Hidden Uncertainty in the Logical Representation of Desires

Jérôme Lang
IRIT / UPS
31062 Toulouse Cedex
France
lang@irit.fr

Leendert van der Torre
CWI
Amsterdam
The Netherlands
torre@cwi.nl

Emil Weydert
University of Applied Sciences
Luxembourg
Luxembourg
emil.weydert@ist.lu

Abstract

In this paper we introduce and study a logic of desires. The semantics of our logic is defined by means of two ordering relations representing preference and normality as in Boutilier's logic QDT. However, the desires are interpreted in a different way: "in context A, I desire B" is interpreted as "the best among the most normal $A \wedge B$ worlds are preferred to the most normal $A \wedge \neg B$ worlds". We study the formal properties of these desires, illustrate their expressive power on several classes of examples and position them with respect to previous work in qualitative decision theory.

1 Introduction

Autonomous agents reason frequently about preferences, desires and goals. For example, Cohen and Levesque [1990] explore principles governing the rational balance among an agent's beliefs, goals, actions and intentions, Rao and Georgeff [1991] show how different types of 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; 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 and performing a sequence of actions to achieve these goals.

In logical formalizations of preferences, desires and goals serve as a computationally useful partial specification or heuristic approximation of the relative preference over the possible results of a plan [Doyle, 1991]. In this paper, we focus on desires although our observations may be relevant for goals as well. There are three different interpretations of sentences like "I desire B" or "in context A, I desire B".

First, desires may be formalized using only utility or preference. However, this leads to the problem that even if a utility function or a preference relation over worlds is fixed, there is no unique way to extend it to formulas or, equivalently, sets of worlds. This distinction between utilities and probabilities is the main reason why formalizing desires is more problematic than formalizing beliefs. Moreover, it does not seem to correspond to the meaning of desire in natural language. For instance, suppose that I exceptionally get sunburned when the weather is hot and sunny. Then, expressing "I desire a hot and sunny weather" intuitively means that the most normal or typical worlds satisfying *hot* $\wedge$ *sunny* are preferred to the most normal worlds verifying $\neg$ *(hot* $\wedge$ *sunny)*, but it does not mean that I like exceptional effects of *hot* $\wedge$ *sunny* such as *sunburn*.

Secondly, desires may express a combination of utility and probability in a classical decision-theoretic context, assuming a probability distribution over worlds. "In context A, I desire B" can then be interpreted as a raise of expected utility: *my expected utility given* $A \wedge B$ *is higher than my expected utility given* $A \wedge \neg B$. A further development in terms of gain of expected utility is proposed in [Brafman and Friedman, 2001]. However, it is well-known that autonomous agents do not always have appropriate probabilistic information on the possible worlds. The probabilistic approach asks for costly, specific information and in practice often leads to an arbitrary choice of probability values.

Thirdly, desires may implicitly refer to the relative *plausibility* or *normality* of worlds. When an agent states "I desire B", *i.e.*, "I desire B to be satisfied", the agent often focuses on typical, normal worlds. The standard choice made in *qualitative decision theory*, see [Doyle and Thomason, 1999], is that uncertainty is described by a *(total) pre-order* expressing, e.g., plausibility, normality, or typicality. The main justification is that ordinal uncertainty is less committing, cognitively speaking, than numerical uncertainty.

The research question of this paper is how we can formalize desires to provide a realistic interaction between preference and normality, and such that realistic examples like examples 2.4, 3.3, and 3.5 can be formalized. E.g., in the latter an airline company desires for an overbooked plane that an individual passenger shows up, but at the same time it desires that not all passengers show up. We start with Boutilier's notion of so-called ideal goals [1994], but then we define conditional desires that refer to both preference and normality, and that turn out to be more satisfactory.

The layout of this paper is as follows. Section 2 recalls the basics of Boutilier's semantics and discusses the choices upon which it is based; we then explain how to extend it to represent desires in a more appropriate way. The details of this modified interpretation of desires are given in Section 3. We study several examples and in particular, we illustrate how our semantics interprets sets of desires that are usually considered contradictory.

2 Boutilier's qualitative decision theory logic

A first class of solutions to combine preference and normality makes use of a trade-off between qualitative utilities and probabilities as in [Pearl, 1993], but this is not well-suited to situations with non-extreme utilities. A second approach is given in [Boutilier, 1994], which makes use of two ordering relations, one for preference and one for normality, representing the ordinal counterparts of utility and probability. However, here we see – informally speaking – a lack of interaction between preference and normality.

