

What is a Normative Goal?

Towards Goal-based Normative Agent Architectures

Mehdi Dastani¹ and Leendert van der Torre²

¹ Institute of Information and Computer Sciences, Utrecht University
mehdi@cs.uu.nl

² Department of Artificial Intelligence, Vrije Universiteit Amsterdam
torre@cs.vu.nl

Abstract. In this paper we are interested in developing goal-based normative agent architectures. We ask ourselves the question what a normative goal is. To answer this question we introduce a qualitative normative decision theory based on belief (B) and obligation (O) rules. We show that every agent which makes optimal decisions – which we call a BO rational agent – acts *as if* it is maximizing the set of normative goals that will be achieved. This is the basis of our design of goal-based normative agents.

1 Introduction

Simon [26] interpreted goals as utility aspiration levels, in planning goals have a notion of desirability as well as intentionality [14], and in the Belief-Desire-Intention or BDI approach [6, 23] goals have been identified with desires. Moreover, recently several approaches have been introduced to extend decision making and planning with goal generation. For example, Thomason’s BDP logic [27] extends the BDI approach with goal generation and planning, and Broersen *et.al.*’s BOID architecture [4] elaborates on the goal generation mechanism for more general class of cognitive agents. But what is this thing called goal? Although there are many uses of goals in planning and more recently in agent theory, the ontological status of goals seems to have received little attention.

In this paper we try to find out what a normative goal is by comparing normative decision systems with knowledge-based systems in which decisions are considered to be the result of planning of goals. Of course, such a comparison is complicated by the fact that there are many different kinds of normative and knowledge-based systems. We therefore restrict ourselves to the characterizations illustrated in Figure 1. This figure should be read as follows. First, for our comparison normative systems are decision-making agents which perform practical reasoning [29]. They can formally be described by a reasoning mechanism based on (defeasible) deontic logic which describes the relation between a set of beliefs (*B*) including observations, a set of obligations (*O*), and decisions or actions. If we replace the set of obligations by a set of desires (*D*), then many qualitative decision theories such as [27, 19] also fit this description. Second, knowledge-based systems have as input a knowledge base (*KB*) including observations and goals, and they have as output actions or plans. Knowledge-based systems

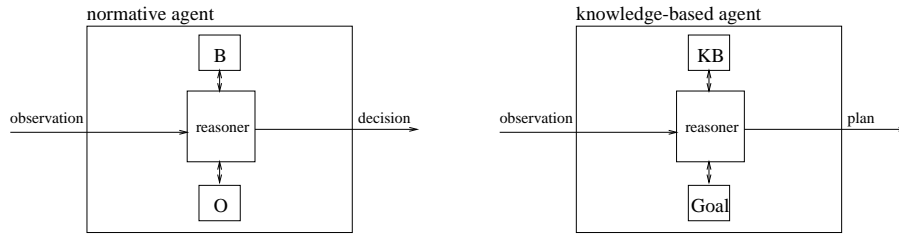


Fig. 1. Agent

have been advocated by Newell and Simon [22, 26] and have been implemented by for example SOAR [17] and ACT [1]. Moreover, also more recent BDI systems like PRS [15] fit this description. The main reasoning task of the knowledge-based system is planning.

How can we compare these two kinds of decision making systems? First we unify the beliefs with the knowledge base, because both represent the motivational attitude of the system. Moreover, we unify decisions with actions and plans. The main problem is the unification of the motivational attitude, the obligations (or, in other qualitative decision theories, the desires) and the goals. Rao and Georgeff [23] proposed, at a very high level of abstraction, that desires and goals can be unified. However, this has been criticized by several authors [16, 13, 10]. An argument against the latter unification is that desires can conflict whereas goals in Rao and Georgeff's framework cannot. Another argument is that goals may be adopted from another agent, whereas desires cannot be adopted. Moreover, desires are more stable than goals [8].

Thomason [27] proposes a logical theory in which desires are a more primitive concept than goals, in the sense that goals can be inferred from desires. Broersen et al. [4] extend this argument to obligations and propose an architecture in which goals can be inferred from desires, intentions and obligations and in which goal generation gets a prominent place. For our comparison, we define goal generation as a theory with input beliefs, observations and obligations, and as output goals. Now we can use the output of goal generation as input for the knowledge-based system to infer decisions or actions. The idea can be paraphrased by:

Goal-based decision making is goal generation plus goal-based planning

This decomposition of decision making in goal generation and planning raises several questions, such as:

- How to represent beliefs? How to represent obligations? In this paper we represent beliefs and obligations by rules, following the dominant tradition in deontic logic (see e.g. [20, 21]).
- How to develop a normative decision theory based on belief and obligation rules? In this paper we introduce a qualitative decision theory, based on belief (B) and obligation (O) rules.

- How can this decision theory be decomposed into goal generation and goal-based planning? How to define a notion of normative goals in this theory? In this paper, we show how these questions can be answered for our qualitative decision theory.

Our main aim in this paper is not to convince the reader that this decision theory is the best option available. It has the advantage that it is a simple theory, it is definitely not the most advanced one. Our aim is to show how, given a decision theory, a distinction can be made between goal generation and goal-based planning. The motivation of our study is to give formal foundations for goal-based normative agent architectures, such as the one depicted in Figure 2. This figure should be read as follows. The input of

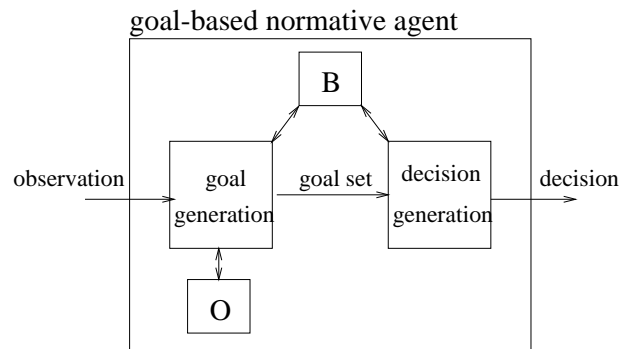


Fig. 2. Goal-based agent

the system is an observation and its output is a decision. There are two components, which we call goal generation and decision generation. Goal generation has a goal set as its output, which is the input for decision generation. Decision generation is for example the reasoner (or the planner) of the classic knowledge-based system depicted in Figure 1. Decision making is based on two sets of rules, represented by components *B* for belief rules and *O* for obligation rules. In particular, both goal generation and decision generation use belief rules, but only goal generation uses obligation rules. This represents that the motivational attitude encoded in *O* is transformed by goal generation in the goal set. In this paper, the difference between obligation rules and normative goal set is that obligation rules are pairs of propositional formulas, whereas a goal set is a set of propositional formulas, or (when we distinguish positive and negative goals) two sets of propositional sentences.

