Goals, Desires, Utilities and Preferences

Leendert van der Torre[1] and Emil Weydert[2]

[1] Vrije Universiteit
de Boelelaan 1081a, 1081 HV Amsterdam, the Netherlands
http://www.cs.vu.nl/~torre/
torre@cs.vu.nl
[2] Max Planck Institute for Computer Science
Im Stadtwald, D-66123 Saarbrücken, Germany
http://www.mpi-sb.mpg.de/~weydert/
weydert@mpi-sb.mpg.de

Abstract. In this paper we study the logic of goals, which are formalized as desires with an utilitarian semantics. In our framework goals have a dual character, because they are constraints on utility functions as well as constructors of these utility functions. The non-monotonic reasoning related to the constructors reflects that goals are used as heuristic approximations of preferences in decision making and planning. Moreover, our framework is based on bipolar additive preferences, where bipolarity means that goals can either result in a gain of utility if achieved, or a loss of utility if not achieved. The framework is used to illustrate different types of context-dependence and conflicts of goals.

1 Introduction

Decision theory has become widely accepted in the AI community as a useful framework for planning and decision making [9]. In the context of *qualitative* decision theory [15, 18, 2] recently several *logics* for goals and desires have been proposed [8, 7, 3, 15, 17, 16, 14, 12, 20–22]. Most of these formalizations follow the thesis of Wellman and Doyle [8, p.698] that the relative preference over the possible results of a plan constitutes the fundamental concept underlying the objectives of planning an decision making, with goals serving as a computationally useful *heuristic approximation of these preferences* [6]. For example, goals are used to communicate desires in a compact and efficient way [10]. In this paper we identify goals with desires. In particular, we follow Lang's proposal [14] of an explicit construction of the agent's preferences via the definition of a class of utility functions satisfying the desires. Surprisingly, our study of goals as desires with an utilitarian semantics shows that the formalization of goals is much harder than we expected beforehand, because this formalization of goals differs from well-known and studied formalisms like conditional logic, counterfactuals and deontic logic. There are several reasons why this becomes necessary.

Conflict. To begin with, goals only impose partial preferences, i.e. preferences given some objective and given some context. As a consequence, goals with overlapping contexts can conflict, because objectives can conflict. (For example, to minimize time a tank has to be filled quickly, but to minimize loss it must be filled slowly [4]. This cannot easily be formalized in standard formalisms.)

Paradox. There are several problems with reasoning with preferences, such as for example von Wright's independent preferences problem [25], the formalization of the specificity principle (in case of conflict, specific rules override more general ones) and Arrow's impossibility results [1]. Some of these problems will be discussed later in this paper.

Multipolarity. Goals can result in either a gain or a loss of utility, maybe even according to different incommensurable dimensions of utility, which we call multi-polarity.

Constraints and construction. The use of goals as heuristic approximations of preferences in the framework of decision making and planning has several implications. First of all, they induce constraints on the utility distributions. But in addition, they provide constructive operators to build up the most plausible interpretations of the given goals in the context of other goals.

We follow Lang's proposal of an explicit construction of the agent's preferences, but our work deviates from Lang's in two respects. The similarity is that the agent's preferences are constructed as the addition of local utility functions assigned to each desire. The local utility function formalizes a weighted reason to prefer or disprefer states, and the addition of the local utility functions combines the reasons and weights. We show how this is in particular useful to analyze conflicting and contradictory goals. The first distinction is that our framework extends Lang's framework in important ways. Our framework can be used to analyze conflicting goals, because they are consistent, and goals can be combined in much more different ways, because of the multi-polarity. A comparative table is given at the end of this paper. The second and more important distinction is that Lang uses his framework to discuss the relation between desires, utilities, knowledge and feasibility, whereas we use the framework to discuss the relation between goals, desires, utilities and preferences. In particular, he argues that a distinction has to be made between necessary and contingent knowledge, following a suggestion from [19], and he discusses how incomplete knowledge influences decisions. In contrast, we discuss how the meaning of goals depends on the context of other goals.

A well-known example of the context-dependence of goals we discuss in this paper is related to the specificity principle. In most cases the strengths of goals are unrestricted, but if a goal is more specific than and conflicting with another goal, then the strength of the former is higher than the strength of the latter. In general, strengths of goals are restricted to resolve conflicts. Moreover, there are types of context-dependence that do not depend on the strengths of goals or conflicts between goals, as shown below. In our framework we therefore distinguish between the strengths of goals and what we call their impact. The following example illustrates the two different types of context-dependence and the distinction between strength and impact. Later in this paper, in Example 3 and 4, these examples are given an intuitive interpretation and they are discussed in detail. Goals 'if a then achieve b' are written as $a \rightarrowtriangle b$, the utility of a certain state a is written as $u(a)$ and we consider goals $a \rightarrowtriangle b$ that only result in utility loss $u(a \wedge \neg b) = -\alpha$ (called loss goals), where α is a positive number.

Example 1 (Strength vs impact). Consider the two sets $S_1 = \{\top \rightarrowtriangle \neg s, i \rightarrowtriangle s\}$ and $S_2 = \{(p \vee c) \rightarrowtriangle p, (c \vee h) \rightarrowtriangle c\}$, with background knowledge $\top \leftrightarrow (p \vee c \vee h)$, $\neg(p \wedge c)$, $\neg(p \wedge h)$, $\neg(c \wedge h)$ (the three variables p, c and h are mutually exclusive and exhaustive), and with for each set strengths s_1 and s_2, and impacts (utility losses) α_1 and α_2. We later show that under certain conditions (as satisfied by e.g. Lang's setting) in both sets the meaning of the second goal depends on the first goal. This is represented in Figure 1. The second goal of S_1 is conflicting with and more specific than the first one, and it therefore overrides the first one according to the specificity rule. We have $s_2 > s_1$, $\alpha_1 = s_1$, $\alpha_2 = s_2$ and thus $u(\neg i \wedge \neg s) = 0 > u(i \wedge s) = u(\neg i \wedge s) = -\alpha_1 = -s_1 > u(i \wedge \neg s) = -\alpha_2 = -s_2$. However, the second goal of S_2 may not be stronger than the first one, because the second goal is not more specific than the first. Nevertheless, its impact is higher than the impact of the first goal. In our framework the impact of the second goal can be higher than the impact of the first goal, when we have $s_2 = s_1$, $\alpha_1 = s_1$, $\alpha_2 = s_1 + s_2$ and thus $u(p) = 0 > u(c) = -\alpha_1 = -s_1 > u(h) = -\alpha_2 = -s_1 - s_2$. It is the loss goal $c \vee h \rightarrowtriangle c$ that illustrates the distinction between strength and impact. Its strength (s_2) is the difference between achieving the goal ($u(h)$) and not achieving the goal ($u(c)$), and its impact (α_2) is the utility of not achieving the goal ($u(h)$).

