Chapter 3

Decisions, Deliberation, and Agent Types
CDT – QDT – BDI – 3APL – BOID

Mehdi Dastani
Utrecht University
mehdi@cs.uu.nl

Leendert van der Torre
CWI
torre@cwi.nl

Abstract

In this paper we investigate the relation between decisions, deliberation and agent types. In particular, we are interested how deliberation leads to decisions, and how agent types classify patterns of deliberation. We therefore consider Classical and Qualitative Decision Theories (CDT and QDT), the Beliefs-Desire-Intention (BDI) model, 3APL systems, and Belief-Obligation-Intention-Desire (BOID) systems. The first two are based on a decision rule which expresses a notion of rationality, whereas the latter three are based on deliberation processes and agent types.

1 Introduction

In recent years the interest in models of decision making for autonomous agents has increased. A reason for this growing interest is the notion of agent autonomy which is often defined as the ability of agents to decide which actions to perform at any moment in time. Different proposals have been made, rooted in different research traditions and objectives. Each proposal explains the decision making behavior of agents in terms of different sets of underlying concepts. For example, classical decision theory [22] explains the decision making behavior in terms of a probability and a utility function, and a decision rule. Other approaches criticize the representation of classical decision theory as being non-practical and unrealistic. According to them it is hard to translate all factors that influence the decision making behavior of an agent in terms of two functions that assign numbers to actions and states. In contrast, they aim at explaining the decision making behavior of agents in terms of qualitative concepts such as preference and likelihood ordering, or cognitive concepts such as beliefs, desires, intentions, and obligations. These concepts are claimed to be intuitive and easily accessible. In these approaches, the problem of deciding an action to perform is split into two problems. The first problem is to decide which goal should be achieved, and the second problem is to decide which action to perform to achieve the selected goal. The first problem is called goal selection and the second problem is called planning. The problem with these approaches is the lack of a clear formulation of decision rules that combines the underlying qualitative concepts to decide which action to perform at any moment in time.

There are many conceptualizations and formalizations of decision making. In [8] we compare classical decision theory with qualitative decision theory, knowledge-based systems

and belief-desire-intention models developed in artificial intelligence and agent theory. They all contain representations of information and motivation. Examples of informational attitudes are probability distributions, qualitative abstractions of probabilities, knowledge, and beliefs. Examples of motivational attitudes are utility functions, qualitative abstractions of utilities, goals, and desires. Each of them encodes a set of alternatives to be chosen from. This ranges from a small predetermined set, a set of decision variables, through logical formulas, to branches of a tree representing events through time. Moreover, they have a way of formulating how a decision is made. Classical and qualitative decision theory focus on the optimal decisions represented by a decision rule. Knowledge-based systems and belief-desire-intention models focus on a model of the representations used in decision making, inspired by cognitive notions like belief, desire, goal and intention. Relations among these concepts express an agent type, which constrains the deliberation process. We also consider the relation between decision processes and intentions, and the relation between game theory and norms and commitments.

In this paper we are interested in the relation between decisions, deliberation and agent types. In particular, we are interested in the question:

> How does deliberation based on agent types lead to decisions?

This question breaks down in three sub-questions:

1. What is an agent decision? To answer this question we discuss different approaches and concepts used to explain the decision making behavior of agents. The first is classical decision theory [15, 22] and the second it qualitative extension (QDT) [1, 17].

2. What is agent deliberation and how does it lead to a decision? To answer this question we discuss two other approaches. The third approach we discuss is based on an abstract model of the mental attitudes of an agent: beliefs, desires and intentions (BDI) [2, 5, 19]. The fourth (3APL) is similar to BDI except that the decision rules is replaced by a process called the deliberation process. It is this process that determine which actions should be performed for a given set of underlying cognitive concepts.

3. How do agent types classify patterns of deliberation? The fifth approach we discuss to answer this question is also based on mental attitudes extended with obligations (BOID) [3].

The layout of this paper is as follows. First we discuss decisions in classical and qualitative decision theory. Then, we discuss the BDI approach, 3APL systems and deliberation process. In the last section we consider the role of agent types in Rao and Georgeff's BDI approach and the BOID approach.

2 Decisions

2.1 Classical Decision Theory

In classical decision theory, a decision is a choice made by some entity of an action from a set of alternatives. It has nothing to say about actions – either about their nature or about how a set of them becomes available to the decision maker. A decision is good if it is an alternative that the decision maker believes will prove at least as good as other alternative actions. Good decisions are formally characterized as actions that maximize expected utility, a notion involving both belief and goodness. See [12] or [15, 22] for further explanations.

Definition 1 *Let A stand for the set of actions or alternatives. With each action, a set of outcomes is associated. Let W stand for the set of all possible worlds or outcomes. Let U*

be a measure of outcome value that assigns a utility $U(w)$ to each outcome $w \in W$, and let P be a measure of the probability of outcomes conditional on actions, with $P(w|a)$ denoting the probability that outcome w comes about after taking action $a \in A$ in the situation under consideration.