Like classical decision theory, but in contrast to several proposals in the BDI approach [6, 23], the theory does not incorporate temporal reasoning and scheduling.

The layout of this paper is as follows. We first develop a normative logic of decision. This logic tells us what the optimal decision is, but it does not tell us how to find this optimal decision. We then consider the AI solution to this problem [26]: break down the decision problem into goal generation and goal-based decisions.

2 A normative decision theory

The qualitative decision theory introduced in this section is based on sets of belief and obligation rules. There are several choices to be made, where our guide is to choose the simplest option available.

2.1 Decision specification

The starting point of any theory of decision is a distinction between choices made by the decision maker (flip a coin) and choices imposed on it by its environment (head or tail). We therefore assume the two disjoint sets of propositional atoms $A = \{a, b, c, \dots\}$ (the agent's decision variables [18] or controllable propositions [3]) and $W = \{p, q, r, \dots\}$ (the world parameters or uncontrollable propositions). We write:

- L_A, L_W and L_{AW} for the propositional languages built up from these atoms in the usual way, and x, y, \dots for any sentences of these languages.
- Cn_A, Cn_W and Cn_{AW} for the consequence sets, and \models_A, \models_W and \models_{AW} for satisfiability, in any of these propositional logics.
- $x \Rightarrow y$ for an ordered pair of propositional sentences called a rule.

A decision specification given in Definition 1 is a description of a decision problem. It contains a set of belief and obligation rules, as well as a set of facts and an initial decision (or prior intentions). A belief rule ‘the agent believes y in context x ’ is an ordered pair $x \Rightarrow y$ with $x \in L_{AW}$ and $y \in L_W$, and an obligation rule ‘the agent ought y in context x ’ is an ordered pair $x \Rightarrow y$ with $x \in L_{AW}$ and $y \in L_{AW}$. It implies that the agent's beliefs are about the world ($x \Rightarrow p$), and not about the agent's decisions. These beliefs can be about the effects of decisions made by the agent ($a \Rightarrow p$) as well as beliefs about the effects of parameters set by the world ($p \Rightarrow q$). Moreover, the agent's obligations can be about the world ($x \Rightarrow p$, obligation-to-be), but also about the agent's decisions ($x \Rightarrow a$, obligation-to-do). These obligations can be triggered by parameters set by the world ($p \Rightarrow y$) as well as by decisions made by the agent ($a \Rightarrow y$).

The reason that we do exclude decision variables in the consequent of the belief rules is that belief rules are assumed here to be not defeasible: a belief for decision a cannot be defeated by the decision $\neg a$. This condition can be relaxed in an extension of the theory which incorporates defeasible belief rules.

Definition 1 (Decision specification). *A decision specification is a tuple $DS = \langle F, B, O, d_0 \rangle$ that contains a consistent set of facts $F \subseteq L_W$, a finite set of belief rules $B \subseteq L_{AW} \times L_W$, a finite set of obligation rules $O \subseteq L_{AW} \times L_{AW}$ and an initial decision $d_0 \subseteq L_A$.*

2.2 Decisions

The belief rules are used to express the expected consequences of a decision, where a decision d is any subset of L_A that implies the initial decision d_0 , and the set of expected consequences of this decision d is the belief extension of $F \cup d$, as defined in Definition 2 below. Belief rules are interpreted as inference rules. We write $E_R(S)$ for the R extension of S .

Definition 2 (Extension). Let $R \subseteq L_{AW} \times L_{AW}$ be a set of rules and $S \subseteq L_{AW}$ a set of sentences. The consequents of the S -applicable rules are:

$$R(S) = \{y \mid x \Rightarrow y \in R, x \in S\}$$

and the R extension of S is the set of the consequents of the iteratively S -applicable rules:

$$E_R(S) = \bigcap_{S \subseteq X, R(Cn_{AW}(X)) \subseteq X} X$$

The following proposition shows that $E_R(S)$ is the smallest superset of S closed under the rules R interpreted as inference rules.

Proposition 1 (Iteration). Let

- $E_R^0(S) = S$
- $E_R^i(S) = E_R^{i-1}(S) \cup R(Cn_{AW}(E_R^{i-1}(S)))$ for $i > 0$

We have $E_R(S) = \bigcup_0^\infty E_R^i(S)$.

Proof. Follows from analogous results in input/output logic.

The following proposition shows that $E_R(S)$ is monotonic.

Proposition 2 (Monotonicity). We have $R(S) \subseteq R(S \cup T)$ and $E_R(S) \subseteq E_R(S \cup T)$.

Proof. By definition.

Monotonicity is illustrated by the following example.

Example 1. Let $R = \{\top \Rightarrow p, a \Rightarrow \neg p\}$ and $S = \{a\}$, where \top stands for any tautology like $p \vee \neg p$. We have $E_R(S) = \{a, p, \neg p\}$, i.e. the R extension of S is inconsistent. We do *not* have that for example the specific rule overrides the more general one such that $E_R(S) = \{a, \neg p\}$.

We assume that a decision is an arbitrary subset of controllable propositions that implies the initial decision and does not imply a contradiction in its belief consequences.

Definition 3 (Decisions). Let $DS = \langle F, B, O, d_0 \rangle$ be a decision specification. The set of DS decisions is

$$\Delta = \{d \mid d_0 \subseteq d \subseteq L_A, E_B(F \cup d) \text{ is consistent}\}$$

When a decision implies a , then we say that the agent makes decision a , or that it does a . The following example illustrates decisions.