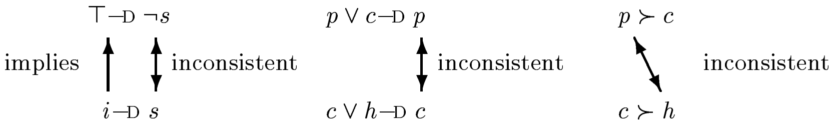

Fig. 1. Context-dependent goals: specificity and conflict-free

The first type of context-dependence (set S_1) is a consequence of resolving conflicts and the second type of context-dependence (set S_2) follows from the relation between desires and preferences; it is *not* a consequence of resolving conflicts. As represented in Figure 1, the representation of goals suggests that S_2 is a conflict, because the consequents of the goals are contradictory. However, this is not a sufficient reason for two goals to conflict. In contrast, the set S_2 is conflict free, because the antecedents of both goals are only true if c is the case, and in that case the first goal can no longer be achieved. This is especially clear if the goals are rewritten as a set of preferences. A goal $b \rightarrowtriangle a$ is formalized by a preference of the achieved goal $a \wedge b$ over the non-achieved goal $\neg a \wedge b$, which can be written as $b \rightarrowtriangle a =_{def} a \wedge b \succ \neg a \wedge b$. Hence, the set S_2 can also be written as $\{p \succ c, c \succ h\}$. The representation of preferences reveals that S_2 is conflict free, because $p \land h$ is inconsistent. There are not two states such that the first is preferred to the second one by the first preference, and vice versa by the second preference. The reason that the impact of $(c \lor h) \relbar\!\mathrm{D}\ c$ is higher than the impact of $(p \lor c) \relbar\!\mathrm{D}\ p$ also follows from the formalization of the goals as preferences. The context-dependent meaning of the goals is such that the derived utility function is $u(p) > u(c) > u(h)$ and reflects the preferences.

The paper is organized as follows. First we discuss the use of goals as heuristic approximations of preferences in decision making and planning and we propose a framework of bipolar additive preferences, where the bipolarity gives more flexibility when representing and combining preferences. Then, we discuss the first type of context dependence by discussing the specificity example. While discussing conflicting goals we also illustrate that goals are constraints on utility functions as well as constructors of utility functions and we illustrate how additivity is used to formalize different types of conflicts. Finally we discuss the second type of context-dependence of goals by discussing several conflict-free sets of goals.

2 The formal framework

2.1 The context of the framework

In the usual approaches to planning in AI, a planning agent is provided with a description of some state of affairs, a *goal state* or a set of goal states, and charged with the task of discovering (or performing) some sequence of actions to achieve that goal. Conditional or context-sensitive goals are formalized with basic concepts provided by decision theory [4, 8, 3]. 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 [13]. This is usually referred to as the foundations of decision theory. The formalization of goals is only concerned with the value and utility functions themselves, because on first approximation they only have an utilitarian or preference-based semantics.

Decision theory does not only provide a semantic framework for conditional goals, but also a more general setting for planning. 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. Haddawy and Hanks [10] 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. Also, in practice, the utilities and probabilities are not known, and we therefore cannot directly use decision theory and maximum utility. Like Haddawy and Hanks our goal is to integrate the two approaches into a theory that preserves decision theory's model of choice while exploiting the computational techniques of symbolic planners as well as qualitative nonmonotonic inference techniques. In order to do so we must confront the relationship between a symbolic planner's goals and a decision-theoretic agent's utility function. Unlike Haddawy and Hanks, however, we do not restrict ourselves to dead-line goals.

In this paper, we only consider the situation where goals 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 simplified version of it, for computational reasons and because we do not know it ourselves. In our semantical interpretation, we concentrate exclusively on the role of expressing desirability –as opposed to intentionality (commitment to pursue) [5]– recognizing that the result is only a partial account of the use of goals in planning systems. To emphasize this desirability aspect of goals, we write $a \multimap_{\mathrm{D}} b$ for the conditional goal 'if a then achieve b.'

The relation between goals, defaults, expectations, facts and decisions is represented in Figure 2. The input of the one-shot (i.e. static) decision problem consists of goals (or desires), defaults (or beliefs), expectations and facts. The input is transformed – nonmonotonically – 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 focus on the relation between goals and utility functions.