2.1 Definitions

Boutilier [1994] interprets a desire in terms of ideal worlds: *if A then ideally B*, denoted by **I**$(B|A)$, is a conditional desire expressing on the semantic level that among all A-worlds, the most preferred ones also satisfy B. Quoting Boutilier, "from a practical point of view, **I**$(B|A)$ means that if the agent (only) knows A, and the truth value of A is fixed (beyond its control), then the agent ought to ensure B (...) The statement can be *roughly* interpreted as if A, do B."

This definition enables the formalization of conditional desires in the conditional logic QDT, which is based on his logic CO. *Roughly*, the semantics of QDT is the following: a QDT-model is $M = \langle W, \geq_P, \geq_N, val\rangle$ where W is a set of possible worlds, val a valuation function and $\geq_P$ and $\geq_N$ are two total pre-orders over W, i.e., transitive connected relations, where a relation is connected iff $w \geq v$ or $v \geq w$ for all $w, v \in W$. They are also called total weak orders. $\geq_P$ is a *preference* order and $\geq_N$ a *normality* order: $w \geq_P w'$ (resp. $w \geq_N w'$) means that w is at least as preferred (resp. as normal) as w'. The conditional connectives **I**(-|-) and **N**(-|-) have the following truth conditions. The two modalities do not mix preference and normality; in particular, conditional desires **I**(-|-) do not use the normality relation at all.

- $M \models$ **I**$(B|A)$ if and only if either $\forall w \in W,\ M, w \models \neg A$, or there exists a $w \in W$ such that $M, w \models A$ and for all $w' \in W$ with $w' \geq_P w$, $M, w' \models A \rightarrow B$. **I**$(B|A)$ expresses the conditional preference "if A then ideally B".
- $M \models$ **N**$(B|A)$: as **I**$(B|A)$, except that $\geq_P$ is replaced by $\geq_N$. **N**$(B|A)$ expresses the conditional default "if A then normally B".

In the rest of the paper we write **I**(B) for **I**$(B|\top)$ and **N**(B) for **N**$(B|\top)$, where $\top$denotes a tautology.

Boutilier also introduces *ideal goals* that combine preference and normality. However, they use only the upper cluster of the normality relation, nothing else. On the semantic level, if we assume that there are no infinite ascending chains, then the truth conditions for ideal goals are as follows. **IG**$(B|A)$ is true if the best of the most normal A worlds satisfy B.

- $M \models$ **IG**$(B|A)$ if and only if

$$Max(\geq_P, Max(\geq_N, Mod(A))) \subseteq Mod(B)$$

Therefore, Boutilier's interpretation of conditional desires relies on two strong assumptions, which we discuss in more detail below. The first assumption is the optimistic interpretation of desires due to the ideality assumption. The second assumption is the weak interaction between preference and normality due to the focus on on the most normal worlds in a context.

2.2 Assumptions

Ideality

The ideality semantics consists in comparing sets of worlds by looking only at the most preferred worlds of these sets. It corresponds to an optimistic point of view in the sense that less preferred worlds are ignored in this process. What would be possible choices for comparing sets of worlds? If we do not want to bring in probabilistic information or assumptions such as equiprobability, then we are left with the following four basic alternatives, and variations or combinations thereof. Let A and B be two formulas. For the sake of simplicity we only consider the non-degenerate case where $Mod(A \wedge B)$ and $Mod(A \wedge \neg B)$ are both non-empty.

MM (ideality semantics): $\mathbf{I}^{MM}(B|A)$ if and only if the best $A \wedge B$-worlds are preferred to the best $A \wedge \neg B$-worlds.

mm: $\mathbf{I}^{mm}(B|A)$ if and only if the worst $A \wedge B$-worlds are preferred to the worst $A \wedge \neg B$-worlds.

mM: $\mathbf{I}^{mM}(B|A)$ if and only if the worst $A \wedge B$-worlds are preferred to the best $A \wedge \neg B$-worlds.

Mm: $\mathbf{I}^{Mm}(B|A)$ if and only if the best $A \wedge B$-worlds are preferred to the worst $A \wedge \neg B$-worlds.

We have that $\mathbf{I}^{mM}(B|A)$ implies both $\mathbf{I}^{mm}(B|A)$ and $\mathbf{I}^{MM}(B|A)$, and all of them imply $\mathbf{I}^{Mm}(B|A)$. Note that only MM and mm are consistent with the semantics of conditional logics; for mm, just take the reverse preference relation. The Mm variant is extremely weak and therefore tells us next to nothing. It may be useful only when it is extended with a non-monotonic reasoning mechanism. The mM variant is extremely strong and hard to satisfy; it certainly does not reflect the usual intuitive understanding of desires. It may be useful only when paired with normality, e.g., by focusing on the most normal worlds.