The expected utility $EU(a)$ of an action $a \in A$ is the average utility of the outcomes associated with the alternative a, weighing the utility of each outcome by the probability that the outcome results from the alternative a, that is, $EU(a) = \sum_{w \in W} U(w)P(w|a)$. A rational decision maker always maximizes expected utility, i.e. he makes decisions according to the MEU decision rule.

Decision theory is an active research area within economics, and the number of extensions and subtleties is too large to address here. Several other decision rules have been investigated, including qualitative ones, such as maximin, minimax, minregret etc. Finally, classical decision theory has been extended in several ways to deal for example with multiple objective, sequential decisions, multiple agents, distinct notions of risk, etcetera.

Decision theory has become one of the main foundations of economic theory due to so-called representation theorems. Representation theorems, such as the most famous one of Savage [22], typically prove that each decision maker obeying certain innocent looking postulates (about weighted choices) acts *as if* he applies the MEU decision rule with some probability distribution and utility function. Thus, he does not have to be aware of it, and his utility function does not have to represent selfishness. In fact, exactly the same is true for altruistic decision makers. They also act as if they maximize expected utility; they just have another utility function.

2.2 Qualitative Decision Theory

Qualitative decision theory relaxes the assumption of classical decision theory that the decision making agent is able to weigh all possible alternative courses of action before choosing one of them. In realistic situations decision making agents are resource bounded in the sense that they have partial information and do not have resources to compare the utility of possible states. The research on qualitative decision theory aims therefore to develop representation and reasoning schemes for partial information and generic preferences to represent probabilities of states and generic preferences over those states [12]. Typically qualitative orderings are introduced that represent the likelihood (probability) and desirability (utility) of states. In contrast to classical decision theory where a decision rule such as maximum expected utility determines the course of actions, in qualitative decision theory and in the presence of incomplete information, the course of actions are decided based on various strategies such as maximize potential gain or minimize potential loss [1].

According to Doyle and Thomason [12, p.58], quantitative representations of probability and utility and procedures for computing with these representations, do provide an adequate framework for manual treatment of very simple decision problems, but they are less successful in more realistic cases. For example, they argue that classical decision theory does not address decision making in unforeseen circumstances, it offers no means for capturing generic preferences, it provides little help in modeling decision makers who exhibit discomfort with numeric trade-offs, and it provides little help in effectively representing and reasoning about decisions involving broad knowledge of the world.

Doyle and Thomason argue for various formalization tasks. They distinguish the following new tasks: formalisms to express generic probabilities and preferences, properties of decision formulations, reasons and explanations, revision of preferences, practical qualitative decision-making procedures and agent modeling. Moreover, they argue that hybrid representation and reasoning with quantitative and qualitative techniques, as well as reasoning

within context, deserve special attention. Many of these issues are related to subjects studied in artificial intelligence. It appears that researchers now realize the need to reconnect the methods of AI with the qualitative foundations and quantitative methods of economics.

Some first results have been obtained in the area of reasoning about uncertainty, a subdomain of artificial intelligence which mainly attracts researchers with a background in reasoning about defaults and beliefs. Often the formalisms of reasoning about uncertainty are re-applied in the area of decision making. Thus, typically uncertainty is not represented by a probability function, but by a plausibility function, a possibilistic function, Spohn-type rankings, etc. Another consequence of this historic development is that the area is much more mathematically oriented than the planning community or the BDI community.

A typical example is Pearl's qualitative decision theory, in which decisions are based on abstractions of probability functions, and which is extended with a model of causality. Boutilier's logic of qualitative decision theory formalizes decisions as a set of controllable propositions. However, he introduces another concept in his logic, which makes the logic closer to the kind of logics we consider in the following section on deliberation and agent types. This new concept is the concept of goal. This is explained in the following section.

2.3 Introduction of Goals in Decision Theory

Goals serve a dual role in most planning systems, capturing aspects of both desires toward states and commitment to pursuing that state [10]. In goal-based planning, adopting a proposition as a goal commits the agent to find some way to accomplish the goal, even if this requires adopting some sub-goals that may not correspond to desirable propositions themselves [9]. In realistic planning situations goals can be achieved to varying degrees, and frequently goals cannot be achieved completely. Context-sensitive goals are formalized with basic concepts from decision theory [1, 9, 13]. In general, goal-based planning must be extended with a mechanism to choose between which goals must be adopted.

Boutilier [1] proposes a logic and possible worlds semantics for representing and reasoning with qualitative probabilities and utilities, and suggests several strategies for qualitative decision making based on this logic. His semantics is not quantitative (like CDT), but purely qualitative. Consequently, the maximum expected utility (MEU) decision rule is replaced by qualitative rules like Wald's criterion. The conditional preference is captured by a preference ordering (an ordinal value function) that is defined on possible worlds. The preference ordering represents the desirability of worlds. Similarly, probabilities are captured by an ordering, called normality ordering, on possible worlds representing their likelihood.