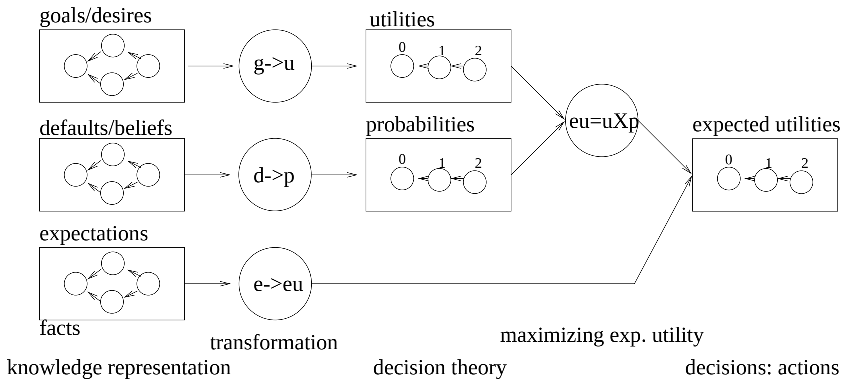

Fig. 2. Goals, defaults, expectations and decisions

Expectations are introduced in Figure 2, because in order to allow for reasonable inferences to be made, a logic that allows tradeoffs of likelihood and preference to be expressed in a qualitative fashion is desirable [3]. Boutilier gives the following example. If I instruct my robot that it should run across the street (instead of crossing at the crosswalk) to save three minutes while fetching my coffee, it can safely deduce that running across the street is worth the risk if a courier deadline is involved. I have implicitly calibrated part of its preference and normality rankings with each other. Goals which only consider utility rankings, as discussed in this paper, are expectations under the assumption that all worlds are equally plausible.

2.2 The framework

The present account of monotonic and nonmonotonic reasoning with and about goals or desires is part of a general qualitative decision theoretic framework based on ranking functions (or possibility functions). It generalizes Lang's approach, enabling us to implement different useful philosophies for generating and handling utility functions. For the sake of clarity, in this paper, we are going to concentrate on one-dimensional utility functions, i.e. with a single scale to measure goodness and badness.

Let $\mathcal{L}$ be a propositional language and W the set of all the propositional valuations (worlds) over $\mathcal{L}$. We are now going to introduce different types of conditional desire or goal expressions of the form $a \multimap\!\mathrm{D}_p^i\ b$ or $a \multimap\!\mathrm{D}^i\ b$, where $a, b \in \mathcal{L}$ are propositional formulas, p is a strictly positive real number or ∞ (or it is dropped altogether), and $i \in \{+, -\}$ denotes the polarity. $a \multimap\!\mathrm{D}_p^+\ b$ is meant to express the desire of (prima facie) strength p to achieve b in the context a. $a \multimap\!\mathrm{D}_p^-\ b$ stands for the desire of (prima facie) strength p to avoid $\neg b$ in the context a. The first is called a gain goal, the second one a loss goal. The nonindexed goal $a \multimap\!\mathrm{D}^i\ b$ may be interpreted as an indexed one, $a \multimap\!\mathrm{D}_p^i\ b$, where $0 < p \leq \infty$ has been left undetermined. Let $\mathcal{L}(\multimap\!\mathrm{D}\ )$ be the set of all goal formulas over $\mathcal{L}$. A goal base Γ (a desire specification in Lang's terminology) is a finite subset of $\mathcal{L}(\multimap\!\mathrm{D}\ )$.

Utility functions are mappings $u : W \to [-\infty, \infty]$ from worlds to real numbers and infinities. They constitute our basic semantic entities. Each u induces a preorder $\geq_u$ on W defined by $w_1 \geq_u w_2$ iff $u(w_1) \geq u(w_2)$. On a first level, goals may be interpreted as constraints on utility functions. We define a model set function $\mathcal{M}$ which associates with every goal base Γ the set of those utility functions sanctioned by Γ. Here we assume classical monotonic semantics. In particular $\mathcal{M}(\Gamma) = \bigcap\{\mathcal{M}(\{\gamma\}) \mid \gamma \in \Gamma\}$. It is not difficult to see that this model class function is monotonic. We write $\mathcal{M}(\gamma) = \mathcal{M}(\{\gamma\})$ and $u \models \Gamma$ for $u \in \mathcal{M}(\Gamma)$. $\multimap\!\mathrm{D}_p^+$ and $\multimap\!\mathrm{D}_p^-$ will only be distinguished in the context of the nonmonotonic semantics.

- $u \models a \multimap\!\mathrm{D}_p^{+,-}\ b$ iff $\max\{u(w) \mid w \models a \wedge \neg b\} + p \leq \max\{u(w) \mid w \models a \wedge b\}$
- $u \models a \multimap\!\mathrm{D}^{+,-}\ b$ iff $\max\{u(w) \mid w \models a \wedge \neg b\} < \max\{u(w) \mid w \models a \wedge b\}$

Because $p > 0$, this means that the the preferred a states are b states. A more cautious possibility would be to replace the second max by min. Note that both definitions make strongly conflicting desires inconsistent. That is, if $\Gamma = \{\top \multimap\!\mathrm{D}_p^i\ a, \top \multimap\!\mathrm{D}_p^i\ \neg a\}$, then $\mathcal{M}(\Gamma) = \emptyset$. If we want to avoid this trivialization, we have to weaken the standard model set notion $\mathcal{M}$. In the following, for any $a \in \mathcal{L}$, we are going to identify $u(a)$ with $\max\{u(w) \mid w \models a\}$.

We are mainly interested in the non-monotonic semantics of goals. It is determined by a model set function function $\mathcal{PM}$ which maps every goal base Γ to a set of preferred or most plausible utility functions verifying Γ, at least in conflict-free contexts. The idea we are going to exploit is that goals do not only provide constraints on utility rankings, but also define constructors, building blocks, to be used in the context of the given construction philosophy implemented by $\mathcal{PM}$. For nonmonotonic model set functions, of course, we do no longer have $\mathcal{PM}(\Gamma) = \bigcap\{\mathcal{PM}(\{\gamma\}) \mid \gamma \in \Gamma\}$. Nonetheless, it seems reasonable to begin by looking at the simplest kind of goal bases, namely $\mathcal{PM}(\{a{-}\mathrm{D}_p^+\ b\})$, $\mathcal{PM}(\{a{-}\mathrm{D}_p^-\ b\})$, and the indexfree versions.

The idea is to start from the uniform, neutral utility ranking u_0, with $u_0(w) = 0$ for all $w \in W$, and to modify it only as much as necessary to arrive at a ranking in $\mathcal{M}(\gamma)$, i.e. to verify γ. For singleton goal bases, this is relatively easy to achieve. For parametrized goals (with index p) we get even uniquely determined preferred models. We adopt the following stipulations.