The mm variant has an underlying *pessimistic semantics*. It makes sense if "I desire B in context A" is interpreted as "in context A, given that I expect the worst outcomes to occur, I am happy to see B true". However, this semantics also does not fit well the intuitions behind the specification of desires. Consider the following example of two desires.

Example 2.1 (Game) *Assume that an agent plays a game where two coins are tossed; he wins if both coins are heads (h) and loses otherwise. His preference ordering can be:* $(h_1, h_2) >_P (h_1, \neg h_2) \sim_P (\neg h_1, h_2) \sim_P (\neg h_1, \neg h_2)$; *the normality ordering $\geq_N$ is the one where all worlds are equally normal. But, surprisingly, neither $\mathbf{I}^{mm}(h_1)$ nor $\mathbf{I}^{mm}(h_2)$ are satisfied.*

The example supports that Boutilier's ideality semantics best suits the commonsense intuitions concerning desired outcomes of actions, and we therefore do not want to give it up.

At a first glance, it may seem paradoxical to favor an optimistic interpretation of desires while most papers in qualitative decision theory argue in favor of a pessimistic (maxmin) criterion for action selection [Brafman and Tennenholtz, 1997; Dubois *et al.*, 1998]. The paradox is however only in appearence. When specifying **I**$(B|\top)$, for instance, an agent expresses that she has a preference for B, which is not at all the same as saying that she intends to take a decision making B true. This is illustrated by the following classical example.

Example 2.2 *[Boutilier, 1994] Let DS = \{ **I**$(\neg u)$, **I**$(u|r)$ \} where u and r stand for* umbrella *and* raining. **I**$(\neg u)$ *expresses that the agent prefers not carrying an umbrella. This* does not imply *that if the agent has the choice between the actions* take-umbrella *and* leave-umbrella *(whose obvious outcomes are u and $\neg u$), she will choose* leave-umbrella*.*

The example illustrates that the interpretation of desires and the action selection criterion are independent issues and we can consistently interpret desires in Boutilier's ideality semantics while using a pessimistic criterion for action choice. This issue is discussed further in Section 4.

Weak interaction between normality and preference

This assumption is more problematic than the previous one. Boutilier's interpretation of ideal goals makes a very rough use of the normality relation, since it consists in focusing first, and once for all, on the most normal worlds, independently of the desires expressed, and then in interpreting desires by the ideality semantics.

This has unfortunate consequences. Ideal goals are different from desires because we have the counterintuitive property that if normally p then p is an ideal goal. Moreover, Boutilier's semantics makes intuitively coherent sets of desires inconsistent, such as the *dog and fence* example of defeasible deontic logic [van der Torre and Tan, 1997].

Example 2.3 (Dog and fence)
DS = \{**I**$(\neg f)$, **I**$(f|d)$, **I**(d)\}

1. *John does not want a fence around his cottage;*
2. *If John owns a dog, then he wants a fence;*
3. *John wants to own a dog.*

Example 2.3 is inconsistent, and would still be inconsistent if we would replace the **I** modality by **IG**.

Here is another example which is inconsistent in Boutilier's semantics. Suppose that I am going to my travel agent just one day before Christmas vacation, when normally all flights are fully booked.

Example 2.4 (Airplane ticket)
DS = \{**I**(r), **I**(a), $\neg$ **I**$(r \wedge a)$, **N**$(\neg r \wedge \neg a)$ \}

1. *I desire to have an airplane ticket to Rome;*
2. *I desire to have an airplane ticket to Amsterdam;*
3. *I do not desire to have both an airplane ticket to Rome and an airplane ticket to Amsterdam;*
4. *Normally, I will neither get an airplane ticket to Rome nor to Amsterdam.*

Example 2.4 is inconsistent, and would still be so if we replaced the **I** modality by **IG**. However, this set of desires has an intuitive interpretation: when I think of having a ticket to Rome, I think of the most normal world where I have a ticket to Rome, in which I do not have a ticket to Amsterdam, and vice versa, and I prefer this world to the most normal world where I do not have a ticket to Rome, in which I do not have a ticket to Amsterdam either. See another interpretation of this set of desires in example 3.6.

3 Hidden uncertainty

Example 2.4 illustrates that, even if the intuitive expression of the desires does not mention normality or uncertainty issues, it implicitly refers to these. This is why we talk of *hidden uncertainty* in the specification of desires.

