS *such that* a *is logically equivalent to* a' *and* $a \wedge b$ *is logically equivalent to* $a' \wedge b'$.

1. *with standard, strict and minimal distinguishedness,* DS *generates the same sets of distinguished utility models as* DS^* *obtained from* DS *by replacing* $\{D_i, D_j\}$ *by the unique desire* $D^p_{max(s,s')}(b|a)$, *and as a consequence,* $\geq_{DS}$ *is identical to* $\geq_{DS^*}$.

2. *with superstrict distinguishedness,* DS *generates the same sets of distinguished utility models as* DS^{**} *obtained from* DS *by replacing* $\{D_i, D_j\}$ *by the unique desire* $D^p_{s+s'}(b|a)$, *and as a consequence,* $\geq_{DS}$ *is identical to* $\geq_{DS^{**}}$.

Proof:

1. assume that DS is consistent (otherwise the result is trivially satisfied) and suppose without loss of generality that $s_i \geq s_j$. Let $x \neq ssd$ be the distinguishedness criterion.
(i) let us consider a x-distinguished utility model $u = u_1 + \dots u_n$ of DS. We have to show that u is also a x-distinguished utility model of DS^*. Let us take $u'_k = u_k$ for every $k \neq i, j$, and $u'_i = u_i + u_j$. Then $u' = \sum_{k \neq j} u'_k = u$ is a x-distinguished utility model of DS^*.
(ii) let us consider a x-distinguished utility model $u' = u'_1 + \dots u'_{j-1} + u'_{j+1} + \dots + u'_n$ of DS^*. We have to show that u is also a x-distinguished utility model of DS. Let us take $u_k = u'_k$ for every $k \neq j$, and $u_i = u''j = \frac{1}{2}.u'_j$. Then $u = \sum_{k=1}^n u_k = u'$ is a x-distinguished utility model of DS.

2. let now consider the case of superstrict distinguishedness, i.e. $x = ssd$, and again let us assume that DS is consistent (otherwise the result is trivially satisfied). Without loss of generality, assume that $i < j$.
(i) let us consider a superstrictly-distinguished utility model $u = u_1 + \dots u_n$ of DS. We have to show that u is also a superstrictly-distinguished utility model of DS^{**}. We have
$$u_i + u_j(\omega) = \begin{cases} \beta + \beta' & \text{iff } \omega \models a \wedge b \\ 0 & \text{iff } \omega \models \neg a \\ -(\alpha + \alpha') & \text{iff } \omega \models a \wedge \neg b \end{cases}$$
with $p = \frac{\alpha}{\alpha+\beta} = \frac{\alpha'}{\alpha'+\beta'}$.
These last equalities yield $p(\alpha + \beta) = \alpha$ and $p(\alpha' + \beta') = \alpha'$, thus $p(\alpha + \alpha' + \beta + \beta') = \alpha + \alpha'$ and therefore $\frac{\alpha+\alpha'+\beta+\beta'}{\alpha+\alpha'} = p$. This implies easily that a local utility function u_k attached to $D^p_{s+s'}(b|a)$

in DS has exactly the same form as $u_i + u_j$ with the very same constraints, the rest being equal due to the logical equivalences. Hence $u = u_1 + \ldots + u_{i-1} + (u_i + u_j) + u_{i+1} + \ldots + u_{j-1} + u_{j+1} + \ldots + u_n$ is a superstrictly distinguished utility model of DS_{**}.
(ii) if $u = u_1 + \ldots + u_{i-1} + u_{i,j} + u_{i+1} + \ldots + u_{j-1} + u_{j+1} + \ldots + u_n$ is a superstrictly-distinguished utility model of DS^{**} then u can be rewritten $u = u_1 + \ldots + u_{i-1} + \lambda.u_{i,j} + u_{i+1} + \ldots + (1-\lambda).u_{i,j} + u_{j-1} + u_{j+1} + \ldots + u_n$ where $\lambda = \frac{s_i}{s_i + s_j}$, which shows that u is also a superstrictly-distinguished utility model of DS.

♣

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

More generally, when two logically equivalent desires with *distinct* polarities p and p' (such as $p < p'$ without loss of generality), coexist in DS, then $p'' = \frac{\alpha+\alpha'+\beta+\beta'}{\alpha+\alpha'} \in [p, p']$ and p' tends to p when $\frac{\alpha+\beta}{\alpha'+\beta'}$ tends to $+\infty$, and to p' when $\frac{\alpha+\beta}{\alpha'+\beta'}$ tends to 0. Thus, DS behaves as if there was a single desire with an *interval polarity* (p, p').

PROPOSITION 5. (specificity). *Let* $DS = \{D^p_s(b|a), D^{p'}_{s'}(c|a')\}$ *with* $\vdash a' \rightarrow a$, $\not\vdash a \rightarrow a'$, *and* $b \wedge c$ *inconsistent. Then,* whatever s and s', *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 $(\beta, -\alpha)$ to $D^p_s(b|a)$ and $(\beta', -\alpha')$ to $D^{p'}_{s'}(c|a')$; distinguished utilities have the form:
$u(\omega) = \beta - \alpha'$ if $\omega \models \neg a' \wedge b \wedge \neg c$
$u(\omega) = \beta' - \alpha$ *if* $\omega \models \neg a' \wedge \neg b \wedge c$
$u(\omega) = -(\alpha + \alpha')$ if $\omega \models \neg a' \wedge \neg b \wedge \neg c$
$u(\omega) = \beta$ if $\omega \models \neg(a \wedge \neg a') \wedge b$
$u(\omega) = -\alpha$ if $\omega \models \neg(a \wedge \neg a') \wedge \neg b$
$u(\omega) = 0$ if $\omega \models \neg a \wedge \neg a'$
The constraints $u(a \wedge b) \rhd u(a \wedge \neg b) + s$ and $u(a' \wedge c) \rhd u(a' \wedge \neg c) + s'$ yield respectively $\beta \rhd \beta' - \alpha + s$ and $\beta' - \alpha \rhd \beta - \alpha' + s'$. This second inequality implies $\beta' - \alpha > \beta - \alpha'$, i.e., $u(a' \wedge \neg b \wedge c) > u(a' \wedge b \wedge \neg c)$.
♣

This result extends easily to all other types of distinguishedness.