- **Parametrized gain goals :**
  $\mathcal{PM}(\{a{-}\mathrm{D}_p^+\ b\}) = \{u\}$ with
  $u(w) = 0$ for $w \models a \wedge \neg b$, $u(w) = p$ for $w \models a \wedge b$, $u(w) = 0$ for $w \models \neg a$.
- **Parametrized loss goals :**
  $\mathcal{PM}(\{a{-}\mathrm{D}_p^-\ b\}) = \{u\}$ with
  $u(w) = -p$ for $w \models a \wedge \neg b$, $u(w) = 0$ for $w \models a \wedge b$, $u(w) = 0$ for $w \models \neg a$.

So, for each parametrized goal γ, we have a distinguished utility function u_γ. Non-parametrized goals, that is $a{-}\mathrm{D}^+ b$ or $a{-}\mathrm{D}^- b$, may be interpreted, either as parametrized goals with variable parameters (but assume their finiteness), or as parametrized goals with a default parameter, w.l.o.g. 1.

- **Non-parametrized gain goals :**
  $\mathcal{PM}(\{a{-}\mathrm{D}^+\ b\}) = \{u_{a-\mathrm{D}_p^+\ b} \mid 0 < p < \infty\}$, alternatively,
  $\mathcal{PM}(\{a{-}\mathrm{D}^+\ b\}) = \{u_{a-\mathrm{D}_1^+\ b}\}$.
- **Non-parametrized loss goals :**
  $\mathcal{PM}(\{a{-}\mathrm{D}^-\ b\}) = \{u_{a-\mathrm{D}_p^-\ b} \mid 0 < p < \infty\}$, alternatively,
  $\mathcal{PM}(\{a{-}\mathrm{D}^-\ b\}) = \{u_{a-\mathrm{D}_1^-\ b}\}$.

The naive approach for generalizing this local solution to arbitrary goal sets Γ would be to construct the distinguished utility function(s) of Γ by adding the local utility functions of each goal in the set. However, in general, this simple construction does not satisfy the constraints of the monotonic semantics, as follows directly from Example 1. For the set $S_2 = \{(p \vee c){-}\mathrm{D}_{s1}^-\ p, (c \vee h){-}\mathrm{D}_{s2}^-\ c\}$ (the three variables p, c and h are mutually exclusive and exhaustive), we would have $u(p) = 0$, $u(c) = -s_1$ and $u(h) = -s_2$. But the constructed utility function does not satisfy $(c \vee h){-}\mathrm{D}_{s2}\ c$, because $u(c) - u(h) < s_2$.

The impact of a goal $a{-}\mathrm{D}_p^{+,-}\ b$ in the context of an utility function $u \in \mathcal{PM}(\Gamma)$ is $u(a \wedge \neg b)$. It may be different from its strength, as illustrated in the previous example, but it is still related to it. To incorporate this kind of examples, we propose a more complex type of construction which makes a distinction between strength and utility impact. But first, we introduce some additional notation. For a goal $\gamma = (a{-}\mathrm{D}_p^{+,-}\ b)$, let $\gamma^+ = a \wedge b$ (achievement), $\gamma^- = a \wedge \neg b$

(non-achievement), and $\gamma^0 = \neg a$ (neutrality). For each $a \in \mathcal{L}$, let i_a be the utility ranking u with $u(w) = 1$ for $w \models a$ and $u(w) = 0$ for $w \not\models a$.

There are three basic principles guiding our choice of a preferred model set function $\mathcal{PM}$ for nonmonotonic reasoning with arbitrary goal bases, notably uniformity, constructibility, and minimality. Supraclassicality, that is $\mathcal{PM}(\Gamma) \subseteq \mathcal{M}(\Gamma)$, is not present because we do not want to exclude strongly conflicting goal sets with $\mathcal{M}(\Gamma) = \emptyset$. To begin with, we would like to avoid unmotivated utility gradients. For unary goal sets $\{\gamma\}$, we have stipulated that $\mathcal{PM}(\gamma)$ only includes utility rankings which are constant within γ^+, γ^- and γ^0. On a more general level, we proceed as follows.

Uniformity. Worlds $w, w' \in W$ belonging to the same $\gamma^+, \gamma^-, \gamma^0$ for each $\gamma \in \Gamma$ should be handled equivalently by each $u \in \mathcal{PM}(\Gamma)$, that is $u(w) = u(w')$.

Uniformity immediately follows from the next postulate, which we consider the most important one.

Constructibility. $u \in \mathcal{PM}(\Gamma)$ implies that there are $x_\gamma^+ \geq 0, x_\gamma^- \leq 0, x_\gamma^0$ with $u = \Sigma_{\gamma \in \Gamma}(x_\gamma^+ i_{\gamma^+} + x_\gamma^- i_{\gamma^-} + x_\gamma^0 i_{\gamma^0})$.[1]

Note that this approach brings in a strong element of nonmonotonicity, because the set of constructible utility rankings *increases* when the set of desires increases. If $\Gamma = \{a \text{—}\mathrm{D}_p^+\, b\}$, our standard approach for unary goal bases gives us $(x_\gamma^+, x_\gamma^-, x_\gamma^0) = (p, 0, 0)$, and if $\Gamma = \{a \text{—}\mathrm{D}_p^-\, b\}$, we get $(x_\gamma^+, x_\gamma^-, x_\gamma^0) = (0, -p, 0)$. If we do not insist on extreme robustness we may always assume $x_\gamma^0 = 0$. But to obtain reasonable results, we have to restrict the set of admissible (x_γ^+, x_γ^-) even further. We therefore propose a minimality principle. Let $Constr(\Gamma) = \{\Sigma_{\gamma \in \Gamma} x_\gamma^+ i_{\gamma^+} + x_\gamma^- i_{\gamma^-} \mid x_\gamma^+, x_\gamma^- > 0\}$ be the set of all simply constructible u. We assume that the constraints are minimally realized in $\mathcal{PM}(\Gamma)$. In particular, this works for single goal examples.