Definition 2 *The possible worlds semantics for this logic is based on models of the form* $M = \langle W, \leq_P, \leq_N, V \rangle$ *where* W *is a set of worlds (outcomes),* $\leq_P$ *is a transitive and connected preference ordering relation on* W*,* $\leq_N$ *is a transitive and connected normality ordering relation on* W*, and* V *is a valuation function.*

In Boutilier's logic, conditional preferences can be represented by means of ideal operator I. We have that a model M satisfies the formula $I(\varphi|\psi)$ if the preferred or best or minimal ψ worlds are φ worlds. For example, let u be the proposition 'agent has umbrella' and r be the proposition 'it rains', then $I(u|r)$ expresses that in the most preferred rain-worlds the agent has an umbrella. Similarly, probabilities are represented in this logic by means of a normative conditional connective $\Rightarrow$. For example, let w be the proposition 'the agent is wet' and r be the proposition 'it rains', then $r \Rightarrow w$ expresses that the agent is wet at the most normal rain-worlds. Its semantics is derived from Hansson's deontic logic with

modal operator O for obligation [14].[1] A similar type of semantics is used by Lang [16] for a modality D to model desire. An alternative approach represents conditional modalities by so called 'ceteris paribus' preferences, using additional formal machinery to formalize the notion of 'similar circumstances', see e.g. [11, 13, 23, 24].

In general, a goal is any proposition that the agent attempts to make true. A rational agent is assumed to attempt to reach the most preferred worlds consistent with its default knowledge. Given the ideal operator and the default conditional, a goal is defined as follows.

Definition 3 *Given a set of facts KB, a goal is any proposition φ such that*

$$M \models \mathcal{I}(\varphi \mid Cl(KB)) \tag{1}$$

where $Cl(KB)$ is the default closure of the facts KB defined as follows:

$$Cl(KB) = \{\varphi \mid KB \Rightarrow \varphi\} \tag{2}$$

We assume (for simplicity of presentation) that Cl(KB) is finitely specifiable and take it to be a single propositional sentence.[2]

3 Deliberation

We discuss the BDI approach and a cognitive agent system called 3APL that aims at developing computational specifications for more concrete agents. They illustrate how deliberation can lead to decisions.

3.1 Belief, Desire, and Intention (BDI)

A criticism to classical and qualitative decision theory is that they are inadequate for real time applications in dynamic environments. In such applications, the decision making agent should select and execute actions at each moment in time. As the dynamic environment changes either during the selection or the execution of actions, the decision making agent need to deliberate to either reconsider the previously decisions, also called intentions, or continuing the commitment to those decisions (intentions). This deliberation process is argued by Bratman [2] to be crucial to realize stable behavior of decision making agents with bounded resources. This approach has led to BDI systems [19] that extends decision theory with the additional deliberation component called intention.

In BDI systems, the partial information on states is reduced to dichotomous values (0-1); the propositions are believed or not. This abstraction is called the beliefs of the decision making agent. Similarly, the partial information about the objectives of the decision making agent is reduced to dichotomous values (0-1); the propositions are desired or not. This abstraction is called the desires of the decision making agent. Finally, the partial information about the previous decisions, to which the agent is still committed to, is represented by dichotomous values (0-1); the proposition are intended or not.

3.2 3APL

3APL is a computational specification for cognitive agents. The specification language consists of two parts: the object-level and the meta-level specification language. The expressions

[1] In default logic, an exception is a digression from a default rule. Similarly, in deontic logic an offense is a digression from the ideal.

[2] A sufficient condition for this property is that each "cluster" of equally normal worlds in $\leq_N$ corresponds to a finitely specifiable theory. This is the case in, e.g., System Z.

of the object-level language specifies the cognitive attitudes of agents such as beliefs, goals, and plans (intentions), and additional ingredients such as actions that can be performed by the agent and reasoning rules that it can apply. The expressions of the meta-level language, also called the deliberation language [6, 7], specify the deliberation process that determines which actions should be performed at each moment in time. These expressions specify many decisions such as which goal to select, which reasoning rules to apply, which plan to execute, if and which goal to revise, and if and which plan to revise.

The deliberation language is a many-sorted language [6, 7]. It includes sorted terms that refer to object-level entities such as beliefs, goals, actions, plans, and rules, predicates that express relations between terms, and statements that selects goals, plans, reasoning rules, and generate plans or examine the consequences of the plans. For now it is important to illustrate only the statements of the deliberation language through which various types of deliberation activities can be implemented.