The following proposition generalizes Example 4 in Section 3.1.

PROPOSITION 6. (absence of drowning effect).
Let $DS = \{D^p_s(b|a), D^{p'}_{s'}D(c|a'), D^{p''}_{s''}(e|a)\}$ *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') satify not only c but also e, thus the precedence of $D_{s'}^{p'}(c|a')$ over $D_s^p(b|a)$ does not inhibit $D_{s''}^{p''}(e|a)$: there is no "drowning effect" (*Benferhat et al., 1993*).*

The proof (omitted) is similar to the proof of Proposition 5 and holds for all types of distinguishedness.

4. Desires and knowledge

Up to now, we have only considered desire specifications, concerned with ideal worlds, and ignored the impact of knowledge about the real world. On a general, formal level, knowledge allows us to restrict the set of possible worlds to a subset of W. However, here it is important to distinguish between three kinds of knowledge.

- *Background knowledge* is meant to express which worlds are physically impossible.
- *Contingent knowledge (facts)* is meant to tell us which (physically possible) worlds do not correspond to the actual state of affairs.
- *Feasibility knowledge* is meant to tell us which (physically possible) worlds the agent is able to reach.

The distinction between background knowledge and contingent knowledge is well known from modal and default conditional logic. It has also been made by van der Torre (van der Torre, 1994) for the treatment of violated obligations in deontic logic. Because preference relations are usually meant to be relevant across situations, in addition to the desire specification, only background knowledge should be taken into account when choosing utility distributions or defining a preference relation over the set of worlds. Contingent knowledge only comes in when the resulting preferences are exploited for taking a decision, more precisely, when looking at the preferred worlds among those which are feasible from the actual situation (more details are given in Section 4.2). Computing the preference relation may therefore be seen as a kind of "compilation", i.e. it is the same whatever the contingent knowledge is.

The distinction between background knowledge and feasible knowledge is well known from logic of action, because it is generally not the case that all physically possible worlds are *feasible from the actual situation.* We therefore introduce in Section 4.2 a third notion, namely feasibility (ability): from the initial situation, there are some decisions the agent can make; each possible decision enables us to reach another – or the same – world, which 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 by performing a suitable action. This is where contingent knowledge will be taken into account.

4.1. World knowledge

DEFINITION 12. (background knowledge, physically possible worlds).

Let DS be a desire specification and BK (the background knowledge base) a finite set of formulas. Worlds satisfying BK are called physically possible. $\langle DS, BK\rangle$ is referred to as a background specification. The semantic framework changes only insofar as in the context of $\langle DS, BK\rangle$, we restrict utility functions to the physically possible worlds, whereas the truth conditions stay unaffected. That is,

$$u \models \langle DS, BK\rangle \quad \textit{iff} \quad Dom(u) = \{\omega \mid \omega \models BK\} \textit{ and } u \models DS.$$

Local utility functions, preference relations, and all kinds of distinguishedness with respect to BK are defined accordingly.

All results of Section 3.2 are easily generalizable to the case where background knowledge is taken into account. We illustrate its impact by the following example, which also shows that constraints between utility losses are not only used to resolve conflicts (as in Example 7).

EXAMPLE 13. (transitivity). *Suppose, there are two desires, notably that going to a party (p) is preferred over going to the cinema (c), and going to the cinema is preferred over staying home (h). Intuitively, we have three variables representing different ways to spend the evening. These are mutually exclusive and exhaustive, which can be expressed with background knowledge.*

$-DS = \{D^p_{>0}(p|p \vee c), D^p_{>0}(c|c \vee h)\}$

$-BK = \{p \vee c \vee h, \neg(p \wedge c), \neg(p \wedge h), \neg(c \wedge h)\}$

The distinguished utility functions now have the following form:

ω			$u(\omega)$
p	$\neg c$	$\neg h$	β_1
$\neg p$	c	$\neg h$	$\beta_2 - \alpha_1$
$\neg p$	$\neg c$	h	$-\alpha_2$

The constraints lead to $\alpha_1 > \beta_2 - \alpha_1 > -\alpha_2$*, thus* p *is preferred to* h *whatever the polarity. If we take* $p = 0.5$ *then the constraints are equivalent to* $\alpha_1 > \alpha_2$*. Now, let us also consider the desire that staying at home is preferred over going to work, and encode these desires with explicit constant strength bounds.*

$$-DS = \{D^p_{\geq 1}(p|p \vee c), D^p_{\geq 1}(c|c \vee h), D^p_{\geq 1}(h|h \vee w)\}$$

$$-BK = \{p \vee c \vee h, \neg(p \wedge c), \neg(p \wedge h), \neg(c \wedge h)\}$$

It follows that the distinguished utility functions and the minimal distinguished utility functions have the following form, which of course can be restricted further by the strength constraints:

ω				$u(\omega)$	$u(\omega)(\min)$
p	$\neg c$	$\neg h$	$\neg w$	β_1	β_1
$\neg p$	c	$\neg h$	$\neg w$	$\beta_2 - \alpha_1$	$\beta_1 - 1$
$\neg p$	$\neg c$	h	$\neg w$	$\beta_3 - \alpha_2$	$\beta_1 - 2$
$\neg p$	$\neg c$	$\neg h$	w	$-\alpha_3$	$\beta_1 - 3$

The polarity does not influence the construction of the preference relation on worlds, because the relative ordering stays constant. If $p = 0.5$*, we get* $\beta_1 = 2$*.*

More examples illustrating the polarity and strength parameter are given in (van der Torre and Weydert, 2000).