Minimality. $\forall u \in \mathcal{PM}(\Gamma) \forall v \in \mathcal{PM}(\Gamma) \exists \gamma \in \Gamma \exists \gamma' \in \Gamma \; u \neq v \vdash u(\gamma^+) + u(\gamma^-) < v(\gamma^+) + v(\gamma^-) \;\wedge\; u(\gamma'^+) + u(\gamma'^-) > v(\gamma'^+) + v(\gamma'^-)$

Let $Minconstr(\Gamma)$ be the set of all those utility rankings which are minimally constructible over Γ. Our summarizing requirement is given below.

Minimal constructibility. $\mathcal{PM}(\Gamma) \subseteq Minconstr(\Gamma)$

In the following, we are going to base our considerations on the most cautious implementation of this principle. $\mathcal{PM}_{mc}(\Gamma) = Minconstr(\Gamma)$. In fact, as long as we stay away from inconsistent Γ, we may consider the natural strengthening. $\mathcal{PM}_{mcc}(\Gamma) = Minconstr(\Gamma) \cap \mathcal{M}(\Gamma)$.

In the following two sections we analyze how conflicting desires are handled in this framework, and how preferences are combined.

[1] If u, v are utility rankings and $x \in [-\infty, \infty]$, then $xu + yv$ is meant to be a linear combination of functions.

3 Conflicting goals

In this section we show how the impact of a goal can be restricted to resolve conflicts. In particular, we show how the impact of a goal depends on the impact of more general and conflicting goals. However, first we illustrate the dual character of goals in our framework by discussing a simple example of conflicting goals.

3.1 Goal = constraint + constructor

Figure 2 illustrates that goals are used to determine utility functions. In particular, each goal induces a constraint on utility functions. However, that is only part of the story, for the following three reasons.

- First, the constraint has no impact on decisions if the most preferred states are highly unlikely.
- Second, there seem to be rather few logical laws governing preference which have a strong intuitive support [2].
- Third, this definition does not reflect that goals are computationally useful heuristic approximations of preferences over the possible results of a plan.

We therefore introduce a second consequence of goals, which we call constructors. In case of conflict, the constructors represent weighted reasons to prefer or disprefer states, and conflicts are resolved by combining these reasons and weights [11]. It is here that the additivity of utilities plays a crucial role.[2] The constructors were formally defined in our framework in Section 2.2, and the underlying mechanism is illustrated in the following example of the two goals to be wealthy and healthy. Complicating factors such as strongly conflicting and conditional goals are not yet considered.

Example 2 (Independent goals). Consider the two independent loss goals 'to be wealthy' $\top \multimap \mathrm{D}^{-}_{s1}\, w$ and 'to be healthy' $\top \multimap \mathrm{D}^{-}_{s2}\, h$, where $\top$ stands for a tautology and the second goal is more important ('stronger') than the first one $s_2 > s_1$. This simple example is already interesting, because a definition like 'p is a goal if the utility of all the p worlds is higher than the utility of each $\neg p$ world' has the problem that the two goals conflict when comparing 'wealthy and unhealthy' $(w \wedge \neg h)$ and 'poor and healthy' $(\neg w \wedge h)$, as was already observed by Von Wright [25]. First, consider the constraint and constructor of the goal to be wealthy.

[2] This approach based on additive utilities can be compared with an approach based on ceteris paribus preferences [8, 7], in which it is attempted to construct frames such that preferences do not conflict. They are closely related, because mutual preferential independence (essentially ceteris paribus) entails additive value functions [13]. We weaken the constraint part (referring only to preferred states), and put all of the more global effects of goals in the (nonmononic) constructor part.

Constraint: the utility of the preferred w world is at least s_1 higher than the utility of the preferred $\neg w$ world. Consequently, the most preferred worlds are w worlds;

Constructor: the utility of the w worlds is $x_1^+ = 0$ and the utility of $\neg w$ worlds is $x_1^- = -s_1$.

Likewise, the constraint for the second goal entails that the most preferred worlds are h worlds and the heuristic is that the utility of h worlds is $x_2^+ = 0$ and the utility of $\neg h$ worlds is $x_2^- = -s_2$. These preferred utility functions in the absence of other goals are the local utility functions of a goal. Now consider the combination of the two goals as represented in Figure 3, where the numbers stand for worlds in a preference ordering: $w_1 \models w \wedge h$, $w_2 \models \neg w \wedge h$, $w_3 \models w \wedge \neg h$ and $w_4 \models \neg w \wedge \neg h$.

Constraint: the most preferred worlds are $w \wedge h$ worlds, and their utility is at least s_1 higher than the utility of the preferred $\neg w$ worlds and at least s_2 higher than the utility of the preferred $\neg h$ worlds.

Constructor: the utility of a world is the sum of the local utilities:

$w \models w \wedge h$	$\neg w \wedge h$	$w \wedge \neg h$	$\neg w \wedge \neg h$
$u(w) = 0$	$-s_1$	$-s_2$	$-s_1 - s_2$

We have that $\neg w \wedge h$ is preferred to $w \wedge \neg h$, as desired.