Definition 4 *Let s be any sort, t_s be a term of sort s, and V_s be a variable of sort s. The set of basic statements of the deliberation language is defined as follows:*

- $V_s := t_s$
- $selgoal(t_{sg}, f_c, V_{ig})$
  $selint(t_{si}, f_c, V_{ii})$
  $selplan(t_{sp}, f_c, V_{ip})$
  $selrule(t_{sxrr}, t_{sx}, V_{ixrr})$ *for* $x \in \{g, i, p\}$
- $update(t_{sx}, t_{iy})$ *for* $x, y \in \{b, g, i\}$
  $reviseplan(t_{ip}, t'_{ip})$
- $plan(t_{ii}, t_N)$
  $replan(t_{ip}, t_{spr}, f_c, t_N)$
  $btplan(t_{ip}, t_{spr}, t_N)$
  $explan(t_{ip})$

The set of deliberation statements is defined as follows:

- *Basic statements are deliberation statements*
- *If $\varphi \in DF$ is a deliberation formula, and α and β are deliberation statements, then the following are deliberation statements:*
  α ; β ,
  `IF` φ `THEN` α `ELSE` β ,
  `WHILE` φ `DO` α

The first statement $V_s := t_s$ is designed to assign a sorted term t_s to a variable V_s of the same sort. The following statements are all selecting some item from a particular set of those items. The statement $selgoal(t_{sg}, f_c, V_{ig})$ selects an individual goal from the set of goals denoted by the term t_{sg}. The term denoting the selected individual goal is assigned to variable V_{ig}. The function f_c maps goals to boolean values indicating whether the goal formula satisfies the criterium c. The statement $selint(t_{si}, f_c, V_{ii})$ selects an individual intention from the set of intentions denoted by the term t_{si}. The term denoting the selected individual intention is assigned to variable V_{ii}. The function f_c indicates whether the intention satisfies the criterium c. The statement $selplan(t_{sp}, f_c, V_{ip})$ selects an individual plan from the set of plans denoted by the term t_{sp}. The term denoting the selected plan should satisfy criterion f_c and is assigned to the variable V_{ip}. The statement $selrule(t_{sxrr}, t_{sx}, V_{ixrr})$ selects a rule from the set of (goal, intention, or plan) reasoning rules denoted by the terms t_{sxrr}

and assigns the term that denotes the rule to the variable V_{ii}. The selected rule should be applicable to a formula from the set denoted by the term t_{sx}.

The criterium c used in the selection functions can be used to define a preference ordering between the goals, intentions, plans and rules. So, in fact this is the place where a relation with qualitative decision theory can be made. The same argument can be made for the other selection functions. The main advantage over the classical decision theoretic approach is that the deliberation uses several independent preference orderings over different concepts. The combination of all these orderings leads to a decision on which action will be performed next. However, unlike decision theory where all factors have to be combined into one function that determines the best action, we explicitly program this combination. Besides this advantage of having all factors explicitly available (and thus easily adjustable) the combination of these factors into a decision can be made situation dependent, therefore allowing for an adjustable preference ordering of the agent.

The statement $update(t_{sx}, t_{iy})$ updates a mental base (belief, goal, or intention base) denoted by the term t_{sx} with the formula denoted by the term t_{iy}. The statement $reviseplan(t_{ip}, t'_{ip})$ removes the plan that is denoted by the term t_{ip} from the plan base, and adds the plan that is denoted by the term t'_{ip} to it.

The final set of basic statements are all related to updating the plans of the agent in some way. The statement $plan(t_{ii}, t_n)$ generates a plan expression with maximum length t_n to achieve intention t_{ii}. The generated plan expression is assigned to the plan base of the agent. The statement $replan(t_{ip}, t_{spr}, f_c, t_N)$ uses the set of planning rules t_{spr} and generates a new plan expression to replace the plan expression t_{ip}. The new plan expression satisfies the criteria f_c and has maximum length t_N. The statement $btplan(t_{ip}, t_{spr}, t_N)$ does the same as replan except that *btplan* uses an order among planning rules and generates the next plan expression according to that order. Finally, $explan(t_{ip})$ executes the individual plan expression denoted by the term t_{ip}. We assume that the execution of a plan has some external effects. The internal mental effects should be realized using the update statement explicitly.

In this paper, we do not consider the formal semantics of this deliberation language since we are only interested in the implementation of the deliberation process and how autonomous agent properties related to the agent's decision making ability can be implemented. The semantics for this language is an extension of the semantics of the meta-language already presented in [6].

4 Agent types

Agent types are patterns of deliberation. The most famous ones are the realism and commitment strategies defined in BDI systems. Agent types that are introduced in the BDI tradition can be considered as specification of abstract decision patterns that determine which state can be chosen as the goal state. These decision patterns can be extended in two directions. The first direction is to consider additional cognitive concepts that influence the decision making behavior such as obligation and norms. The goal state can then be chosen not only from desires, but also from obligation or norm states since obligation and norms are external motivational attitudes. Like in the BDI tradition, the set of goal states can be constrained by a number of axioms that determine which states can be chosen as the goal state. Another direction to extend the specification of abstract decision patterns is to consider the specifications of more concrete decision patterns or agent types. The latter extension results in what is called the specification of deliberation process which is usually formulated as a iterative procedure that repeats itself infinitely.