3.1 Definitions

We now introduce a notion of desires with hidden uncertainty that better fits the intuitive meaning of desires. To simplify the definition, we assume – in contrast to Boutilier – the existence of maxima. We may guarantee this by stipulating, e.g., that there are no infinite ascending chains. **D**$(B|A)$ means that there are maximally normal $A \wedge B$-worlds which are strictly preferred to all the most normal $A \wedge \neg B$-worlds.

Definition 3.1 *Truth conditions for desires are as follows. $M \models$**D**$(B|A)$ if and only if $\forall w' \in Max(\geq_N, Mod(A \wedge \neg B))$ $\exists w \in Max(\geq_N, Mod(A \wedge B))$ such that $w >_P w'$.*

In other words, "in context A, I desire B" is interpreted as "the best among the most normal $A \wedge B$ worlds are preferred to the most normal $A \wedge \neg B$ worlds". As usual, we write **D**(B) instead of **D**$(B|\top)$.

There exists also a probabilistic interpretation of desires which may help to clarify their meaning. The basic idea is to use the ϵ-semantics for normality, a super-ϵ-semantics – e.g. replacing ϵ by ϵ^ϵ – for utility, and the resulting concept of expected utility for interpreting desires.

In a sense, our definition of desires is still reminiscent of the ideality semantics, because for a desire **D**$(B|A)$ we only consider the most normal $A \wedge B$ and the most normal $A \wedge \neg B$ worlds. However, the fact that we do not only consider the most normal A worlds makes a crucial difference. This is illustrated by the properties and examples below.

3.2 Properties

The definition guarantees that the conditional desire always holds if the implication is strict, i.e., if $A \wedge \neg B$ is inconsistent. When all worlds are equally normal, desires represent pure preference in the sense that both modalities **D**() and **I**() coincide. However, in general none of the following implications **D**$(A|B) \Rightarrow$ **I**$(A|B)$, **I**$(A|B) \Rightarrow$ **D**$(A|B)$, **D**$(A|B) \Rightarrow$ **IG**$(A|B)$ and **IG**$(A|B) \Rightarrow$ **D**$(A|B)$ holds. Furthermore it avoids – as opposed to some of its variants – the validation of conditional desires of the form **D**$(\neg A|A)$ for consistent A.

The following example illustrates that we have no longer right weakening (strictly speaking, left weakening, given our notation) for the desires, i.e. **D**(A) does not imply **D**$(A \vee B)$.

Example 3.1 *Let M be the following model in which $\neg p$ is always more plausible than p, whereas p is always preferred to $\neg p$.*

$\geq_N$: $\{(\neg p, q), (\neg p, \neg q)\} >_N \{(p, q), (p, \neg q)\}$;

$\geq_P$: $(p, \neg q) >_P (p, q) >_P (\neg p, \neg q) >_P (\neg p, q)$

Then M satisfies **D**(p) *but not* **D**$(p \vee q)$.

This absence of right weakening, a controversial property when reasoning with obligations in deontic logics, is also very natural for desires: take p = "the woman of my dreams falls in love with me" and q = "I receive my electricity bill" (needless to say where the most normal worlds are).

The violation of right weakening also explains the distinction between the two conflicts in the following example. The example also illustrates that the conjunction (or AND) rule is not valid for **D**.

Example 3.2 *$\{$**D**(p), **D**$(\neg p)\}$ is inconsistent, but $\{$**D**(p), **D**$(\neg p \wedge q)\}$ is consistent when q is exceptional, as is witnessed by the following model:*

$\geq_N$: $\{(p, \neg q), (\neg p, \neg q)\} >_N \{(p, q), (\neg p, q)\}$

$\geq_P$: $(\neg p, q) >_P (p, \neg q) >_P \{(\neg p, \neg q), (p, q)\}$

Interestingly, many rules valid for **I**$(B|A)$ do no longer hold for **D**$(B|A)$, as illustrated by the two examples above. Or, more precisely, they hold only conditionally. The reason is a kind of context-dependence or "higher-order nonmonotonicity". For instance, if $Mod(A \wedge B)$ increases or decreases, it does not follow that $\max(\geq_N, Mod(A \wedge B))$ increases or decreases as well. One way to design proof rules is to ensure with additional clauses that this implication holds, which has been done for decision-theoretic defaults [Brafman and Friedman, 2001]. However, our logic of desires also validates inference rules which do not directly depend on such conditions.