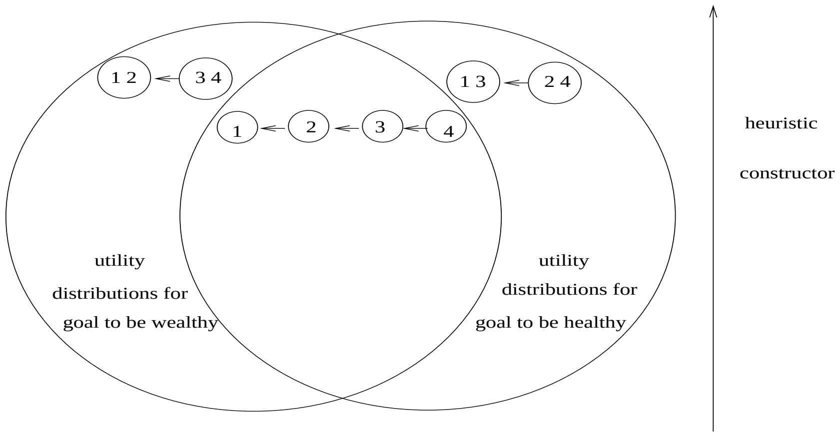

Fig. 3. Combining wealthy and healthy

The construction illustrated in Figure 3 is a kind of non-monotonic reasoning. However, it is different in spirit from the non-monotonic reasoning techniques used by for example Boutilier [3] (worlds gravitate towards the ideal) and Tan and Pearl [17, 16] (most compact models), because it is more constructive. The heuristic operators can be interpreted dynamically as follows. Assume that the initial state is the state in which all worlds have utility 0. Given a loss goal for p with strength s, the heuristic of a local utility function can be interpreted as the shift of all $\neg p$ worlds by $-s$. The addition of the two local utility functions for wealthy and healthy is the application of both operators.

3.2 Specificity

The second important benchmark problem we consider is the formalization of a specificity set, i.e. a set of two conflicting desires where one of the desires is more specific than the other one. It has been observed by several authors (see e.g. [2]) that ceteris paribus preferences cannot consistently formalize a specificity set.

Example 3 (Specificity). Consider the goals 'not having surgery' $\top{-}\!\mathrm{D}_{s1}\, \neg s$ and 'to have surgery if it improves one long-term's health' $i{-}\!\mathrm{D}_{s2}\, s$ [2]. The constructable utility rankings are as follows, with the following two constraints.

$w \models \neg i \wedge \neg s$	$i \wedge s$	$\neg i \wedge s$	$i \wedge \neg s$
$u(w) = x_1^+ + x_2^0$	$x_1^- + x_2^+$	$x_1^- + x_2^0$	$x_1^+ + x_2^-$

$\max(x_1^+ + x_2^0, x_1^+ + x_2^-) \geq s_1 + \max(x_1^- + x_2^+, x_1^- + x_2^0)$
$x_1^- + x_2^+ \geq s_2 + x_1^+ + x_2^-$

Minimal constructible. From the second constraint follows that the impact of more specific and conflicting goals is higher than their strengths: $x_2^+ - x_2^- > s_2 + x_1^+ - x_1^-$. The minimal construction is $x_1^+ - x_1^- = s_1$ and $x_2^+ - x_2^- = s_1 + s_2$.

Moreover, consider a third goal $i \wedge j{-}\!\mathrm{D}_{s3}\, \neg s$.

$w \models \neg i \wedge \neg s$	$i \wedge \neg j \wedge s$	$i \wedge j \wedge \neg s$
$u(w) = x_1^+ + x_2^0 + x_3^0$	$x_1^- + x_2^+ + x_3^0$	$x_1^+ + x_2^- + x_3^+$
$w \models \neg i \wedge s$	$i \wedge \neg j \wedge \neg s$	$i \wedge j \wedge s$
$u(w) = x_1^- + x_2^0 + x_3^0$	$x_1^+ + x_2^- + x_3^0$	$x_1^- + x_2^+ + x_3^-$

$\max(x_1^+ + x_2^0 + x_3^0, x_1^+ + x_2^- + x_3^+, x_1^+ + x_2^- + x_3^0) \geq s_1 + \max(x_1^- + x_2^+ + x_3^0, x_1^- + x_2^0 + x_3^0, x_1^- + x_2^+ + x_3^-)$
$\max(x_1^- + x_2^+ + x_3^0, x_1^- + x_2^+ + x_3^-) \geq s_2 + \max(x_1^+ + x_2^- + x_3^+, x_1^+ + x_2^- + x_3^0)$
$x_1^+ + x_2^- + x_3^+ \geq s_3 + x_1^- + x_2^+ + x_3^-$

Minimal constructible. For loss goals, the minimal constructible ranking is $x_1^- = -s_1$, $x_2^- = -s_1 - s_2$ and $x_3^- = s1 - s_2 - s_3$.

In the following section we consider the second type of context-dependence of goals.[3]

[3] The standard definition of goals is quite strong. Consider the two sets of goals $S_1 = \{\top{-}\!\mathrm{D}\, \neg f, d{-}\!\mathrm{D}\, f, \top{-}\!\mathrm{D}\, d\}$ and $S_2 = \{\top{-}\!\mathrm{D}\, j, \top{-}\!\mathrm{D}\, f, \top{-}\!\mathrm{D}\, \neg(j \wedge f)\}$. The set S_1 formalizes the goals of an animalloving construction robot when f is read as 'there is a fence around the cottage' and d as 'there is a dog.' S_2 formalizes that Sue likes John and she also likes Fred. She might prefer to be married to John over not, and also prefer to be married to Fred over not, though she would not like it to be married to both of them [2]. Both sets are inconsistent, allthough intuitively they are consistent.

4 Combining preferences

In this section we discuss the relation between goals and preferences. In particular, we discuss a very restricted but nevertheless interesting subset of goals called non-conflicting goals.

Definition 1. *A set of goals $S = \{a_1 \multimap b_1, \ldots, a_n \multimap b_n\}$ is called conflict free if the graph defined by $\{(w_1, w_2) \mid \exists a_i \multimap b_i \in S s.t. w_1 \models a_i, w_2 \models b_i\}$ is cycle free.*

Example 3 illustrates that more specific and overriding goals have a higher utility loss than more general goals, and therefore override them. Now, the question can be raised whether we have that a goal $a \multimap b$ has a higher utility loss than $a_1 \multimap b_1$ *only if* it is more specific and conflicting. In general, the question can be raised whether goals have a higher utility loss than other goals only to resolve conflicts. The following example shows that this is not the case, because it also occurs in conflict free goal sets. In the example we assume that we can represent so-called background knowledge, expressing that some worlds are physically impossible, see [14, 19].