4.1 BDI-CTL

As a typical example of a BDI model, we discuss Rao and Georgeff's initial BDI-CTL logic [19]. In this rather complicated semantics of the BDI-CTL logic, the belief, desire, and intention of decision making agents are represented by the relation $\mathcal{B}$, $\mathcal{D}$, and $\mathcal{I}$, respectively. In BDI-CTL, the formalization of mental states (belief-desire-intention) is extended with Computational Tree Logic CTL and CTL* in order to represent how the mental states of the decision making agent change through time. Time is considered as an abstraction of actions that the agent can perform.

Definition 5 (Semantics of BDI logic [19]) *An interpretation M is defined to be a tuple $M = \langle W, E, T, <, U, \mathcal{B}, \mathcal{D}, \mathcal{I}, \Phi \rangle$, where W is the set of worlds, E is the set of primitive event types, T is a set of time points, $<$ a binary relation on time points, U is the universe of discourse, and Φ is a mapping from first-order entities to elements in U for any given world and time point. A situation is a world, say w, at a particular time point, say t, and is denoted by w_t. The relations $\mathcal{B}$, $\mathcal{D}$, and $\mathcal{I} \subseteq W \times T \times W$ map the agent's current situation to her beliefs, desire, and intention-accessible worlds, respectively.*

BDI does not involve an explicit notion of actions, but instead models possible events that can take place through time by the branching time. The branching time structure models possible sequences of events and determines the alternative cognitive worlds that an agent can bring about through time. Thus, each branch in the temporal structure represents an alternative the agent can choose. Uncertainty about the effects of actions is not modeled by branching time, but by distinguishing between different belief worlds. In effect, they model all uncertainty about the effects of actions as uncertainty about the present state, a well known representation from decision theory [22]. This translation is discussed at length in [18].

The consequence of the fact that we no longer have pre-orders, like in Boutilier's logic, but only an unstructured set, is that we can now only represent monadic expressions like $B(\varphi)$ and $D(\varphi)$, no longer dyadic expressions like $I(\varphi|\psi)$. We have that a world at a time point of the model satisfies $B(\varphi)$ if φ is true in all belief accessible worlds at the same time point. Since there is no such ordering on the possible worlds in BDI, each desire world can in principle be chosen in BDI as the goal world which need to be achieved. It may be clear that it is not an intuitive idea to select a desire world as the goal world in a random way since desire worlds can be in principle not belief worlds. Selecting a desire world which is not believed results in wishful thinking and thus suggests an unrealistic decision. Therefore, BDI proposes a number of constraints under which each desire world can be chosen as a goal world. These constraints are usually characterized by some axioms called realism, strong realism or weak realism [20, 4].

In particular, the realism constraint states that agent's desire should be consistent with its beliefs. Note that this constraint is the same as in QDT where the goal worlds should be consistent with the belief worlds. Formally, the realism axiom states that the set of desire accessible worlds should be a subset of the set of belief accessible worlds, i.e.

$$B(\varphi) \rightarrow D(\varphi)$$

and, moreover, the belief and desire worlds should have identical branching time structure (the same alternatives), i.e.

$$\forall w \forall t \forall w' \text{ if } w' \in \mathcal{D}_t^w \text{ then } w' \in \mathcal{B}_t^w$$

The definition of realism includes an additional axiom to reduce the set of desire worlds and to guarantee that chosen desire world is consistent with the worlds that are already chosen

as the goal worlds. Formally, this axiom states that the intention accessible worlds should be a subset of desire accessible worlds, i.e.

$$D(\varphi) \rightarrow I(\varphi)$$

and, moreover, the desire and intention worlds should have identical branching time structure (the same alternatives), i.e.

$$\forall w \forall t \forall w' \text{ if } w' \in \mathcal{I}_t^w \text{ then } w' \in \mathcal{D}_t^w$$

In addition to these constraints, which are classified as static constraints, there are different types of constraints introduced in BDI resulting in additional agent types. These axioms determine when intentions or previously decided goals should be reconsidered or dropped. In BDI, choosing to drop a goal is thus considered to be as important as choosing a new goal. These constraints, called commitment strategies, involve time and intentions and express the dynamics of decision making. The well-known commitment strategies are 'blindly committed decision making', 'single-minded committed decision making', and 'open-minded committed decision making'. For example, the single-minded commitment strategy states that an agent remains committed to its intentions until either it achieves its corresponding objective or does not believe that it can achieve it anymore.

These static and dynamic properties determine the type of decision making behavior. Thus, in contrast to QDT, where the decision rule is based on the ordering on the possible worlds, in BDI any desire world that satisfies a set of assumed axioms such as realism, can be chosen as the goal world, i.e., BDI systems do not consider decision rules but they consider agent types. Although, there is no such issue as agent type in classical or qualitative decision theory, there are discussions which can be related to agent types. For example, in discussions about risk attitude, often a distinction is made between risk neutral and risk averse.