4.2. Feasible worlds and preferred decisions

At this point we have to choose an action model. Since the paper focuses on preference representation rather on action theories, we prefer to stick here to a simple action model where actions are deterministic and ramification is not handled. Our action model is inspired from Boutilier's (Boutilier, 1994): the set of propositional variables Var is partitioned into two classes: controllable variables ($ContrVar$), whose truth value may be fixed or changed by the agent, and uncontrollable variables ($UncontrVar$), whose truth value is fixed by the outside world. To any controllable variable x, there exist two *atomic decisions* $do(x)$ and $do(\neg x)$ whose effects are that the truth value of x is fixed, or changed, to respectively `true` or `false`. A *consistent, complete decision* $\vec{d}$ is a set of atomic decisions

(i) containing exactly one of both atomic decisions $do(x)$ and $do(\neg x)$ for every $x \in ContrVar$ and

(ii) consistent with BK.

Lastly, some of these consistent complete decisions may not be applicable in some contexts: for this we introduce applicability constraints of the form

`if` φ `then` δ `unapplicable`

where δ is a conjunction of atomic actions and φ is any propositional formula (which may even contain controllable variables – see the Reykjavik example further on). This syntax is inspired by `impossible` δ `if` φ in action languages such as (Gelfond and Lisfchitz, 1993).

We say that a complete, consistent decision $\vec{d}$ is *available* in the initial state $Init$ iff it is consistent with $\{\neg\delta \mid$ `if` φ `then` δ `unapplicable` $\in Constr$ and $Init \models \varphi\}$. The set of all consistent decisions available in the initial state $Init$ w.r.t. a set of constraints $Constr$ is denoted by $Dec(Init, Constr)$.

DEFINITION 13. (decision problem). *A decision problem $\mathcal{P}$ consists of*

- *a* partition $(ContrVar, UncontrVar)$ *of the propositional variables;*
- *a* desire specification DS*;*
- *a* background knowledge base BK*;*
- *a set $Constr$ of* applicability constraints `if` φ `then` δ `unapplicable` *as above;*
- *a* complete initial situation $Init$*, namely, $Init$ is a propositional formula such that for any uncontrollable variable y, we have either $Init \models y$ or $Init \models \neg y$.*

Intuitively, the initial situation $Init$ consists of

1. a complete assignment of all uncontrollable variables, and
2. a partial assignment of the uncontrollable variables representing the choices already made by the agent (see Examples 6, 7).

By default, no controllable variable is initially assigned. The case where the agent has an incomplete knowledge of the initial state of the world will not be considered in this paper.

DEFINITION 14. (effect of a decision on the initial state; feasible worlds). *Let $\vec{d}$ be a consistent decision available in the initial state $Init$. The effect of $\vec{d}$ on $Init$, denoted $Res(\vec{d}, Init)$, is defined as follows:*

•$Res(\vec{d}, Init)$ assigns each controllable variable x to `true` *if $\vec{d}$ contains $do(x)$ and to* `false` *if $\vec{d}$ contains $do(\neg x)$;*

•$Res(\vec{d}, Init)$ assigns each uncontrollable variable as in $Init$, i.e., for each uncontrollable variable y, we have $Res(\vec{d}, Init) \models y$ if and only if $Init \models y$ and $Res(\vec{d}, Init) \models \neg y$ if and only if $Init \models \neg y$.

The set $\{Res(\vec{d}, Init) \mid \vec{d} \in Dec(Init, Constr)\}$ is called the set of feasible worlds *from $Init$ and is denoted by $FW(BK, Init, Constr)$.*

Note that in $Init$ the uncontrollable variables need not to be assigned a truth value. Only the controllable variables must have a fixed truth value (*complete initial state*). $Init$ being complete and actions being deterministic means that the agent has a complete knowledge of the initial world regarding to the uncontrollable atoms and thus that the set of feasible worlds $FW(BK, Init, Constr)$ is known.

What we want to do now 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. *By default, and for the sake of simplicity, from now on the preference relation $\langle DS, BK \rangle$ is induced from (standard) distinguished utility models.*

DEFINITION 15. (preferred feasible worlds and decisions).
Let $\mathcal{P} = \langle ContrVar, UncontrVar, DS, BK, Constr, Init \rangle$ be a decision problem and let $\vec{d}, \vec{d'} \in \mathcal{D}(Init)$. We say that $\vec{d} \sqsupseteq_{\mathcal{P}} \vec{d'}$ iff $Res(\vec{d}, Init) \geq_{\langle DS,BK \rangle} Res(\vec{d'}, Init)$. A decision $\vec{d}$ is a preferred decision iff $\vec{d}$ is $\geq_{\langle DS,BK \rangle}$-maximal in $Dec(Init, Constr)$; in this case $Res(\vec{d}, Init)$ will be called a preferred feasible world.