The first formula of the following proposition illustrates that the conjunction rule holds under normality conditions **N**$(A|B \wedge C)$ and **N**$(B|A \wedge C)$, and the latter two formulas hold without any conditions.

Proposition 3.1 *The following formulas are theorems of our logic:*

1. **D**$(A|C)$ $\wedge$ **D**$(B|C)$ $\wedge$ **N**$(A|B \wedge C)$ $\wedge$ **N**$(B|A \wedge C)$ $\Rightarrow$ **D**$(A \wedge B|C)$;
2. **D**$(A|B \wedge C)$ $\wedge$ **D**$(B|C)$ $\Rightarrow$ **D**$(A \wedge B|C)$;
3. **D**$(A|A \vee B)$ $\wedge$ **D**$(\neg A|\neg A \vee B)$ $\Rightarrow$ **D**$(\neg B|\top)$;

Properties expressed in the dyadic modal logic are often hard to read. We therefore propose an alternative strategy for generating proof rules. The idea is to express desires with a suitable order modality $<_{nd}$ where $A <_{nd} B$ means that, taking into account normality and adopting an optimistic perspective, A is less desirable than B, or inconsistent. Formally speaking, we get the following truth condition for $<_{nd}$ (assuming finitely many worlds).

Definition 3.2 *Truth condition of preference is as follows. $M \models A <_{nd} B$ if and only if $\forall w' \in Max(\geq_N, Mod(A))$ $\exists w \in Max(\geq_N, Mod(B))$ such that $w >_P w'$.*

We have the following relationship between the conditional desires and preferences.

Proposition 3.2 *The following translation rules are valid between $<_{nd}$ and* **D**$(.|.)$.

- $M \models A <_{nd} B$ *iff* $M \models$**D**$(\neg A|A \vee B)$
- $M \models$**D**$(B|A)$ *iff* $M \models (A \wedge \neg B) <_{nd} (A \wedge B)$

Similarly, we can introduce $<_n$ for expressing propositional normality comparisons. Taken together, these modalities allow us to formulate a number of features which are conceptually easier to grasp and to handle than the corresponding conditional notions. The following proposition lists several properties using this alternative representation.

Proposition 3.3 *The following formulas are theorems of the logic. They are called left weakening (L1), right weakening (L2), left strengthening (L3), left impossibility (L4), right possibility (L5), asymmetry (L6), left disjunction (L7), right disjunction (L8), transitivity (L9), left cautious transitivity (L10), right cautious transitivity (L11).*

L1 $(A <_{nd} B) \wedge (C <_n A) \rightarrow ((A \vee C) <_{nd} B)$

L2 $(A <_{nd} B) \wedge \neg(B <_n C) \rightarrow (A <_{nd} (B \vee C))$

L3 $(A <_{nd} B) \wedge \neg((A \wedge C) <_n A) \rightarrow ((A \wedge C) <_{nd} B)$

L4 $(\perp <_{nd} A)$

L5 $(A <_{nd} \perp) \rightarrow (A <_n \perp)$

L6 $(A <_{nd} B) \wedge \neg(A <_n \perp) \wedge \neg(B <_n \perp) \rightarrow \neg(B <_{nd} A)$

L7 $(A <_{nd} C) \wedge (B <_{nd} C) \rightarrow ((A \vee B) <_{nd} C)$

L8 $(A <_{nd} B) \wedge (A <_{nd} C) \rightarrow (A <_{nd} (B \vee C))$

L9 $(A <_{nd} B) \wedge (B <_{nd} C) \rightarrow (A <_{nd} C)$

L10 $(A <_{nd} B) \wedge ((A \vee B) <_{nd} C) \rightarrow (A <_{nd} C)$

L11 $(A <_{nd} (B \vee C)) \wedge (B <_{nd} C) \rightarrow (A <_{nd} C)$

Due to lack of space we must omit proofs and discussion.

3.3 Examples

Let us consider Example 2.3 and 2.4 taking hidden uncertainty into account.