Example 2. Let $A = \{a, b, c\}$, $W = \{p, q, r\}$ and $DS = \langle F, B, O, d_0 \rangle$ with $F = \{p \rightarrow r\}$, $B = \{p \Rightarrow q, b \Rightarrow \neg q, c \Rightarrow p\}$, $O = \{\top \Rightarrow r, \top \Rightarrow a, a \Rightarrow b\}$ and $d_0 = \{a\}$. The initial decision d_0 reflects that the agent has already decided in an earlier stage to reach the obligation $\top \Rightarrow a$. Note that the consequents of all B rules are sentences of L_W , whereas the antecedents of the B rules as well as the antecedents and consequents of the O rules are sentences of L_{AW} . We have due to the definition of $E_R(S)$:

$$E_B(F \cup \{a\}) = \{p \rightarrow r, a\}$$

$$E_B(F \cup \{a, b\}) = \{p \rightarrow r, a, b, \neg q\}$$

$$E_B(F \cup \{a, c\}) = \{p \rightarrow r, a, c, p, q\}$$

$$E_B(F \cup \{a, b, c\}) = \{p \rightarrow r, a, b, c, p, q, \neg q\}$$

Note that $\{a, b, c\}$ is not a DS decision, because its extension is inconsistent.

2.3 Optimal decisions

Given the specification of a decision problem, Definition 3 indicates all possible decisions that can be generated. In the following, we introduce a normative decision theory, which determines the interpretation of the elements of the decision specification. This normative decision theory imposes an ordering on possible decisions based on the obligation rules and provides a way to identify optimal decisions. In particular, the obligation rules are used to compare the decisions. There are various ways to compare decisions based on the obligation rules. For example, one can compare decisions by considering the obligation rules that are violated by them where an obligation rule $x \Rightarrow y$ is called to be violated by a decision if the belief consequences of the decision imply $x \wedge \neg y$. Another way to compare decisions is by considering the reached obligation rules where an obligation rule $x \Rightarrow y$ is called to be reached by a decision if the belief consequences of the decision imply $x \wedge y$. In this paper, we compare decisions by considering the unreached obligation rules. An obligation rule $x \Rightarrow y$ is unreached by a decision if the belief consequences of the decision imply x , but not y . Note that the set of unreached desires is a superset of the set of violated desires.

Definition 4 (Comparing decisions). Let $DS = \langle F, B, O, d_0 \rangle$ be a decision specification and d be a DS decision. The unreached obligations of decision d are:

$$U(d) = \{x \Rightarrow y \in O \mid E_B(F \cup d) \models x \text{ and } E_B(F \cup d) \not\models y\}$$

Decision d_1 is at least as good as decision d_2 , written as $d_1 \geq_U d_2$, iff

$$U(d_1) \subseteq U(d_2)$$

Decision d_1 dominates decision d_2 , written as $d_1 >_U d_2$, iff

$$d_1 \geq_U d_2 \text{ and } d_2 \not\geq_U d_1$$

Decision d_1 is as good as decision d_2 , written as $d_1 \sim_U d_2$, iff

$$d_1 \geq_U d_2 \text{ and } d_2 \geq_U d_1$$

The following continuation of Example 2 illustrates the comparison of decisions.

Example 3 (Continued). We have:

$$U(\{a\}) = \{\top \Rightarrow r, a \Rightarrow b\},$$

$$U(\{a, b\}) = \{\top \Rightarrow r\},$$

$$U(\{a, c\}) = \{a \Rightarrow b\}.$$

We thus have that the decisions $\{a, b\}$ and $\{a, c\}$ both dominate the initial decision $\{a\}$, i.e. $\{a, b\} >_U \{a\}$ and $\{a, c\} >_U \{a\}$, but the decisions $\{a, b\}$ and $\{a, c\}$ do not dominate each other nor are they as good as each other, i.e. $\{a, b\} \not\geq_U \{a, c\}$ and $\{a, c\} \not\geq_U \{a, b\}$.

The following proposition shows that the binary relation \geq_U on decisions is transitive and we can thus interpret it as a preference relation.

Proposition 3 (Transitivity). *The binary relation \geq_U is transitive.*

Proof. Follows from transitivity of subset-relation.

A consequence of this normative decision theory is that the ordering of decisions is influenced only by the subset of obligation rules which is disjoint with the set of belief rules. The following proposition shows that obligations only matter as long as they are different from beliefs.

Proposition 4 (Redundancy). *Let $DS = \langle F, B, O, d_0 \rangle$ and $DS' = \langle F, B, O \setminus B, d_0 \rangle$. Then, for every pair $\langle d_1, d_2 \rangle$ of decisions from DS there exists a pair $\langle d'_1, d'_2 \rangle$ from DS' such that $d_1 \geq_U d_2$ iff $d'_1 \geq_U d'_2$.*

Proof.

By Definition 3 DS and DS' have the same set of decisions. Let $x \Rightarrow y \in B$ and $x \Rightarrow y \in O$ in both DS and DS' . Then, Proposition 1 states that if $x \in E_B(d \cup F)$ then also $y \in E_B(d \cup F)$. Consequently, for DS and DS' the rule $x \Rightarrow y$ cannot be in $U(d)$ and thus this rule cannot change the ordering relation in DS or DS' .

The decision theory prescribes an economic rational decision maker to select the optimal or best decision, which is defined as a decision that is not dominated.

Definition 5 (Optimal decision). *Let DS be a decision specification. A DS decision d is U -optimal iff there is no DS decision d' that dominates it, i.e. $d' >_U d$.*

The following example illustrates optimal decisions.

Example 4. Let $A = \{a, b\}$, $W = \emptyset$ and $DS = \langle \emptyset, \emptyset, \{a \Rightarrow b\}, \emptyset \rangle$. We have:
 $U(d) = \{a \Rightarrow b\}$ if $d \models_{AW} a$ and $d \not\models_{AW} b$, $U(d) = \emptyset$ otherwise

The U -optimal decisions are the decisions d that either do not imply a or that imply $a \wedge b$.

The following proposition shows that for each decision specification, there is at least one optimal decision. This is important, because it guarantees that agents can always act in some way.