Equivalently, ω is a preferred feasible world iff ω is $\geq_{\langle DS,BK \rangle}$-maximal in $FW(BK, Init, Constr)$. 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, $\sqsupseteq_{\mathcal{P}}$ is generally not complete either. Thus there may be several incomparable preferred decisions.

4.3. Examples

The following example illustrates an instance of what Horty (Horty, 1994) calls *overridden rules.*

EXAMPLE 14. (asparagus).

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.* d *is considered uncontrollable, because once you have started participating to the dinner, the truth of* d *cannot be changed. In agreement with intuition, 1. and 2. are encoded by loss desires, while 3. is background knowledge (*$[\neg d, a]$ *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. Thus,*

$$\begin{aligned}&ContrVar = \{f\};\ UncontrVar = \{d, a\};\\ &DS = \{D^1_{\geq 1}(\neg f|d), D^1_{\geq 1}(f|a)\};\\ &BK = \{a \rightarrow d\};\ Init = \{a\};\ Constr = \top.\end{aligned}$$

Let α_1 *and* α_2 *be associated respectively with* $D^1_{\geq 1}(\neg f|d)$ *and* $D^1_{\geq 1}(f|d \wedge a)$*. The distinguished utility functions have the following form:*

$$\begin{array}{ll} u_{DS,BK}([d,a,f]) = -\alpha_1 & u_{DS,BK}(da\bar{f}) = -\alpha_2 \\ u_{DS,BK}([d,\neg a,f]) = -\alpha_1 & u_{DS,BK}([d,\neg a,\neg f]) = 0 \\ u_{DS,BK}([\neg d,\neg a,\times]) = 0 & u_{DS,BK}([\neg d,a,\times]) \textit{ is undefined.} \end{array}$$

The constraints expressed by the conditional desires entail that $\alpha_2 > \alpha_1 > 0$ *(*$D^1_{\geq 1}(f|a)$ *is more specific than* $D^1_{\geq 1}(\neg f|d)$*). Hence, we get the following preference relation on the set of physically possible worlds:*

$$\begin{array}{c} [d,\neg a,\neg f] \approx [\neg d,\neg a,f] \approx [\neg d,\neg a,\neg f] \\ \downarrow \\ [d,a,f] \approx [d,\neg a,f] \\ \downarrow \\ [d,a,\neg f] \end{array}$$

Now, since $Init = \{a\}$*, the set of feasible worlds is* $\{[d,a,f],[d,a,\neg f]\}$*. Since* $[d,a,f] >_{DS,BK} [d,a,\neg f]$*, the only preferred feasible world is* $[d,a,f]$*. Hence the only preferred decision is* $do(f)$*,* 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 $Init = d \wedge \neg a$*, the preferred feasible world would then be* $[d, \neg a, \neg f]$ *and the preferred decision would be* $do(\neg f)$.

EXAMPLE 15. (asparagus and napkin).
Same as Example 14 plus the conditional desire $D^1_{\geq 1}(n|d)$ *(n means* you eat with a napkin – *controllable).*
The new desire is again a loss desire.

$$DS = \{D^1_{\geq 1}(\neg f|d), D^1_{\geq 1}(f|a), D^1_{\geq 1}(n|d)\};$$
$$BK = \{a \rightarrow d\};\ Init = \{a\}$$

The preferred feasible world is $[d, a, f, n]$*, and consequently the preferred decision is* $\{do(f), do(n)\}$.

EXAMPLE 16. *(where a desire is more specific than a set of 2 desires): p uncontrollable, a and b controllable.*

$$DS = \{D^1_{\geq 1}(a|\top), D^1_{\geq 1}(b|\top), D^1_{\geq 1}(\neg a \wedge \neg b|p)\};$$
$$BK = \emptyset;\ Init = \{p\}.$$

Let α_1, α_2 *and* α_3 *be associated respectively with* $D^1_{\geq 1}(a|\top)$, $D^1_{\geq 1}(b|\top)$ *and* $D^1_{\geq 1}(\neg a \wedge \neg b|p)$.

$u_{\langle DS,BK\rangle}([p, a, b]) = -\alpha_3$	$u_{\langle DS,BK\rangle}([p, a, \neg b]) = -(\alpha_2 + \alpha_3)$
$u_{\langle DS,BK\rangle}([p, \neg a, b]) = -(\alpha_1 + \alpha_3)$	$u_{\langle DS,BK\rangle}([p, \neg a, \neg b]) = -(\alpha_1 + \alpha_2)$
$u_{\langle DS,BK\rangle}([\neg p, a,]b) = 0$	$u_{\langle DS,BK\rangle}([\neg p, a, \neg b]) = -\alpha_2$
$u_{\langle DS,BK\rangle}([\neg p, \neg a, b]) = -\alpha_1$	$u_{\langle DS,BK\rangle}([\neg p, \neg a, \neg b]) = -(\alpha_1 + \alpha_2)$

The constraints induce that $\alpha_3 \geq 1 + \alpha_1 + \alpha_2 > \alpha_1 + \alpha_2$*, which means that* $D^1_{\geq 1}(\neg a \wedge \neg b|p)$ is more specific than the set of desires $\{D^1_{\geq 1}(a|\top), D^1_{\geq 1}(b|\top)\}$. *Since* $Init = \{p\}$*, the set of feasible worlds is* $\{[p, a, b], [p, a, \neg b], [p, \neg a, \neg b], [p, \neg a, \neg b]\}$. *Now, it can be checked that the preferred feasible world is* $[p, \neg a, \neg b]$, i.e., $D^1_{\geq 1}(\neg a \wedge \neg b|p)$ *overwhelms the set of two desires* $\{D^1_{\geq 1}(a|\top), D^1_{\geq 1}(b|\top)\}$. *The preferred decision is thus* $\{do(\neg a), do(\neg b)\}$.