Example 4. Consider the goals $S = \{p \succ_{s1} c = (p \vee c) \multimap_{s1} p$ and $c \succ_{s2} h = (c \vee h) \multimap_{s2} c\}$ with background knowledge $\top \leftrightarrow (p \vee c \vee h)$, $\neg(p \wedge c)$, $\neg(p \wedge h)$, $\neg(c \wedge h)$ (the three variables are mutually exclusive and exhaustive). Intuitively the three variables represent different ways to spend the evening: p can be read as going to a party, c as going to the cinema and h as staying home.

$w \models$	p	c	h
$u(w) =$	$x_1^+ + x_2^0$	$x_1^- + x_2^+$	$x_1^0 + x_2^-$

$x_1^+ + x_2^0 \geq s_1 + x_1^- + x_2^+$
$x_1^- + x_2^+ \geq s_2 + x_1^0 + x_2^-$

Loss goals. The minimal construction is $x_1^- = -s_1$ and $x_2^- = -s_1 - s_2$. The impact of the second desire is higher than its strength, although the second goal is not more specific nor conflicting with the first.[4] In the context of the first goal, the loss of utility of the second goal is the sum of the strengths of both goals.

Gain goals. The minimal construction is $x_1^+ = s_1 + s_2$ and $x_2^+ = s_2$.

The preferences are closed under transitivity in the sense that the strength of the derived preference $p \succ h$ is $s_1 + s_2$. Obviously, this transitive closure only holds non-monotonically. Moreover, consider the goals $p \succ c$, $c \succ h$, $h \succ w$, where w can be read as going to work.

[4] It is illuminating to compare the desires in Example 4 with the two goals $(p \vee h) \succ c = (p \vee c \vee h) \multimap (p \vee h)$ and $c \succ h = (c \vee h) \multimap c$. The second goal is more specific than the first one and we have $x_2^+ - x_2^- > x_1^+ - x_2^-$, just like in Example 3. The logic does not distinguish between this set of goals and the set of goals in Example 4, because the violation of the first goal is in both cases c. The distinction between the two goals $(p \vee c \vee h) \multimap (p \vee h)$ and $(p \vee c) \multimap p$ is that the first is fulfilled if $p \vee h$ is true, whereas the second is fulfilled if p is true.

p	c	h	w
$x_1^+ + x_2^0 + x_3^0$	$x_1^- + x_2^+ + x_3^0$	$x_1^0 + x_2^- + x_3^+$	$x_1^0 + x_2^0 + x_3^-$

Loss goals. The minimal construction is $x_1^- = -s_1$, $x_2^- = -s_1 - s_2$ and $x_3^- = -s_1 - s_2 - s_3$.

Gain goals. The minimal construction is $x_1^+ = s_1 + s_2 + x_3$, $x_2^+ = s_2 + s_3$ and $x_3^+ = s_3$.

The following more complex examples of non-conflicting goals illustrate the distinction between gain and loss goals.

Example 5. Consider the set of goals $S = \{p \succ_{s1} (c \vee h), h \succ_{s2} w\}$, with background knowledge $\top \leftrightarrow (p \vee c \vee h \vee w)$ and the four variables are mutually exclusive, as illustrated in figure 4 below.

$w \models p$	c	h	w
$u(w) = x_1^+ + x_2^0$	$x_1^- + x_2^0$	$x_1^- + x_2^+$	$x_1^0 + x_2^-$

$x_1^+ + x_2^0 \geq s_1 + \max(x_1^- + x_2^0, x_1^- + x_2^+)$
$x_1^- + x_2^+ \geq s_2 + x_1^0 + x_2^-$

Loss goals. The minimal construction is $x_1^- = -s_1$, and $x_2^- = -s_1 - s_2$.

Gain goals. The minimal construction is $x_1^+ = \max(s_1, s_2)$ and $x_2^+ = s_2$.

Finally, consider the goals $S = \{p \succ_{s1} h, (c \vee h) \succ_{s2} w\}$, also represented in Figure 4.

$w \models p$	c	h	w
$u(w) = x_1^+ + x_2^0$	$x_1^0 + x_2^+$	$x_1^- + x_2^+$	$x_1^0 + x_2^-$

$x_1^+ + x_2^0 \geq s_1 + x_1^- + x_2^+$
$\max(x_1^0 + x_2^+, x_1^- + x_2^+) \geq s_2 + x_1^0 + x_2^-$

Loss goals. The minimal construction is $x_1^- = -s_1$ and $x_2^- = \min(-s_1, -s_2)$.

Gain goals. The minimal construction is $x_1^+ = s_1 + s_2$ and $x_2^+ = s_2$.

The examples show that the impact of a goal not only depends on its strength, but also on its polarity.

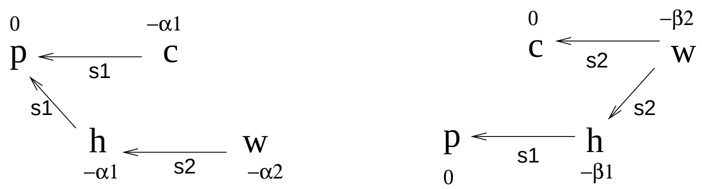

Fig. 4. Loss goals

5 Conclusions

This paper is another step towards a logic for qualitative decision theory. We analyzed goals as desires with an utilitarian semantics in a framework with two important properties. First, the use of goals in the general framework of decision making under uncertainty, as illustrated in Figure 2, implies that goals not only induce constraints but also constructors. These constructors enable a specific type of non-monotonic reasoning [3, 17, 16]. Second, the framework is based on additivity and bipolarity. Our framework is very general. Finding restrictions that give intuitive results for the examples and an extension to multi-dimensional utility functions is subject of further research.