Proposition 5 (Existence). *Let DS be a decision specification. There is at least one U -optimal DS decision.*

Proof. Since the facts F are consistent, there exists at least one DS decision. Since the set of desire rules is finite there do not exist infinite ascending chains in \geq_U , and thus there is an U -optimal decision.

For a given decision specification, there may be more than one optimal decisions. Therefore, we introduce an alternative to our notion of optimality by adding minimality in the definition of optimal decisions. The following Definition 6 introduces a distinction between smaller and larger decisions. A smaller decision implies that the agent commits itself to less choices. A minimal optimal decision is an optimal decision such that there is no smaller optimal decision.

Definition 6 (Minimal optimal decision). A decision d is a minimal U -optimal DS decision iff it is an U -optimal DS decision and there is no U -optimal DS decision d' such that $d \models d'$ and $d' \not\models d$.

The following example illustrates the distinction between optimal and minimal optimal decisions.

Example 5. Let $DS = \langle F, B, O, d_0 \rangle$ with $F = \emptyset$, $B = \{a \Rightarrow x, b \Rightarrow x\}$, $O = \{\top \Rightarrow x\}$, $d_0 = \emptyset$. Optimal decisions are $\{a\}$, $\{b\}$ and $\{a, b\}$, of which only the former two are minimal.

The following proposition illustrates in what sense a decision theory based on optimal decisions and one based on minimal optimal decisions are different.

Proposition 6 (Minimality). There is an U -optimal DS decision d , such that there is no minimal U -optimal DS decision d' with $d \sim_U d'$.

Proof. Consider the decision specification in Example 4, $DS = \langle \emptyset, \emptyset, \{a \Rightarrow b\}, \emptyset \rangle$. The unique minimal U -optimal decision is $d_1 = \emptyset$. The decision $d_2 = \{a, b\}$ is also U -optimal, but we do not have $d \sim_U d'$.

The following example illustrates that the minimal decision d_0 is not necessarily an optimal decision.

Example 6. Let $A = \{a\}$, $W = \emptyset$ and $DS = \langle \emptyset, \emptyset, \{\top \Rightarrow a\}, \emptyset \rangle$. We have $U(\emptyset) = \{\top \Rightarrow a\}$ and $U(\{a\}) = \emptyset$. Hence, doing a is better than doing nothing.

The notions U -optimality and minimal U -optimality, which are in fact properties of decisions, can be used to characterize the type of decision making rational agents. We define two types of rational agents.

Definition 7. A *BO rational agent* is an agent that, confronted with a decision specification DS , selects an U -optimal DS decision. A *BO parsimonious agent* is a *BO rational agent* that selects a minimal U -optimal DS decision.

The logic of belief rules employed in this paper has been called simple-minded output (to be precise, it has been called out_3^+) in input/output logics [20]. The following example illustrates one of its drawbacks. In Savage's terminology [24], the agent does not obey the sure-thing principle.

Example 7. Let $A = \{a\}$, $W = \{p\}$ and $DS = \langle \emptyset, \emptyset, \{p \Rightarrow a, \neg p \Rightarrow a\}, \emptyset \rangle$. Any decision is an optimal decision. There is no preference for decision $\{a\}$. If p holds then a is obliged, and if p is false then a is obliged. However, the agent cannot infer that a is the optimal decision.

However, in this paper we no longer consider the particular properties of the logic of rules and the logic of decision we have proposed thus far, but we turn to the notion of goals. This concept is introduced in the following section.

3 Goal-based normative decision theory

In the previous section, we have explained possible decisions of BO agents, in the sense of Definition 7, and introduced U-optimal property of decisions. In this section, we show that every BO rational agent can be understood as a goal-based agent [22]. This is done by assuming the decisions of a BO agent to be the result of planning of some of its goals. These goals are in turn assumed to be generated by a goal generation mechanism. The question we like to answer is what are the properties of goals such that, when they are planned based on the belief rules, they result U-optimal decisions. In particular, we aim to define a characterization of goals such that the decisions that achieve those goals are U-optimal decisions and vice versa, i.e. U-optimal decisions are decisions that generate those goals. This demand is what we will call "representation theorem".

3.1 Goal-based optimal decisions

Goal-based decisions in Definition 8 combine decisions in Definition 3 and the notion of goal, which is a set of propositional sentences. Note that a goal set can contain decision variables (which we call to-do goals) as well as parameters (which we call to-be goals).

Definition 8 (Goal-based decision). *Let $DS = \langle F, B, O, d_0 \rangle$ be a decision specification and the goal set $G \subseteq L_{AW}$ a set of sentences. A decision d is a G decision iff $E_B(F \cup d) \models_{AW} G$.*

How to define a goal set for a decision specification? We are looking for goal sets G which have the desirable property that all G decisions are optimal. One way to start is to consider all derivable goals from an initial decision and a *maximal* set of obligations.

Definition 9 (Derivable goal set). *Let $DS = \langle F, B, O, d_0 \rangle$ be a decision specification. A set of formulas $G \subseteq L_{AW}$ is a derivable goal set of DS iff*

$$G = E_{B \cup O'}(F \cup d_0) \setminus Cn_{AW}(E_B(F \cup d_0))$$

where $O' \subseteq O$ is a maximal (with respect to set inclusion) set such that

1. $E_{B \cup O'}(F \cup d_0)$ is consistent and
2. there is a DS decision d that is a G decision.

However, the following proposition shows that for some derivable goal set G , not all G decisions are U-optimal.

Proposition 7 (U-optimal G decision). *For a derivable goal set G of DS , a G decision does not have to be an U-optimal decision.*

Proof. Reconsider the decision specification in Example 4, $DS = \langle \emptyset, \emptyset, \{a \Rightarrow b\}, \emptyset \rangle$. The derivable goal set is $G = \emptyset$. The decision $d = \{a\}$ is a G decision, but it is not U-optimal.

The following proposition shows that the former proposition also holds if we restrict ourselves to minimal optimal decisions.

Proposition 8 (Minimal G decision). *For a derivable goal set G of DS , a minimal G decision does not have to be an U -optimal decision.*

Proof. Consider the decision specification $DS = \langle \emptyset, \{a \Rightarrow p\}, \{\top \Rightarrow p, a \Rightarrow b\}, \emptyset \rangle$. The set $G = \{p\}$ is the only derivable goal set (based on $O' = \{\top \Rightarrow p, a \Rightarrow b\}$). The DS decisions $d_1 = \{a\}$ is a minimal G decision, but only $d_2 = \{a, b\}$ is an U -optimal decision. d_2 is also a G decision.