EXAMPLE 17. (Forrester paradox).

1. You should not kill;

2. If you kill, you should do it gently.

1. and 2. are encoded as loss desires. Note that Init is complete since $UncontrVar = \emptyset$.

$DS = \{D^1_{\geq 1}(\neg k|\top), D^1_{\geq 1}(g|k)\}$ *(g and k are controllable).*
$BK = \{g \to k\}$; $Init = \top$

Let α_1 *and* α_2 *be associated respectively with* $D^1_{\geq 1}(\neg k|\top)$ *and* $D^1_{\geq 1}(g|k)$.
We get $u_{\langle DS,BK\rangle}([\neg k, \neg g]) = 0$; $u_{\langle DS,BK\rangle}([k, g]) = -\alpha_1$; $u_{\langle DS,BK\rangle}([k, \neg g]) = -(\alpha_1 + \alpha_2)$.
The preferred decision is $\{do(\neg k), do(\neg g)\}$.

Assume now that the agent has received from his hierarchy the order to kill. Disobeying the hierarchy induces a maximal penalty, which translates by the desire $D^1_{\geq +\infty}(k|\top)$, *to be added to* DS. *Now the preferred decision becomes* $\{do(k), do(g)\}$.

A variation of this example consists of considering as possible the situation where the individual to be killed is already dead. This translates by adding the new uncontrollable variable d (already dead), and $Constr = \{$ `if` d `then` $do(k)$ *inapplicable,* `if` d `then` $do(g)$ *inapplicable*$\}$ *(it is not possible to kill, nor a fortiori to kill gently, a person who is already dead). Then the* only *available decision (and thus the preferred one whatever the orders of the hierarchy) is* $\{do(\neg k), do(\neg g)\}$.

EXAMPLE 18. (Reykjavik scenario, continued).

DS as in Example 8;
g , r are controllable; $BK = \top$;
$Constraints =$ $\{$ `if` g `then` $do(\neg g)$ `inapplicable`,
`if` r `then` $do(\neg r)$ `inapplicable`$\}$;

The constraints express that once g (respectively r) is true, i.e., once Gorbachev (respectively Reagan) knows the secret, it is not possible to fix the value of g (respectively r) to `false`.

The preference ordering is shown in Example 12. When $Init = \top$, *the preferred decision is* $\{do(\neg r), do(\neg g)\}$. *Now, if the secret has already been told to Reagan (*$Init = r$*), the previous preferred decision is not any longer available, and the new preferred decision becomes* $(do(r), do(g))$.

5. Further research

In this paper we only considered the situation where desires are imposed upon an agent by another one, as in multi agent systems. In particular, the picture we have in mind is that of a robot who, given our requirements (imperatives), tries to figure out the set of admissible utility functions, calculates the expected utilities and acts accordingly. We do not impose our complete preference ordering on the robot, but only a simple approximation of it, for computational reasons and because we do not know it ourselves. In our semantic interpretation, we concentrate exclusively on the role of expressing desirability –as opposed to intentionality (commitment to pursue) (Doyle, 1980)– recognizing that the result is only a partial account of the use of goals in planning systems.

Even in this simple situation we have to take several other concepts into account. The relation between desires, defaults, expectations, facts and decisions is represented in Figure 1. The input of the one-shot (i.e. static) decision problem consists of desires (or goals), defaults (or beliefs), expectations and facts. The input is transformed – non-monotonically – into (qualitative abstractions of) utility and probability distributions, which can be combined in an expected utility ordering suggesting decisions according to some decision rule like maximum utility. In this paper we have focussed on the relation between desires and utility functions.

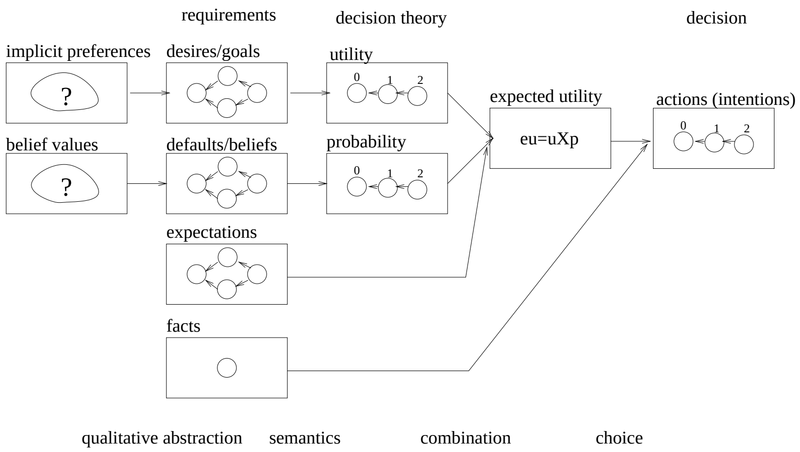

Figure 1. Desires, defaults, expectations and decisions

The main problem of this figure is the combination of desires and defaults for expected utility considerations. Defaults are usually interpreted as constraints over order-of-magnitude abstractions of probability functions, thus to combine these defaults with desires it seems that first we have to interpret desires as constraints over order-of-magnitude abstractions of utility functions and second we have to assume that the two abstractions are of the same order-of-magnitude.