To illustrate bipolarity we introduced a new type of examples in Example 4. These new examples are different from standard benchmark examples used to test logics of defeasible reasoning and logics of deontic reasoning. For example, in default reasoning the standard examples are variants and extensions of Example 3 (the notorious Tweety examples), in deontic (and counterfactual) reasoning the standard examples are of the type 'p ought to be done' and 'if p is not done, then q ought to be done' (so-called contrary-to-duty reasoning) and 'if p is done, then q ought to be done' (related to so-called deontic detachment), and in defeasible deontic reasoning [19, 23, 24] the standard examples are combinations of these classes.[5] However, we think it is an important example for a logic of goals and desires. Moreover, these patterns are also important for applications.[6]

Finally, the following table compares Lang's framework with our extension.

	Lang's framework	Our framework
Polarity	Monopolar	Bipolar
Impact	Loss	Loss and gain
	Impact	Impact and strength
Contradictory	inconsistent	consistent (const)
Knowledge	yes	no
Feasibility	yes	no

[5] The relation between goals and obligations, as formalized in deontic logic, has been observed by several authors, e.g. [15, 3, 14, 2, 18]. Moreover, Lang illustrates his framework by examples of defeasible deontic logic (specificity and contrary-to-duty reasoning). It is left to the reader to verify that the two typical sets $CTD = \{(p \vee c) \multimap_D p, \top \multimap_D (p \vee c)\}$ and $ATD = \{\top \multimap_D p, (c \vee h) \multimap_D c\}$ do not induce constraints on the utility gains and losses, and are thus in that sense less interesting.

[6] The examples contains different goals that describe a linearly ordered set of options (going to a party, going to the cinema, staying home, and least preferred, going to work). The examples show that conditional or context-sensitive goals can be used to formalize objectives that reflect graded criteria, such as time taken to fill a tank or amount of fluid spilled, because they can formalize different degrees of satisfaction of a goal. They are therefore important benchmark problems of a logic of goals, because in realistic planning situations the robot's objectives can be satisfied to varying degrees, and an agent will frequently encounter goals that it cannot achieve [4].

References

1. K.J. Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58:328–346, 1950.
2. F. Bacchus and A.J. Grove. Utility independence in a qualitative decision theory. In *Proceedings of KR'96*, pages 542–552, 1996.
3. C. Boutilier. Toward a logic for qualitative decision theory. In *Proceedings of the KR'94*, pages 75–86, 1994.
4. T. Dean and M. Wellman. *Planning and Control*. Morgan Kaufmann, San Mateo, 1991.
5. J. Doyle. A model for deliberation, action and introspection. Technical Report AI-TR-581, MIT AI Laboratory, 1980.
6. J. Doyle. Rationality and its rules in reasoning (extended abstract). In *Proceedings of the AAAI'91*, pages 1093–1100, 1991.
7. J. Doyle, Y. Shoham, and M.P. Wellman. The logic of relative desires. In *Sixth International Symposium on Methodologies for Intelligent Systems*, Charlotte, North Carolina, 1991.
8. J. Doyle and M.P. Wellman. Preferential semantics for goals. In *Proceedings of AAAI-91*, pages 698–703, Anaheim, 1991.
9. V. Ha and P. Haddawy. Problem-focused incremental elicitation of multi-attribute utility models. In *Proceedings of the UAI'97*, 1997.
10. P. Haddawy and S. Hanks. Representations for decision-theoretic planning: Utility functions for dead-line goals. In *Proceedings of the KR'92*, Cambridge, MA, 1992.
11. J. Hage. *Reasoning with Rules. An Essay on Legal Reasoning and its Underlying Logic*. Law and philosophy library. Kluwer, 1997.
12. Z. Huang and J. Bell. Dynamic goal hierarchies. In *Intelligent Agent Systems, Theoretical and Practical Issues*, LNAI 1209, pages 88–103. Springer, 1997.
13. R.L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Trade-offs*. Wiley and Sons, New York, 1976.
14. J. Lang. Conditional desires and utilities - an alternative approach to qualitative decision theory. In *Proceedings of the ECAI'96*, pages 318–322, 1996.
15. J. Pearl. From conditional ought to qualitative decision theory. In *Proceedings of the UAI'93*, pages 12–20, 1993.
16. S.-W. Tan and J. Pearl. Qualitative decision theory. In *Proceedings of the AAAI'94*, 1994.
17. S.-W. Tan and J. Pearl. Specification and evaluation of preferences under uncertainty. In *Proceedings of the KR'94*, pages 530–539, 1994.
18. R. Thomason and R. Horty. Nondeterministic action and dominance: foundations for planning and qualitative decision. In *Proceedings of the TARK'96*, pages 229–250. Morgan Kaufmann, 1996.
19. L. van der Torre. Violated obligations in a defeasible deontic logic. In *Proceedings of the ECAI'94*, pages 371–375. John Wiley & Sons, 1994.
20. L. van der Torre. Labeled logics of conditional goals. In *Proceedings of the ECAI'98*, pages 368–369. John Wiley & Sons, 1998.
21. L. van der Torre. Phased labeled logics of conditional goals. In *Logics in Artificial Intelligence. Proceedings of the 6th European Workshop on Logics in AI (JELIA'98)*, LNAI 1489, pages 92–106. Springer, 1998.
22. L. van der Torre. Defeasible goals. In *Symbolic and Quantitative Approaches to Reasoning and Uncertainty. Proceedings of the ECSQARU'99*, LNCS 1638. Springer, 1999.

23. L. van der Torre and Y. Tan. Cancelling and overshadowing: two types of defeasibility in defeasible deontic logic. In *Proceedings of the IJCAI'95*, pages 1525–1532. Morgan Kaufman, 1995.
24. L. van der Torre and Y. Tan. The many faces of defeasibility in defeasible deontic logic. In D. Nute, editor, *Defeasible Deontic Logic*, pages 79–121. Kluwer, 1997.
25. G.H. von Wright. *The logic of preference.* Edinburgh University Press, 1963.