Finally, the following proposition shows that there are also derivable goal sets G such that there exist no G decision at all.

Proposition 9 (Existence). *For a derivable goal set G of DS , G decisions do not have to exist.*

Proof. Consider the decision specification $DS = \langle \emptyset, \emptyset, \{\top \Rightarrow p\}, \emptyset \rangle$. The set $G = \{p\}$ is the only derivable goal set (based on $O' = \{\top \Rightarrow p\}$). However, the only DS decisions is $d = \emptyset$ and G is not a d decision.

Given this variety of problems, we do not try to repair the notion of derivable goal set. Instead, we define goals with respect to an optimal decision.

Definition 10 (Achievable goal set). *Let $DS = \langle F, B, O, d_0 \rangle$ be a decision specification. A set of formulas $G \subseteq L_{AW}$ is an achievable goal set of DS iff there is an U -optimal DS decision d such that*

$$G = \{x \wedge y \mid x \Rightarrow y \in O', E_{B \cup O'}(F \cup d) \models_{AW} x \wedge y\}$$

where

$$O' = \{x \Rightarrow y \in O \mid E_B(F \cup d) \not\models_{AW} x \text{ or } E_B(F \cup d) \models_{AW} x \wedge y\}$$

Note that the obligation rules in O' are those rules from O that are satisfied (i.e. reached) by the U -optimal decision d . The additional application of these rules in $E_{B \cup O'}(F \cup d)$ will not effect the consequences of U -optimal decision d . This results in the following simpler proposition.

Proposition 10 (Achievable goal set). *Let $DS = \langle F, B, O, d_0 \rangle$ be a decision specification. A set of formulas $G \subseteq L_{AW}$ is an achievable goal set of DS iff there is an U -optimal DS decision d such that*

$$G = \{x \wedge y \mid x \Rightarrow y \in O, E_B(F \cup d) \models_{AW} x \wedge y\}$$

Proof. Follows directly from $E_{B \cup O'}(F \cup d) = E_B(F \cup d)$, where O' is as defined in Definition 10.

The following two properties show that the notion of achievable goal set is not an enough characterization of goals such that the representation theorem cannot be proven. In particular, the following proposition shows that we can define one half of the representation theorem for achievable goal sets.

Proposition 11. *For an U -optimal decision d of DS there is an achievable goal set G of DS such that d is a G decision.*

Proof. Follows directly from $E_{B \cup O'}(F \cup d) = E_B(F \cup d)$.

However, the following proposition shows that the other half of the representation theorem still fails.

Proposition 12. *For an achievable goal set G of DS , a G decision does not have to be an U -optimal decision.*

Proof. Consider the decision specification $DS = \langle \{-q\}, \{a \Rightarrow p, b \Rightarrow p\}, \{\top \Rightarrow p, b \Rightarrow q\}, \emptyset \rangle$. The set $G = \{p\}$ is the only achievable goal set (based on $O' = \{\top \Rightarrow p, b \Rightarrow q\}$). The DS decisions $d_1 = \{a\}$ and $d_2 = \{b\}$ are both (minimal) G decisions, but only d_1 is an optimal decision.

The counterexample in Proposition 12 also shows that we cannot prove the second half of the representation theorem, because we only consider positive goals (states the agent wants to reach) and not negative goals (states the agents wants to evade). The theory is extended with positive and negative goals in the following subsection.

3.2 Positive and negative goals

In this section we show that the representation theorem works both ways if we add negative goals, which are defined in the following definition as states the agent has to avoid. They function as constraints on the search process of goal-based decisions.

Definition 11 (Goal-based decision). *Let $DS = \langle F, B, O, d_0 \rangle$ be a decision specification, and the so-called positive goal set G^+ and negative goal set G^- subsets of L_{AW} . A decision d is a $\langle G^+, G^- \rangle$ decision iff $E_B(F \cup d) \models_{AW} G^+$ and for each $g \in G^-$ we have $E_B(F \cup d) \not\models_{AW} g$.*

Based on this definition of goal decision, we can extend the definition of achievable goal set with negative goals.

Definition 12 (Positive and negative achievable goal set). *Let $DS = \langle F, B, O, d_0 \rangle$ be a decision specification. The two sets of formulas $G^+, G^- \subseteq L_{AW}$ are respectively a positive and negative achievable goal sets of DS iff there is an optimal DS decision d such that*

$$G^+ = \{x \wedge y \mid x \Rightarrow y \in O', E_B(F \cup d) \models_{AW} x \wedge y\}$$

$$G^- = \{x \mid x \Rightarrow y \in O', E_{B \cup O'}(F \cup d) \not\models_{AW} x\}$$

where

$$O' = \{x \Rightarrow y \in O \mid E_B(F \cup d) \not\models_{AW} x \text{ or } E_B(F \cup d) \models_{AW} x \wedge y\}$$

For $\langle G^+, G^- \rangle$ decisions, we consider minimal optimal decisions. The following example illustrates the distinction between optimal $\langle G^+, G^- \rangle$ decisions and minimal optimal $\langle G^+, G^- \rangle$ decisions.

Example 8. Let $A = \{a, b\}$, $W = \emptyset$ and $DS = \langle \emptyset, \emptyset, \{a \Rightarrow b\}, \emptyset \rangle$. The optimal decision is \emptyset or $\{a, b\}$, and the related goal sets are $\langle G^+, G^- \rangle = \langle \emptyset, \{a\} \rangle$ and $\langle G^+, G^- \rangle = \langle \{a \wedge b\}, \emptyset \rangle$. The only minimal optimal decision is the former.

The following example illustrates a conflict.