This is the approach followed by e.g. Pearl (Pearl, 1993) and Dubois and Prade (Dubois and Prade, 1995), who call the latter assumption the commensurability assumption.

Decision theory does not only provide a semantic framework for conditional desires, but also a more general setting for planning. In further research we therefore consider how our logic can be used for planning, and in which way it has to be extended. A plan is a sequence of action types with different costs and plausibilities of success. So, all we need – in theory – is decision theory and an intelligent search strategy for such chains of tasks, the goal being to maximize the expected utility. Haddawy and Hanks (Haddawy and Hanks, 1992) observe the distinction between symbolic planning and decision theory, where the latter provides a normative model of choice under uncertainty, but offers no guidance as to how the planning options are to be generated. An important problem is that in practice, the utilities and probabilities are not known, and we therefore cannot directly exploit the tools of decision theory.

6. Related work

In this section we discuss related work from decision theory, agent theory and planning. Since our approach also works on many examples from deontic logic, in Section 6.2 we discuss some links between conditional desires and dyadic obligations in deontic logic.

6.1. Preferences

Modeling preferences of rational agents has been tackled for long by decision theory with utility-based or relational models, but generally without focusing on representational issues. For example, all possible states with their utility value are represented explicitly. On the contrary, AI languages for knowledge representation can also be used for preference modeling, to which they offer a way to represent preferences compactly and implicitly.

In the agency literature there is a lot of research on logics of desires and goals in a modal possible worlds framework, see e.g (Cohen and Levesque, 1990; Rao and Georgeff, 1991). An advantage of this simple formalization is that it easily captures the relation between for example beliefs, desires and intentions. On the other hand, as argued by Pearl (Pearl, 1993), it obscures other relations, such as for example the relation between desires and actions. Moreover, as argued by Boutilier (Boutilier, 1994), it is difficult to formalize context-sensitive or conditional goals. In addition, we think that an approach from first (decision-theoretic) principles is necessary to make underlying assumptions explicit, to give a satisfactory account of non-monotonic reasoning about preferences, and to analyze different types of conflicts (for which we can extend our decision-theoretic formalism with notions from multi attribute utility theory (Keeney and Raiffa, 1976; Bacchus and Grove, 1996)).

In the context of *qualitative* decision theory recently several *logics* for desires and goals have been proposed, which follow the thesis of Wellman and Doyle (Doyle and Wellman, 1991, p.698):

> The relative preference over the possible results of a plan constitutes the fundamental concept underlying the objectives of planning and decision making, with desires and goals serving as a computationally useful *partial specification or heuristic approximation of these preferences* (Doyle, 1991).

Our approach is complementary to Boutilier's (Boutilier, 1994): while he focuses only on the definition of optimal actions from a given preference relation, we also focus on the practical generation of this preference relation. Boutilier also 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 (Tan and van der Torre, 1996) about the two-phase treatment of violated obligations. Other approaches to qualitative decision theory include (Doyle and Wellman, 1991), (Tan and Pearl, 1994b) (where conditional desires are interpreted very differently, via a *ceteris paribus* assumption) and (Dubois and Prade, 1995) who give a possibility theory based view of decision theory.

6.2. Desires vs obligations

Two formalisms related to the logic of desires are *the logic for qualitative decision theory* (QDT) and *deontic logic*, because they distinguish between what *ideally* is the case (obligations and desires) from what *actually* is the case (facts). Conditional desires ("if a then I ideally would like b to be true") are similar to conditional obligations $O(b|a)$ of dyadic deontic logics ("if a is true, then b should be true"), in particular when these deontic logics are based on a preference-based semantics (see (Makinson, 1993; van der Torre and Tan, 1999a) for a survey) : "$O(b|a)$ holds iff b holds in all preferred a-worlds (Hansson, 1971)". Dyadic deontic logics were developed to handle so-called "contrary-to-duty (CTD) obligations" (see Examples 8, 12, 14, 17, 18). Defeasible obligations are handled in a non-monotonic framework by Horty (Horty, 1994; van der Torre, 1994). Alternative approaches to the handling of CTD obligations were proposed by Prakken and Sergot (Prakken and Sergot, 1996), by Cholvy and Cuppens (Cholvy and Cuppens, 1995) 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 van der Torre and Tan (van der Torre and Tan, 1999b) 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 from loss desires gives the intended results on the examples of CTD-obligations taken from the deontic logics literature. Thus, our work on loss desires 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:

1. In deontic logics, obligations are generally considered exogenous (they are imposed by a legal or moral code) while desires in logics for QDT are endogenous (directly specified by the agent). Note that this difference should not necessarily lead to a need of different treatments for obligations and desires (as also argued by Boutilier (Boutilier, 1994) and Tan & van der Torre (Tan and van der Torre, 1996)).

2. More importantly, the main purpose of a deontic logic is deriving new obligations (and permissions) from an initial specification, while QDT focuses on the search for optimal acts and 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") to find the best possible actions (An alternative logical approach to "goalness" has been proposed in (Wainer, 1994)). 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 (Makinson, 1993) for a survey). Our approach has the advantage of simplicity since we avoid the multiplication of modal operators, using instead a notion of controllability.

The relation between decision making and normative reasoning has also raised the following two questions.