4.2 BOID

In the BOID approach [3] the actions are decided by first generating the goals that should be achieved, and then planning them. The goals are generated based on the derivation of so-called extensions in default logic [21]. Default logic extends the inference rule modus ponens with two new mechanisms. First, there is a consistency constraint on the inference process, such that rules are applied only if they do not lead to an inconsistency. Second, the application of defeasible rules may result in conflicting outputs and thus in conflicting goal sets. They lead to alternative sets of logic formulas. However, to resolve the conflict we have to consider the whole extension, because agents should consider the effects of goals before they commit to them. In the BOID approach, goal generation is based on prioritized default logic. The specification of goal generation process - the instantiation of a default theory - contains the specification of a set of facts, a set of rules, and the specification of a priority function ρ on the rules. To keep the formal details in this paper to a minimum we assume that individual extensions do not contain disjunctive information, that is, we assume that extensions are sets of positive or negated atomic formulas called literals.

The specification of goal generation process is given in Definition 6. The goal generation process starts with a set of observations Obs, which cannot be overridden, and initial sets of default rules for B, O, I, and D. The procedure then determines a sequence of sets of extensions $S_0, S_1, \ldots$. The first element in the sequence is the set of observations: $S_0 = \{Obs\}$. A set of extensions S_{i+1} is calculated from a set of extensions S_i by checking for each extension E in S_i whether there are rules that can extend the extension. There can be none, in which case nothing happens. Otherwise each of the consequents of the applicable

rules with highest ρ-value are added to the extension separately, to form distinct extensions in S_{i+1}. The operator $Th(S)$ refers to the logical closure of S, and the syntactic operation $Lit(b)$ extracts the set of literals from a conjunction of literals b. In practice not the whole set of extensions is calculated, but only those that are calculated before the agent runs out of resources.

Definition 6 (Generate goal process) *Let $\Delta = \langle Obs, B, O, I, D, \rho \rangle$ be the specification of the goal generation process for propositional logic L, and let an extension E be a set of L literals (an atom or its negation). We say that:*
- a rule $(a \hookrightarrow b)$ is strictly applicable to an extension E, iff $a \in Th(E)$, $b \notin Th(E)$ and $\neg b \notin Th(E)$;
- $\max(E, \Delta) \subseteq B \cup O \cup I \cup D$ is the set of rules $(a \hookrightarrow b)$ strictly applicable to E such that there does not exists a $(c \hookrightarrow d) \in B \cup O \cup I \cup D$ strictly applicable to E with $\rho(c \hookrightarrow d) > \rho(a \hookrightarrow b)$;
- $E \subseteq L$ is an extension for Δ iff $E \in S_n$ and $S_n = S_{n+1}$ for the following procedure:

$i := 0$; $S_i := \{Obs\}$;
repeat
  $S_{i+1} := \emptyset$;
  for all $E \in S_i$ ***do***
    if *exists* $(a \hookrightarrow b) \in B \cup O \cup I \cup D$ *strictly applicable to* E ***then***
      for all $(a \hookrightarrow b) \in \max(E, \Delta)$ ***do***
        $S_{i+1} := S_{i+1} \cup \{ E \cup Lit(w) \}$;
      end for
    else
      $S_{i+1} := S_{i+1} \cup \{E\}$;
    end if
  end for
  $i{:=}i{+}1$;
until $S_i = S_{i-1}$;

Agent types are used to distinguish, classify and compare agent decision making behavior. In this paper, we consider agent types that are defined in terms of goal generation process. These agent types are based on overriding, such that in realistic agents beliefs override other mental attitudes, and in social agents obligations override desires. Agent types based on goal generation are conflict resolution methods. An agent has a conflict if the goal generation process in Definition 6 derives multiple extensions. A conflict is resolved if the priority function is adapted such that no alternative extensions are generated. A mental attitude conflicts with another mental attitude if two rules from different attitudes are applicable, but applying both leads to an inconsistent set. A rule overrides another rule if it has a higher priority. Finally, agent types based on goal generation are formalized in Definition 7 as constraints on the set of available priority functions. When the process starts the selected priority function obeys the constraints corresponding to the agent type. An agent type is called primitive if it contains only one constraint, and complete if it induces a total strict ordering on the components.

Definition 7 *An agent type based on goal generation is a consistent set of constraints on priority function ρ of the form $X \succ Y$ with $X, Y \in \{B, O, I, D\}$ defined as follows:*
$\forall rule\_x \in X \ \forall rule\_y \in Y \ \rho(rule\_x) > \rho(rule\_y)$
A primitive agent type contains a single constraint. A complete agent type is a maximal consistent set of constraints.