Example 9. Let $W = \{p\}$, $A = \{a\}$, $DS = \langle F, B, O, d_0 \rangle$ with $F = \emptyset$, $B = \emptyset$, $O = \{\top \Rightarrow a \wedge p, \top \Rightarrow \neg a\}$, $d_0 = \emptyset$. We have optimal decision $\{\neg a\}$ with goal set $\langle G^+, G^- \rangle = \langle \{\neg a\}, \emptyset \rangle$. The decision $\{a\}$ does not derive goal set $\langle G^+, G^- \rangle = \langle \{a \wedge p\}, \emptyset \rangle$. One of the possible choices is $\{a\}$, which is however sub-optimal since we cannot guarantee that the first obligation is reached.

The following two propositions show that $\langle G^+, G^- \rangle$ goal set is the right characterization of goals such that the representation theorem can be proven. The first part of the representation theorem is analogous to Proposition 11.

Proposition 13. *For an U-optimal decision d of DS there is an achievable goal set $\langle G^+, G^- \rangle$ of DS such that d is a $\langle G^+, G^- \rangle$ decision.*

Proof. See Proposition 11.

In contrast to achievable goal set G , the second part of the representation theorem can be proven for $\langle G^+, G^- \rangle$ goal set.

Proposition 14. *For an achievable goal set $\langle G^+, G^- \rangle$ of DS , a $\langle G^+, G^- \rangle$ decision is an U-optimal decision.*

Proof. $\langle G^+, G^- \rangle$ is achievable and thus there is an U-optimal DS decision such that $E_B(F \cup d) \models_{AW} G^+$ and for all $g \in G^-$ we have $E_B(F \cup d) \not\models_{AW} g$. Let d be any decision d such that $E_B(F \cup d) \models_{AW} G^+$ and for all $g \in G^-$ we have $E_B(F \cup d) \not\models_{AW} g$. Suppose d is not U-optimal. This means that there exists a d' such that $d' >_U d$, i.e. such that there exists an obligation $x \Rightarrow y \in O$ with $E_B(F \cup d) \models_{AW} x$, $E_B(F \cup d) \not\models_{AW} y$ and either:

- $E_B(F \cup d') \not\models_{AW} x \wedge y$;
- $E_B(F \cup d') \models_{AW} y$;

However, the first option is not possible due to the negative goals and the second option is not possible due to the positive goals. Contradiction, so d has to be U-optimal.

The representation theorem is a combination of Proposition 13 and 14.

Theorem 1. *A decision d is an U-optimal decision if and only if there is an achievable goal set $\langle G^+, G^- \rangle$ of DS such that d is a $\langle G^+, G^- \rangle$ decision.*

The following example illustrates uncertainty about the world.

Example 10. Let $DS = \langle F, B, O, d_0 \rangle$ with $B = \{a \Rightarrow q\}$ and $O = \{\top \Rightarrow p, p \Rightarrow q\}$. We have two optimal decisions, $d_1 = \emptyset$ and $d_2 = \{a\}$, with corresponding achievable goal sets $\langle G^+, G^- \rangle = \langle \emptyset, \{p\} \rangle$ and goal $\langle G^+, G^- \rangle = \langle \{p, q\}, \emptyset \rangle$ $G = \{p, q\}$. We may select $\{a\}$ whereas we do not know whether p will be the case. If we are pessimistic, we assume p will be false. There is no reason to do $\{a\}$.

The following example illustrates side effects from actions.

Example 11. Let $DS = \langle F, B, O, d_0 \rangle$ with $B = \{a \Rightarrow p, a \Rightarrow q\}$ and $O = \{\top \Rightarrow p, \top \Rightarrow \neg q\}$. We have two optimal decisions, $D_1 = \emptyset$ and $d_2 = \{a\}$, with corresponding achievable goal sets $\langle G^+, G^- \rangle = \langle \emptyset, \{p, \neg q\} \rangle$ and goal $\langle G^+, G^- \rangle = \langle \{p, q\}, \emptyset \rangle$ $G = \{p, q\}$. a implies a desired proposition, but it also violates another desire.

The following example illustrates a zig zag.

Example 12. Let $DS = \langle F, B, O, d_0 \rangle$ with $B = \{a_i \Rightarrow p_i \mid i = 0, 1, 2, 3, \dots\}$ and $O = \{\top \Rightarrow p_0\} \cup \{a_i \Rightarrow p_{i+1} \mid i = 0, 1, 2, 3, \dots\}$.

$$\begin{aligned} G_0^+ &= \{p_0\} \\ d_0 &= \{p_0, a_0\} \\ G_1^+ &= \{p_0, a_0, p_1\} \\ d_1 &= \{p_0, a_0, p_1, a_1\} \\ &\dots \end{aligned}$$

... We can continue to construct new goals and new decisions due to side effects of actions.

4 Agent specification and design

In this section we discuss how the proposed qualitative normative decision and goal theory can be used to guide the design and specification of rational BO agents in a compositional way. The general idea of compositional design and specification is to build agents using components. They may be either primitive components or composed of other components, such that the specification of agents can be broken down into the specification of components and their relations. Here we give some preliminary ideas and explain how the proposed qualitative normative decision and goal theory supports a specific compositional design for rational BO agent.

The qualitative decision theory, as proposed in section 2, specifies the decision making of an agent in terms of its observations and its mental attitudes such as beliefs and obligations. The specified agent can therefore be considered as consisting of components that represent agent's beliefs and obligations and a reasoning component that generates agent's decisions based on its observations and mental attitudes. The abstract design of such a BO agent is illustrated in Figure 1 and copied in Figure 3. For this design of BO agents, notions such as optimal decisions and minimal optimal decisions can be used to specify the reasoning component and thus the decision making mechanism of the agent.

The following example illustrates an agent with certain beliefs and obligations, the possible decisions that the agent can make, and how the notions from qualitative normative decision theory can be used to specify the subset of decisions that the agent can make.