The practical problem: how do norms influence decision making? The question here is whether norms influence one's decision making only via the associated rewards and penalties, or also directly. Different types of agents can be defined, for example agents that only maximize their own utility and agents that follow the norms of the society (Castelfranchi et al., 2000). In the latter case, desires reflect individual behavior whereas social norms represent social group behavior (Jennings and Campos, 1997).

The philosophic problem: is normative reasoning a kind of decision making? For example, von Wright (von Wright, 1983) argued that there is no genuine logic of norms and that the norm giving activity can only be judged under various aspects and standards of rationality.

7. Conclusion

In this paper we have proposed different procedures to induce the preference relation of the agent from a set of initial conditional desires and different types of knowledge. A set of desires induces a set of distinguished utility functions by weighting and adding up the utility losses and gains of the individual desires, and these distinguished utility functions induce a qualitative partial preference ordering on worlds. In the most complicated procedure, we use three different notions:

- *desirability*: some worlds are more desirable than others for the agent. This notion only concerns the preferences of the agent, only, and has nothing to do with the actual state of affairs;

- *physical possibility*: some worlds are physically possible, some are not. This notion concerns the outside world and has nothing to do with preference (except that it fixes the domain of the utility function);

- *feasibility*: some worlds can be reached by the agent, some cannot. This notion concerns the decisions the agent can make; each possible decision enables him to reach another – or the same – world, which must be physically possible.

In this most sophisticated procedure, we have turned the problem specification 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 by performing a suitable action. This is where contingent knowledge is taken into account.

We have used a *logical framework* to represent the derivation of the preference relation of the agent from a set of initial conditional desires. We assume that desires induce constraints on utility functions, which we represent with monotonic logic, and we assume that they induce a way to construct the distinguished or preferred utility functions, which we represent with non-monotonic extensions. Moreover, we have defined a particular way to *use* this logic. The input is a set of desires (and knowledge) and the output is a set of distinguished utility functions, i.e. the input is a set of formulas and the output is a set of models. We also added a definition to reduce the set of utility functions obtained from a set of desires to a single preference structure. The main advantage to use a logical framework in the derivation of preference orders is that it explicates subtle distinctions, like the different ways to add up and weigh losses and gains of utility, the different notions of redundancy, etc. The logical representation is compact and transparent, and can also be used for communicating desires between agents. Other advantages of the logical representation are that it facilitates the comparison between our representation of desires and related formalisms and non-monotonic extensions developed for e.g. defaults and obligations (e.g. (Makinson, 1993; Lehmann and Magidor, 1992)), and that it facilitates the introduction of more sophisticated mechanisms to determine the distinguished utility function (by incorporating mechanisms developed for non-monotonic reasoning, such as e.g. (Weydert, 1998)).

References

Bacchus, F. and A. Grove: 1996, 'Utility independence in a qualitative decision theory'. In: *Proceedings of Fifth International Conference on Knowledge Representation and Reasoning (KR'96)*. pp. 542–552.

Bell, J. and Z. Huang: 1997, 'Dynamic Goal Hierarchies'. In: *Intelligent Agent Systems, Theoretical and Practical Issues*. pp. 88–103.

Belzer, M.: 1986, 'A logic of deliberation'. In: *Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI'86)*. pp. 38–43.

Benferhat, S., C. Cayrol, D. Dubois, J. Lang, and H. Prade: 1993, 'Inconsistency management and prioritized syntax-based entailment'. In: *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI'93)*. pp. 640–645.

Boutilier, C.: 1994, 'Towards a logic for qualitative decision theory'. In: *Proceedings of the Fourth International Conference on Knowledge Representation and Reasoning (KR'94)*. pp. 75–86.

Castelfranchi, C., F. Dignum, C. Jonker, and J. Treur: 2000, 'Deliberate normative agents: principles and architecture'. In: *Intelligent Agents VI. Proc. of the Sixth International Workshop on Agent Theories, Architectures and Languages, ATAL'99*.

Cholvy, L. and F. Cuppens: 1995, 'Solving normative conflicts by merging roles'. In: *Proceedings of the Fifth International Conference on Artificial Intelligence and Law (ICAIL'95)*. Washington.

Cohen, P. and H. Levesque: 1990, 'Intention is choice with commitment'. *Artificial Intelligence* **42 (2-3)**, 213–261.

Dean, T. and M. Wellman: 1991, *Planning and Control*. San Mateo: Morgan Kaufmann.

Doyle and Thomason: 1999, 'Background to qualitative decision theory'. *AI magazine* **20:2**, 55–68.

Doyle, J.: 1980, 'A model for deliberation, action and introspection'. Technical Report AI-TR-581, MIT AI Laboratory.

Doyle, J.: 1991, 'Rationality and its rules in reasoning (extended abstract)'. In: *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI'91)*. pp. 1093–1100.

Doyle, J., Y. Shoham, and M. Wellman: 1991, 'The logic of relative desires'. In: *Sixth International Symposium on Methodologies for Intelligent Systems*. Charlotte, North Carolina.

Doyle, J. and M. Wellman: 1991, 'Preferential semantics for goals'. In: *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI'91)*. pp. 698–703.

Dubois, D. and H. Prade: 1995, 'Possibility theory as a basis for qualitative decision theory'. In: *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95)*. pp. 1924–1930.

Gelfond, M. and V. Lisfchitz: 1993, 'Representing action and change by logic programs'. In: *Journal of Logic Programming*, Vol. 17. pp. 301–322.

Goldszmidt, M., P. Morris, and J. Pearl: 1993, 'A maximum entropy approach to nonmonotonic reasoning'. *IEEE transactions on pattern analysis and machine intelligence* **15:3**, 220–232.