Example 3.3 (Dog and fence, continued)
*$\{$**D**(d), **D**$(\neg f)$, **D**$(f|d)\}$ is consistent. Here are four classes of models, which illustrate that for each normality ordering there are various preference orderings.*

class 1: *dogs are exceptional (and nothing else)*

$\geq_N$: $\{(\neg d, f), (\neg d, \neg f)\} >_N \{(d, f), (d, \neg f)\}$
$\geq_P$: *any complete preordering satisfying constraints*
$(d, f) >_P (d, \neg f)$
$(d, f) >_P (\neg d, \neg f) >_P (\neg d, f)$

class 2: *not having a fence is exceptional (and nothing else)*

$\geq_N$: $\{(\neg d, f), (d, f)\} >_N \{(\neg d, \neg f), (d, \neg f)\}$
$\geq_P$: *any complete preordering satisfying constraints*
$(\neg d, \neg f) >_P (\neg d, f)$, $(d, f) >_P (d, \neg f)$
$(d, f) >_P (\neg d, f)$

class 3: *dogs and not having a fence are both exceptional*

$\geq_N$: $(\neg d, f) >_N \{(d, f), (\neg d, \neg f)\} >_N (d, \neg f)$
$\geq_P$: *any complete preordering satisfying constraints*
$(\neg d, \neg f) >_P (\neg d, f)$, $(d, f) >_P (\neg d, f)$
$(d, f) >_P (d, \neg f)$

class 4: *neither dogs nor not having a fence is exceptional*

$\geq_N$: $(d, \neg f) >_N (\neg d, f) > (d, f) >_N (\neg d, \neg f)$
$\geq_P$: *any complete preordering satisfying constraints*
$(d, \neg f) >_P (\neg d, f)$, $(d, f) >_P (\neg d, f)$

Example 3.4 (Airplane ticket, continued)
*DS = $\{$**D**(r), **D**(a), $\neg$ **D**$(r \wedge a)$, **N**$(\neg r \wedge \neg a)$ $\}$. Here is a set of models satisfying DS:*

$\geq_N$: $(\neg r, \neg a) >_N \{(r, \neg a), (\neg r, a)\} >_N (r, a)$;

$\geq_P$: *any order where $(r, \neg a), (\neg r, a)$ are strictly more preferred than the other worlds.*

The following example is a reformulation of the lottery paradox in terms of desires.

Example 3.5 (overbooking)
The agent is an airline company which has sold 301 tickets for a flight on an airplane of 300 seats. For each seat occupied the company gains 100 utility units, but if all 301 persons show up, then the company looses 1000 utility units. The agent may consistently express $\{$**D***(show-up(1)), ...,* **D***(show-up(301)),* **D***(¬ (show-up(1)* $\wedge$ *...* $\wedge$ *show-up(301))*$\}$*, because, individually, passenger #i showing up makes the expected utility of the company increase (slightly), due to the fact that it is* very unlikely *that all passengers show up.*

3.4 Normality and update

There are several perspectives on normality. One consists in viewing normality as distance to the current situation. When an agent figures out a "normal A-world" he often figures out the closest A-world to the actual world. This is in accordance with the principle used for evaluating counterfactuals. Update could come very intuitively into this framework.

The normality ordering is then defined by the proximity to the current world, which is defined by a faithful proximity relation in the sense of [Katsuno and Mendelzon, 1991], i.e., a collection of weak orders $\{\leq_w, w \in W\}$, where faithfulness is the condition: $w <_w w'$ for all $w' \neq w$; $w_1 \leq_w w_2$ means w_1 is closer to w than w_2. We simply have $w_1 \geq_N w_2$ iff $w_1 \leq_w w_2$, and the set of most normal worlds (*i.e.*, closest worlds to w) satisfying A is the *update* of w, denoted by $w \diamond A$. The simplest and most frequent choices are: $w_1 \leq_w w_2$ iff *Diff*$(w, w_1) \subseteq$ *Diff*(w, w_2) and $w_1 \leq_w w_2$ iff $|$*Diff*$(w, w_1)| \leq |$*Diff*$(w, w_2)|$, where *Diff*(w, w') is the set of variables assigned a different value by w and w'.

Let us consider example 2.4 but with a slightly different interpretation. We do not have to suppose here that flights are normally fully booked. We just assume that *in the current situation, the agent does not have any airplane ticket*.

Example 3.6 (Airplane ticket, continued)
DS = $\{$**I***(r),* **I***(a),* ¬ **I***(r* $\wedge$ *a),* ¬ *r* $\wedge$ ¬ *a* $\}$

1. *I desire to have an airplane ticket to Rome;*
2. *I desire to have an airplane ticket to Amsterdam;*
3. *I do not desire to have both an airplane ticket to Rome and an airplane ticket to Amsterdam;*
4. *In the current situation I do not have any airplane ticket.*

If the proximity relation is such that (r, a) is closer to $(\neg r, a)$ and $(r, \neg a)$ than to $(\neg r, \neg a)$, and equally close to $(\neg r, a)$ and $(r, \neg a)$, then we get the following normality ordering: $(\neg r, \neg a) >_N (\neg r, a) \sim_N (r, \neg a) >_N (r, a)$, because the initial situation is $(\neg r, \neg a)$. Normality reflects proximity to the initial situation.