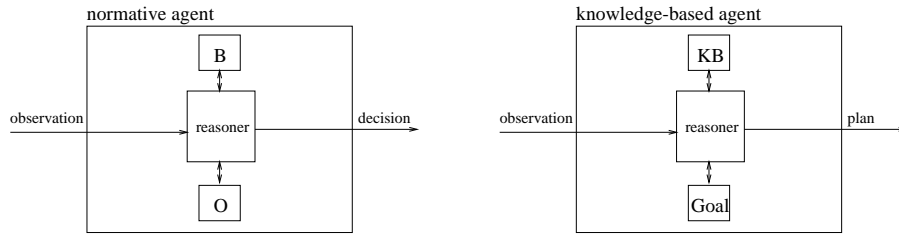


Fig. 3. Agent

Example 13. Consider an agent who believes that he works and that if he sets an alarm clock he can wake up early to arrive in time at his work, i.e.

$$B = \{\top \Rightarrow \text{Work}, \text{SetAlarm} \Rightarrow \text{InTime}\}$$

The agent has also the obligation to arrive early at his work and he has to inform his boss when he does not work, i.e.

$$O = \{\text{Work} \Rightarrow \text{InTime}, \neg\text{Work} \Rightarrow \text{InformBoss}\}$$

In this example, the propositions SetAlarm and InformBoss are assumed to be decision variables (the agent has control on setting the alarm clock and informing his boss), while Work and InTime are assumed to be world parameters (the agent has no control on its working status and the starting time). Moreover, we assume that the agent has no observation and no intentions. One can specify the agent as a rational BO agent in the sense that it makes optimal decisions. Being specified as a rational BO agent, he will decide to use the alarm clock though he has in principle many possible decisions including \emptyset , {SetAlarm}, {InformBoss}, and {SetAlarm, InformBoss}.

The goal-based decision theory, as proposed in section 3, explains the decision making of a rational BO agent as if it aims at maximizing achieved normative goals. In particular, the goal-based decision theory explains how normative goals of an agent can be specified based on its decision specification. The specified reasoning component of the rational BO agent can therefore be decomposed and designed as consisting of two reasoning components: one which generates normative goals and one which generate decisions to achieve those goals. This decomposition suggests an agent design as illustrated in Figure 2 and copied in Figure 4. According to this agent design, a BO agent generates first its normative goals based on its observation, its beliefs, obligations and its intentions. The generated goals are subsequently the input of the decision generation component.

Following the design decomposition, the specification of a BO agent can now also be decomposed and defined in terms of the specification of its goal and decision generation mechanisms. In particular, the goal generation mechanism can be specified in terms of agent's observations and its mental state on the one hand and its goals on the other hand. The decision generation component can then be specified in terms of agent's goals and mental state on the one hand and its decisions on the other hand.

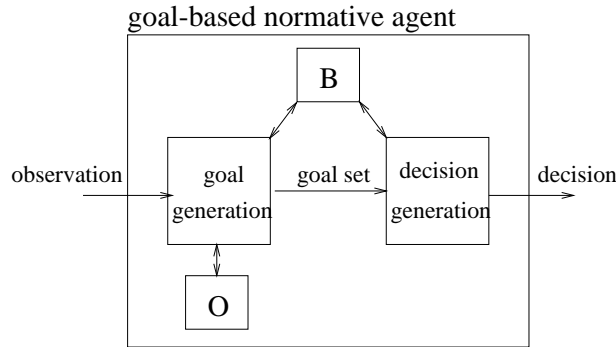


Fig. 4. Goal-based agent

For example, consider again the working agent that may have in principle many goal sets consisting of \emptyset , Work, Intime, SetAlarm, and InformBoss. This implies that the goal generation component may generate one of these possible goal sets. Using the notions from goal-based decision theory one may specify the goal generation mechanism in order to generate achievable goal sets which when planned by the decision generation component will result optimal decisions.

In summary, we believe that the qualitative normative decision theory and goal-based decision theory can be used to provide compositional specification and design of rational BO agents. This leads to a transparent agent specification and design structure. Moreover, it leads to support for reuse and maintainability of components and generic models. The compositional specification and design of agents enable us to specify and design agents at various levels of abstraction leaving out many details such as representation issues and reasoning schemes. For our rational BO agents we did not to explain how decisions are generated; we only specified what decisions should be generated. At one lower level we decomposed the reasoning mechanism and specified goal and decision generation mechanisms. We also did not discuss the representation of individual components such as the belief or the obligation components. The conditional rules in these components specify the input/output relation.

5 Further research

A distinction has been made between goal generating norms and action filtering norms (Castelfranchi and Conte, personal communication). It is an open problem whether these two kinds of norms can be formalized by our decision theory. It seems that obligation rules are only used to generate goals, and we therefore need another type of norms which filters actions. However, ought-to-do norms (i.e. obligations with a decision variable in the head) seem to act as or that we

In this paper we have restricted our discussion to beliefs and obligations, and the question can be raised how the decision theory can be extended with desires and intention, i.e. to the full BOID.

Moreover, we have restricted our analysis to a single autonomous agent. However, norms become useful in particular when several agents are considered in a multi agent system. We have made some preliminary observations based on a qualitative game theory in [11, 9].

We have not specified any details of the obligation rules. For example, whether there are any sanctions related to the norms [2].

6 Related research

We draw inspiration from Savage's classical decision theory [24]. The popularity of this theory is due to the fact that Savage shows that a rational decision maker, which satisfies some innocent looking properties, acts *as if* it is maximizing its expected utility function. This is called a representation theorem. In other words, Savage does not assume that an agent has a utility function and probability distribution which the agent uses to make decisions. However, he shows that if an agent bases his decisions on preferences and some properties of these preferences, then we can assume that the agent bases his decisions on these utilities and probabilities together with the decision rule which maximizes its expected utility. The main advantage is that Savage does not have to explain what a utility function *is*, an ontological problem which had haunted decision theory for ages.

Likewise, we want to develop a qualitative normative decision theory in which a normative agent acts *as if* it is trying to maximize achieved normative goals. This is what we call a goal-based representation theorem. It implies that agents can be formalized or verified as goal-based reasoners even when the agent does not reason with goals at all. In other words, goal-based representations do not have to be descriptive. A consequence of this indirect definition of goals is that the theory tells us what a goal *is*, such that we do not have to explain its ontological status separately. We call an agent which minimizes its unreachd obligations a BO rational agent, and we define goals as a set of formulas which can be derived by beliefs and obligations in a certain way. Our central result thus says that *BO rational agents act as if they maximize the set of goals that will be achieved*.

The theories in Thomason's BDP [27] and Broersen et al.'s BOID [4] are different, because they allow multiple belief sets. This introduces the new problem of blocking wishful thinking discussed extensively in [5].