Haddawy, P. and S. Hanks: 1992, 'Representations for Decision-Theoretic Planning: Utility Functions for dead-line goals'. In: *Proceedings of Third International Conference on Knowledge Representation and Reasoning (KR'92)*. Cambridge, MA.

Hansson, B.: 1971, 'An analysis of some deontic logics'. In: R. Hilpinen (ed.): *Deontic Logic: Introductory and Systematic Readings*. Dordrecht, Holland: D. Reidel Publishing Company, pp. 121–147.

Horty, J.: 1994, 'Moral Dilemmas and Nonmonotonic Logic'. *Journal of Philosophical Logic* **23**, 35–65.

Jennings, N. and J. Campos: 1997, 'Towards a social level characterisation of socially responsible agents'. *IEEE Proceedings on Software Engeneering* **144 (1)**, 11–25.

Keeney, R. and H. Raiffa: 1976, *Decisions with Multiple Objectives: Preferences and Value Trade-offs*. New York: John Wiley and Sons.

Lang, J.: 1996, 'Conditional desires and utilities - an alternative approach to qualitative decision theory'. In: *Proceedings of the Twelth European Conference on Artificial Intelligence (ECAI'96)*. pp. 318–322.

Lehmann, D.: 1996, 'Generalized qualitative probability: Savage revisited'. In: *Proceedings of Twelth Conference on Uncertainty in Artificial Intelligence (UAI'96)*. pp. 381–388.

Lehmann, D.: 1998, 'Non-standard numbers for qualitative decision making'. In: *Proceedings of Seventh Conference on Theoretical Aspects of Rationality and Knowledge (TARK'98)*. pp. 161–174.

Lehmann, D. and M. Magidor: 1992, 'What does a conditional knowledge base entail?'. *Artificial Intelligence* **55**, 1–60.

Makinson, D.: 1993, 'Five Faces of Minimality'. *Studia Logica* **52**, 339–379.

Makinson, D.: 1994, 'General Patterns in Nonmonotonic Reasoning'. In: Gabbay, Hogger, and Robinson (eds.): *Handbook of Logic in Artificial Intelligence and Logic Programming*, Vol. 3. Oxford: Oxford University Press, pp. 35–110.

Neapolitan, R.: 1990, *Probabilistic Reasoning in Expert Systems.* John Wiley and Sons.

Pearl, J.: 1993, 'From conditional ought to qualitative decision theory'. In: *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence (UAI'93).* pp. 12–20.

Prakken, H. and M. Sergot: 1996, 'Contrary-to-duty obligations'. *Studia Logica* **57**, 91–115.

Rao, A. and M. Georgeff: 1991, 'Modeling rational agents within a BDI architecture'. In: *Proceedings of Second International Conference on Knowledge Representation and Reasoning (KR'91).* pp. 473–484.

Tan, S.-W. and J. Pearl: 1994a, 'Qualitative decision theory'. In: *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI'93).*

Tan, S.-W. and J. Pearl: 1994b, 'Specification and Evaluation of Preferences Under Uncertainty'. In: *Proceedings of the Fourth International Conference on Knowledge Representation and Reasoning (KR'94).* pp. 530–539.

Tan, Y. and L. van der Torre: 1996, 'How to combine ordering and minimizing in a deontic logic based on preference'. In: *Deontic Logic, Agency and Normative Systems. Proceedings of Third Workshop on Deontic Logic in Computer Science (Δeon'96).* pp. 216–232.

Thomason, R.: 2000, 'Desires and defaults: a framework for planning with inferred goals'. In: *Proceedings of Seventh International Conference on Knowledge Representation and Reasoning (KR'2000).* pp. 702–713.

Thomason, R. and R. Horty: 1996, 'Nondeterministic action and Dominance: foundations for planning and qualitative decision'. In: *Proceedings of Theoretical Aspects of Reasoning about Knowledge (TARK'96).* pp. 229–250.

van der Torre, L.: 1994, 'Violated Obligations in a Defeasible Deontic Logic'. In: *Proceedings of the Eleventh European Conference on Artificial Intelligence (ECAI'94).* pp. 371–375.

van der Torre, L.: 1998, 'Labeled logics of goals'. In: *Proceedings of the Thirteenth European Conference on Artificial Intelligence (ECAI'98).* pp. 368–369.

van der Torre, L. and Y. Tan: 1999a, 'Contrary-To-Duty Reasoning with Preference-based Dyadic Obligations'. *Annals of Mathematics and Artificial Intelligence* **27**, 49–78.

van der Torre, L. and Y. Tan: 1999b, 'Diagnosis and Decision Making in Normative Reasoning'. *Artificial Intelligence and Law* **7**, 51–67.

van der Torre, L. and E. Weydert: 2000, 'Parameters for utilitarian desires in a qualitative decision theory'. *Applied Intelligence.*

von Neumann, J. and O. Morgenstern: 1944, *theories of Games and Economic Behavior.* Princeton University Press.

von Wright, G.: 1983, *Norms Truth and Logic. Practical Reason.* Oxford: Blackwell.

Wainer, J.: 1994, 'Yet another semantics for goals and goal priorities'. In: *Proceedings of the Eleventh European Conference on Artificial Intelligence (ECAI'94).* pp. 269–273.

Weydert, E.: 1998, 'System JZ: How to build a canonical ranking model of a default knowledge base'. In: *Proceedings of the Seventh International Conference on Knowledge Representation and Reasoning (KR'98).* pp. 190–201.