Of course, this assumes that a preliminary step has been done so as to translate the proximity to the current world into a normality ordering; this issue has been considered several times in the literature. See for instance [Grahne, 1991] for the intertranslation between update and conditional logics, and [Herzig, 1998] for a review of logics for belief update.

4 Related research

4.1 Two tasks in qualitative decision theory

Qualitative decision theory aims at developing mainly non-numerical – and therefore non-probabilistic – normative frameworks for decision making under uncertainty, e.g. looking for minimal sets of minimal behavioral properties or axioms of a rational agent that correspond to a given action selection criterion. Most approaches, e.g., [Brafman and Tennenholtz, 1996; Lehmann, 2001; Dubois *et al.*, 2002], use ordinal structures for preference and uncertainty.

In a decision-theoretic context we can distinguish at least the task of interpreting desires, which aims at reasoning about the mental state of the agent (what he likes and what he believes) and the task of selecting an action that *uses* the possible mental states induced by the upstream task, like qualitative analogs of maximum expected utility. Both tasks are complementary. The logic developed in this paper does not investigate criteria for action selection and therefore it *is not really a new approach to qualitative decision theory*. Our logic aims at *interpreting desires* as they can be expressed by agents, for instance in an interactive elicitation process. Our logic can infer some information about the normality ordering or the preference ordering of an agent's mental state but cannot predict which action he will perform.

There are several ways in which our logic can be extended with action selection. In example 2.2, when a model contains both r and $\neg r$ normal worlds, **I***(¬ u)* merely expresses that preferred worlds satisfy $\neg u$ while the selection of action *take-umbrella* may well reflect that *the worst among the most normal effects of this action* are preferred to the worst among the most normal effects of *leave-umbrella*. The latter criterion is used by [Brafman and Tennenholtz, 1997] who model agents as pessimistic decision makers. Noticeably, it is similar to our interpretation of desires, except that worst states have to be focused on instead of best states. It would be worth extending our logical framework with a modality **PA***(.|.)*: the preference of taking action α over action β would be expressed by **PA***(*$do(\alpha)|do(\alpha) \vee do(\beta)$*)*, expressing that the worst normal effects of action α are preferred to the worst normal effects of action β. This modality would have the same properties as the **D***(.|.)* modality, except that the preference order has to be reversed: if $M = \langle W, \geq_P, \geq_N, val\rangle$ and $M' = \langle W, reverse(\geq_P), \geq_N, val\rangle$ then $M \models$ **PA***(B|A)* iff $M' \models$ **D***(B|A)*. This extension is left for further research.

4.2 Formalisms of desires

There are several other formalisms that represent desires and pure preference, especially *ceteris paribus constraints* [Doyle and Wellman, 1991; Boutilier *et al.*, 1999]. However, this framework deals with pure preference only and not with uncertainty and normality, except maybe a preliminary attempt in [Tan and Pearl, 1994]. The combination of ceteris paribus and normality is an issue for further research.

Thomason's framework [Thomason, 2000] builds on Reiter's default logic and deals with both normality and preference defaults, but with a procedural strategy which departs from our completely semantical interpretation: goals are derived by first closing the facts under beliefs defaults, and thereafter under desire defaults. The same mechanism is used in BOID architectures [Broersen *et al.*, 2002]. The goals are thus restricted to the most normal state.

Finally, decision-theoretic defaults [Brafman and Friedman, 2001; Poole, 1992] also have a semantics based on normality and preference. For example, Brafman and Friedman give a detailed analysis of the use of expected utility, they explain a drawback of straightforward expected utility, and they introduce defaults based on gain of utility $G(A)$. Moreover, they define weak and strong notions of defaults, where the first does not satisfy right weakening, because $G(A)$ is not informative about the behavior on subsets of A. This seems related to what we here called higher-order nonmonotonicity.

5 Summary

The research question of this paper is how desires can be formalized with a realistic interaction between preference and normality. We start with Boutilier's logic for QDT, in which preference and normality have been combined in a notion of ideal goal. However, there is only weak interaction, such that the dog and fence example as well as the airplane ticket example cannot be represented in a consistent way. We therefore interpret desires in a different way: "in context A, I desire B" is interpreted as "the best among the most normal $A \wedge B$ worlds are preferred to the most normal $A \wedge \neg B$ worlds".