There are twelve primitive agent types, which are listed in Table 1 together with the corresponding constraint. They are ordered in six pairs, each agent type A together with its

inverse $\{X \succ Y \mid Y \succ X \in A\}$. An agent type is a set of primitive agent types. For example,

Constraint	Agent type
$B \succ O$ ($O \succ B$)	Realistic with respect to obligations (dogmatic)
$B \succ I$ ($I \succ B$)	Realistic with respect to intentions (over-committed)
$B \succ D$ ($D \succ B$)	Realistic with respect to desires (wishful thinker)
$O \succ I$ ($I \succ O$)	(Un)Stable with respect to obligations
$O \succ D$ ($D \succ O$)	Social (selfish)
$I \succ D$ ($D \succ I$)	(Un)Stable with respect to desires

Table 1: Twelve primitive agent types

the realistic agent type is $\{B \succ O, B \succ I, B \succ D\}$, the stable agent type is $\{I \succ O, I \succ D\}$, the social stable agent type is $\{I \succ O, I \succ D, O \succ D\}$, etc. Moreover, agent types can be derived. For example, since orderings are transitive we can derive that an agent which is unstable with respect to obligations ($O \succ I$) and stable with respect to intentions ($I \succ D$) is social ($O \succ D$). There are twenty-four complete agent types, of which the six realistic ones are listed in Table 2. The definition of agent types leads to a simple way in which

Constraints	Agent type
$B \succ O\ O \succ I\ I \succ D$	Realistic, unstable-O, stable-D, social
$B \succ O\ O \succ D\ D \succ I$	Realistic, unstable-O, unstable-D, social
$B \succ I\ I \succ O\ O \succ D$	Realistic, stable-O, stable-D, social
$B \succ I\ I \succ D\ D \succ O$	Realistic, stable-O, stable-D, selfish
$B \succ D\ D \succ O\ O \succ I$	Realistic, unstable-O, unstable-D, selfish
$B \succ D\ D \succ I\ I \succ O$	Realistic, stable-O, unstable-D, selfish

Table 2: Six complete realistic agent types

agent types can be compared. Agent type A is at least as general as agent type B if all the priority functions that respect constraints of agent type A, also respect the constraints of agent type B. The generality relation between realistic agent types forms the lattice visualized in Figure 1 if we add a top element. This figure should be read as follows. The level in this hierarchy indicates the generality of agent types. The bottom of this lattice is the realistic agent type. Each higher layer adds additional constraints resulting in more specific agent types. The top of this lattice is the falsum which indicates that adding any additional constraint to the ρ function results in an inconsistent ordering. Just below are the six complete realistic agent types.

There are also other constraints on priority functions. One of them is the following unique extension property, which says that ρ associates with each rule a unique integer. It induces a strict total order on the rules.

Definition 8 (Unique goal set) *A goal generation process that generates unique goal sets is specified by a tuple $\langle Obs, B, O, I, D, \rho\rangle$ with ρ a function from $B \cup O \cup I \cup D$ to the integers such that $\rho(x) = \rho(y)$ implies $x = y$.*

Another constraint on priority function ρ is the following attitude order property.

Definition 9 (Attitude order) *A goal generation process specified by a tuple $\langle Obs, B, O, I, D, \rho\rangle$ induces a strict attitude order when ρ is a function from $B \cup O \cup I \cup D$ to the integers such that for all $X, Y \in \{B, O, I, D\}$ with $X \neq Y$ we have either $X \succ Y$ or $Y \succ X$.*

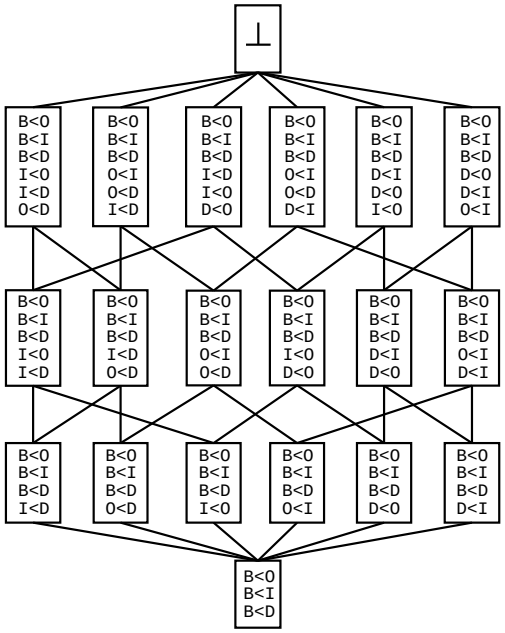

Figure 1: The lattice structure of realistic agent types.

The only agent types that satisfy the attitude order property are the complete agent types. Lack of the attitude order property is illustrated by the realistic agent. It starts with the observations and calculates belief extensions by iteratively applying belief rules. When no belief rule is applicable anymore, then the set of either O, the I, or the D rules are chosen arbitrarily. The chosen set is called the active set of rules. If a rule from the active set of rules is applicable, then the rule is selected and applied. When the rule is applied successfully, the belief rules are attended again and belief rules are applied. If there is no rule from the active set of rules O, I, or D is applicable, then another set of rules becomes the active set of rules. If there is no rule from any of the O, I, or D applicable, then the process terminates – a fixed point is reached – and extensions are calculated.