Conte and Dignum [7] argue that, if you are speaking of normative agents as systems that somehow 'process' norms and decide upon them, then they must first form believe about those norms, whether they then adopt the norms or not. We believe that this is not incompatible with the approach advocated in this paper. However, in our general theory we do not want to commit ourselves to this particular view on norms. Our theory can also be applied, for example, to the game-theoretic notion of norms as advocated by for example [25].

In earlier work such as [28] we used the set of violated and reached obligations to order states, in the sense that we minimized violations and maximized reached obligations. The present definition has the advantage that it is simpler because it is based on a single minimization process only. Note that in the present circumstances we cannot

minimize violations only, because it would lead to the counterintuitive situation that the minimal decision $d = d_0$ is always optimal.

7 Concluding remarks

In this paper we have given an interpretation for goals in a qualitative decision theory based on beliefs and obligation rules, and we have shown that any agent which makes optimal decisions acts as if it is maximizing its achieved goals.

Our motivation comes from the analysis of goal-based architectures, which have recently been introduced. However, the results of this paper may be relevant for a much wider audience. For example, Dennett argues that automated systems can be analyzed using concepts from folk psychology like beliefs, obligations, and goals. Our work may be used in the formal foundations of this ‘intentional stance’ [12].

There are several topics for further research. The most interesting question is whether belief and obligation rules are fundamental, or whether they in turn can be represented by some other construct. Other topics for further research are a generalization of our representation theorem to other choices in our theory, the development of an incremental approach to goals, and the development of computationally attractive fragments of the logic, and heuristics of the optimization problem.

Acknowledgements

Thanks to Guido Boella, Zhisheng Huang and Joris Hulstijn for discussions on the issues raised in this paper.

References

1. J.R. Anderson, M. Matessa, and C. Lebiere. ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, 12(4):439–462, 1997.
2. G. Boella and L. Lesmo. A game theoretic approach to norms. *Cognitive Science Quarterly*, To appear.
3. C. Boutilier. Toward a logic for qualitative decision theory. In *Proceedings of the KR’94*, pages 75–86, 1994.
4. J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre. Goal generation in the BOID architecture. *Cognitive Science Quarterly*, to appear.
5. J. Broersen, M. Dastani, and L. van der Torre. Realistic desires. *Journal of Applied Non-Classical Logics*, to appear.
6. P.R. Cohen and H.J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
7. R. Conte, C. Castelfranchi, and F. Dignum. Autonomous norm-acceptance. In J. Muller, M. Singh, and A. Rao, editors, *Intelligent Agents V (ATAL98)*, volume 1555 of *LNAI*, pages 319–333. Springer, 1999.
8. M. Dastani, Z. Huang, and L. van der Torre. Dynamic desires. In S. Parsons, P. Gmytrasiewicz, and M. Wooldridge, editors, *Game Theory and Decision Theory in Agent-Based Computing*. Kluwer, 2001. Preliminary version appeared in Proceedings of the ICMAS2000 Workshop on game-theoretic and decision-theoretic approaches to agency (GTDT’2000), Boston, 2000.

9. M. Dastani and L. van der Torre. Decisions and games for bd agents. In *Proceedings GTDT'02*, 2002.
10. M. Dastani and L. van der Torre. Specifying the merging of desires into goals in the context of beliefs. In *Proceedings of The First Eurasian Conference on Advances in Information and Communication Technology (EurAsia ICT 2002)*, LNCS. Springer, 2002.
11. M. Dastani and L. van der Torre. What is a joint goal? games with beliefs and defeasible desires. In *Proceedings of NMR02*, 2002.
12. D. Dennett. *The intentional stance*. MIT Press, Cambridge, MA, 1987.
13. F. Dignum, D. Morley, E. Sonenberg, and L. Cavedon. Towards socially sophisticated bdi agents. In *Proceedings of the fourth International Conference on MultiAgent Systems (ICMAS-2000)*, pages 111–118, Boston, 2000. IEEE Computer Society.
14. J. Doyle. A model for deliberation, action and introspection. Technical Report AI-TR-581, MIT AI Laboratory, 1980.
15. M. Georgeff and A. Lansky. Reactive reasoning and planning. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 677–682, 1987.
16. S. Kraus, K. Sycara, and A. Evenchik. Reaching agreements through argumentation: a logical model and implementation. *Artificial Intelligence*, 104:1–69, 1998.
17. J.E. Laird, A. Newell, and P.S. Rosenbloom. SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33(1):1–64, 1987.
18. J. Lang. Conditional desires and utilities - an alternative approach to qualitative decision theory. In *In Proceedings of the European Conference on Artificial Intelligence (ECAI'96)*, pages 318–322, 1996.
19. J. Lang, L. van der Torre, and E. Weydert. Utilitarian desires. *Autonomous Agents and Multi-Agent Systems*, 5:3:329–363, 2002.
20. D. Makinson and L. van der Torre. Input-output logics. *Journal of Philosophical Logic*, 29:383–408, 2000.
21. D. Makinson and L. van der Torre. Constraints for input-output logics. *Journal of Philosophical Logic*, 30:155–185, 2001.
22. A. Newell. The knowledge level. *Artificial Intelligence*, 1982.
23. A. S. Rao and M. P. Georgeff. Decision procedures for bdi logics. *Journal of Logic and Computation*, 8:293–342, 1998.
24. L. Savage. *The foundations of statistics*. 1954.
25. Y. Shoham and M. Tennenholtz. On the emergence of social conventions: Modeling, analysis, and simulations. *Artificial Intelligence*, 94:139–166, 1997.
26. H. A. Simon. *The Sciences of the Artificial*. MIT Press, Cambridge, MA, second edition, 1981.
27. R. Thomason. Desires and defaults: A framework for planning with inferred goals. In *Proceedings KR 2000*, pages 702–713, 2000.
28. L. van der Torre and Y. Tan. Diagnosis and decision making in normative reasoning. *Artificial Intelligence and Law*, 7(1):51–67, 1999.
29. G.H. von Wright. *Practical Reason: Philosophical papers, volume 1*. Basil Blackwell, Oxford, 1983.