We study various formal properties of these desires. We show that our **D**$(A|B)$ does not imply Boutilier's **I**$(A|B)$ or **IG**$(A|B)$, nor vice versa. We show that they do not satisfy weakening nor the the conjunction rule. We show that these properties and many others hold under conditions of normality, and we show that some properties like transitivity and cumulativity hold unconditionally. We illustrate the expressive power on several classes of examples including the dog and fence and airline ticket examples, and we illustrate how update can be introduced in the framework. Finally we position the desires with respect to previous work in qualitative decision theory, where we mention as subjects of further research an extension with an action selection criterion, and the introduction of normality in ceteris paribus preferences.

References

[Bacchus and Grove, 1996] F. Bacchus and A.J. Grove. Utility independence in a qualitative decision theory. In *Proceedings of KR'96*, pages 542–552, 1996.

[Boutilier *et al.*, 1999] C. Boutilier, R. Brafman, H. Hoos, and D. Poole. Reasoning with conditional *ceteris paribus* statements. In *Proceedings of UAI'99*, pages 71–80, 1999.

[Boutilier, 1994] C. Boutilier. Towards a logic for qualitative decision theory. In *Proceedings of KR'94*, pages 75–86, 1994.

[Brafman and Friedman, 2001] R.I. Brafman and N. Friedman. On decision-theoretic foundations for defaults. *Artificial Intelligence*, 133:1–33, 2001.

[Brafman and Tennenholtz, 1996] R. Brafman and M. Tennenholtz. On the foundations of qualitative decision theory. In *Proceedings of AAAI'96*, pages 1291–1296, 1996.

[Brafman and Tennenholtz, 1997] R. Brafman and M. Tennenholtz. Modeling agents as qualitative decision makers. *Artificial Intelligence*, 94:217–268, 1997.

[Broersen *et al.*, 2002] J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre. Goal generation in the BOID architecture. *Cognitive Science Quarterly*, 2(3-4):428–447, 2002.

[Cohen and H.Levesque, 1990] Cohen and H.Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

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

[Doyle and Wellman, 1991] J. Doyle and M. P. Wellman. Preferential semantics for goals. In *Proceedings of AAAI-91*, pages 698–703, 1991.

[Doyle, 1991] J. Doyle. Rationality and its rules in reasoning (extended abstract). In *Proceedings of AAAI'91*, pages 1093–1100, 1991.

[Dubois *et al.*, 1998] D. Dubois, H. Prade, and R. Sabbadin. Qualitative decision theory with Sugeno integrals. In *Proceedings of UAI'98*, pages 121–128, 1998.

[Dubois *et al.*, 2002] D. Dubois, H. Fargier, and P. Perny. On the limits of ordinality in decision making. In *Proceedings of KR'02*, pages 133–144, 2002.

[Grahne, 1991] G. Grahne. Updates and counterfactuals. In *Proceedings of KR'91*, pages 269–276, 1991.

[Herzig, 1998] A. Herzig. *Handbook of defeasible reasoning and uncertainty management*, volume 3, chapter Logics for belief base updating, pages 189–231. Kluwer Academic Publishers, 1998.

[Katsuno and Mendelzon, 1991] H. Katsuno and A. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52:263–294, 1991.

[Lehmann, 2001] D. Lehmann. Expected qualitative utility maximization. *Games and Economic Behavior*, 35(1-2):54–79, 2001.

[Pearl, 1993] J. Pearl. From conditional ought to qualitative decision theory. In *Proceedings of the UAI'93*, pages 12–20, 1993.

[Poole, 1992] D. Poole. Decision-theoretic defaults. In *Proceedings of the Ninth Biennial Canadian Artificial Intelligence Conference*, pages 190–197, 1992.

[Rao and Georgeff, 1991] A. Rao and M. Georgeff. Modeling rational agents within a BDI architecture. In *Proceedings of KR'91*, pages 473–484, 1991.

[Tan and Pearl, 1994] S.W. Tan and J. Pearl. Specification and evaluation of preferences for planning under uncertainty. In *Proceedings of KR'94*, 1994.

[Thomason, 2000] R. Thomason. Desires and defaults: a framework for planning with inferred goals. In *Proceedings of KR'00*, pages 702–713, 2000.

[van der Torre and Tan, 1997] L. van der Torre and Y. Tan. The many faces of defeasibility in defeasible deontic logic. In D. Nute, editor, *Defeasible Deontic Logic*, volume 263 of *Synthese Library*, pages 79–121. Kluwer, 1997.