5 Summary

We consider a range of systems, which are concerned with decisions, deliberation or agent types. We ask ourselves the following question:

> How does deliberation based on agent types lead to decisions?

This question breaks down in three sub-questions. First, what is an agent decision? To answer this question we discuss theories which define what a decision is, so-called decision theories. The first is classical decision theory [15, 22] and the second it qualitative extension (QDT) [1, 17]. In all these theories, a decision is the best option among a set of alternatives. We also illustrate that classical as well as recently proposed qualitative decision theories do not explain how decisions can be found, but in Boutilier's decision theory the concept of goal plays a central role. This is also a crucial concept in the three theories of deliberation studied in this paper.

Second, what is agent deliberation and how does it lead to a decision? To answer this question we discuss two other approaches. The first approach we discuss is based on an

abstract model of the mental attitudes of an agent: beliefs, desires and intentions (BDI) [2, 5, 19]. The second (3APL) is similar to BDI except that the decision rules is replaced by a process called the deliberation process. It is this process that determine which actions should be performed for a given set of underlying cognitive concepts. The discussion on BDI systems and 3APL illustrates how deliberation finds decisions. The deliberation process is based on cognitive concepts like beliefs, desires, goals, intentions and plans. The system generates goals, and thereafter finds plans to achieve these goals.

Third, how do agent types classify patterns of deliberation? The fifth approach we discuss to answer this question is also based on mental attitudes extended with obligations (BOID) [3]. The discussion on BDI-CTL and BOID illustrates how agent types classify deliberation. The deliberation process in BDI and BOID has been characterized in terms of agent types or deliberation patterns. In BDI-CTL the agent types represent how beliefs, goals and intentions are related, and when goals are maintained or dropped. In BOID the agent types represent how the agent resolves his conflicts. For example, a selfish agent lets its desires take precedence over his obligations. Using cognitive concepts and formulating the decision rule in terms of deliberation patterns makes the later two approaches cognitive theories of decision making.

Acknowledgments

Thanks to Jan Broersen, Zhisheng Huang and Joris Hulstijn for many discussions on related subjects.

References

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

[2] M. Bratman. *Intention, plans, and practical reason*. Harvard University Press, Cambridge Mass, 1987.

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

[4] J. Broersen, M. Dastani, and L. van der Torre. Realistic desires. *Journal of Applied Non-Classical Logics*, 12(2):287–308, 2002.

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

[6] M. Dastani, F. de Boer, F. Dignum, and J.-J. Meyer. Programming agent deliberation: An approach illustrated using the 3apl language. In *Proceedings of The Second Conference on Autonomous Agents and Multi-agent Systems (AAMAS'03)*, Melbourne, Australia, 2003.

[7] M. Dastani, F. Dignum, and J.-J. Meyer. Autonomy and agent deliberation. In *Proceedings of The First International Workshop on Computational Autonomy - Potential, Risks, Solutions (Autonomy'03)*, Melbourne, Australia, 2003.

[8] M. Dastani, J. Hulstijn, and L. van der Torre. How to decide what to do? *European Journal of Operations Research*, to appear.

[9] T. Dean and M. Wellman. *Planning and Control*. Morgan Kaufmann, San Mateo, 1991.

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

[11] J. Doyle, Y. Shoham, and M.P. Wellman. The logic of relative desires. In *Sixth International Symposium on Methodologies for Intelligent Systems (ISMIS-91)*, volume LNAI Volume 542, pages 16–31, Charlotte, North Carolina, 1991.

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

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

[14] B. Hansson. An analysis of some deontic logics. *Nous*, 3:373–398, 1969.

[15] R. C. Jeffrey. *The Logic of Decision*. McGraw-Hill, New York, 1965.

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

[17] J. Pearl. From conditional ought to qualitative decision theory. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence (UAI'93)*, pages 12–20. John Wiley and Sons, 1993.

[18] A. Rao and M. Georgeff. Deliberation and its role in the formation of intentions. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence (UAI-91)*, pages 300–307, San Mateo, CA, 1991. Morgan Kaufmann Publishers.

[19] A. Rao and M. Georgeff. Modeling rational agents within a bdi architecture. In *Proceedings of Second International Conference on Knowledge Representation and Reasoning (KR'91)*, pages 473–484. Morgan Kaufmann, 1991.

[20] A. Rao and M. Georgeff. Decision procedures for BDI logics. *Journal of Logic and Computation*, 8:293–342, 1998.

[21] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[22] L. Savage. *The foundations of statistics*. Wiley, New York, 1954.

[23] S.-W. Tan and J. Pearl. Qualitative decision theory. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI'94)*, pages 928–932, 1994.

[24] S.-W. Tan and J. Pearl. Specification and evaluation of preferences under uncertainty. In *Proceedings of the Fourth International Conference on Knowledge Representation and Reasoning (KR'94)*, pages 530–539. Morgan Kaufmann, 1994.