A Logical Framework to Represent and Reason about Graded Preferences and Intentions

Ana Casali
Facultad de Cs. Exactas, Ingeniería y Agrimensura
Universidad Nacional de Rosario
Av. Pellegrini 250, 2000 Rosario, Argentine

Lluís Godo and **Carles Sierra**
Institut d'Investigació en Intel·ligència Artificial
Campus Universitat Autònoma de Barcelona s/n
08193 Bellaterra, Spain.

Abstract

In intentional agents, actions are derived from the mental attitudes and their relationships. In particular, preferences (positive desires) and restrictions (negative desires) are important proactive attitudes which guide agents to intentions and eventually to actions. In this paper, we present a general logical framework to represent and reasoning about gradual notions of desires and intentions, including sound and complete axiomatizations. Some extensions are proposed corresponding to certain additional constraints that the agent can set about the kind of preferences she is dealing with. We also show that the framework is expressive enough to describe how desires, together with other information, can lead agents to intentions.

Introduction

Preferences are essential for making choices in complex situations, for mastering large sets of alternatives, and for coordinating a multitude of decisions. In particular, preferences are the proactive attitudes in intentional agents. From the positive preferences or desires the agent may choose which ones he will intend to achieve through a suitable plan of actions. Negative preferences are also considered in modelling different AI problems and particularly in multi-agent systems. For an intentional agent negative preferences may represent restrictions or rejections over the possible worlds she can reach.

In intentional agents and particularly in BDI architectures, (Rao and Georgeff 1991) desires represent the *ideal* agent preferences regardless of the agent current perception of the environment and regardless of the cost involved in actually achieving them. We deem important to distinguish what is positively desired from what is not rejected. According to the works on bipolarity representation of preferences in (Benferhat *et al.* 2002b) and (Benferhat *et al.* 2006), positive and negative information may be modeled in the framework of possibilistic logic (see also (Bistarelli *et al.* 2007) for another kind of approach). Inspired by this line of work, we suggest to formalize both positive and negative agent desires. Positive desires represent what the agent would like to be the case. Negative desires correspond to what the agent rejects or does not want to occur. Furthermore, positive and negative desires can be graded to represent different levels of preference or rejection, respectively. When dealing with both kinds of preferences it is also natural to express indifference, meaning that we have neither a positive nor a negative preference over an object.

In this work, we present a logical framework (i.e. language, axioms and inference rules) to represent and reason about agent desires in the sense of bipolar graded preferences. Then, we extend this framework to represent intentions and finally we give some insights of how they can be used to lead to agent actions.

With respect to the previous works by Benferhat et al. on bipolar preference representation, we advance the state of the art by giving a sound and complete axiomatic and defining different logical schemas to represent some additional constraints over preferences. Also, we present a logical system for intentions and we show that the framework is expressive enough to describe how desires (either positive or negative), together with other information, can lead agents tointentions.

For this purpose, we first define a modal-like language to express graded positive and negative desires with its corresponding semantics. Indeed, a many-valued modal setting is used to deal with desire degrees, following the approach presented in (Hájek 1998) to reason about probabilities and other uncertainty measures. Then, an appropriate layered structure of axioms is defined to capture the behavior of these preferences. The language with this set of axioms constitute the basic logic framework for representation and reasoning about desires.

Since positive and negative preferences are stated separately, it is worthwhile to consider whether any consistency condition should be imposed among them, as done e.g. in (Benferhat *et al.* 2006). We consider that it may be somewhat controversial and domain dependent to set (normative) general restrictions about e.g. positive (negative) desires both on some goal and its negation, or also between the positive and negative desires on a same given goal. Therefore, besides the basic framework for the bipolar desires representation, some (alternative) additional constraints are analyzed in this work, resulting in different logical schemas or theories.

After representing positive and negative desires we ex-

tend the logical framework to represent the agent intentions. Intentions, as well as desires, represent agent preferences. However, we consider that intentions cannot depend just on the benefit, or satisfaction, of reaching a goal, but also on the world state and the cost of transforming it into a world where the pursued goal becomes true. Then, as in (Cohen and Levesque 1990), intentions result from the agent beliefs and desires and possibly other (utilitarian) information, and hence we do not consider them as a really basic mental attitude. By allowing a many-valued representation of the strength of intentions we are able to attach to intentions a graded measure of the cost/benefit relation involved in the agent actions toward the intended goal. In this direction we provide in the last part of the paper some insights on how the positive and negative desires together with other information, can lead to the best agent intention and eventually to the plan to follow.

Basic Logic for Desires representation

The Language

We start from a basic propositional language $\mathcal{L}$ built from a *finite* set of propositional variables *Var* and the classical propositional logic connectives $\neg, \rightarrow$, with $\wedge, \vee, \equiv, \bot, \top$ being defined from the former ones in the usual way. To represent positive and negative desires over formulas of $\mathcal{L}$, we introduce in such a language two (many-valued) modal-like operators D^- and D^+. If $\varphi \in \mathcal{L}$, then $D^+\varphi$ and $D^-\varphi$ are many-valued formulas, which respectively read as "φ is positively desired" and "φ is negatively desired" (or "φ is rejected"), and whose truth degrees will respectively represent the agent level of satisfaction would φ become true and her level of disgust on φ becoming true. We will use Rational Pavelka logic RPL (see Annex I), i.e. $[0,1]$-valued Łukasiewicz logic expanded with rational truth-constants (Hájek 1998), as the base logic to deal with grades of desires.

More precisely, formulas of the expanded language $\mathcal{L}_D$ are defined as follows:

- If $\varphi \in \mathcal{L}$ then $\varphi \in \mathcal{L}_D$
- If $\varphi \in Sat(\mathcal{L})$[1] then $D^-\varphi, D^+\varphi \in \mathcal{L}_D$
- If $r \in \mathbb{Q} \cap [0,1]$ then $\overline{r} \in \mathcal{L}_D$
- If $\Phi, \Psi \in \mathcal{L}_D$ then $\Phi \rightarrow_L \Psi \in \mathcal{L}_D$ and $\neg_L \Phi \in \mathcal{L}_D$ (other Lukasiewicz logic connectives, like $\oplus, \otimes, \wedge_L, \vee_L, \equiv_L$ are definable from $\neg_L$ and $\rightarrow_L$)

We will call a modal formula *closed* when every propositional variable is in the scope of a D^+ or a D^- operator.

The notation $(D^+\psi, r)$, with $r \in [0,1] \cap \mathbb{Q}$, will be used as a shortcut of the formula $\overline{r} \rightarrow_L D^+\psi$, which specifies that the level of positive desire of ψ is at least r [2]. Analogously for $(D^-\psi, r)$ and $\overline{r} \rightarrow_L D^-\psi$.

[1]Thus excluding to have positive and negative desires on a contradiction ($\bot \notin Sat(\mathcal{L})$).

[2]Note that, according to Lukasiewicz logic semantics (see Annex I), $\overline{r} \rightarrow_L D^+\psi$ gets value 1 whenever $D^+\psi$ gets a value greater or equal to r.

An specification of the agent's desires will consist of a (finite) set of modal formulas or theory $\mathcal{T}$ in the expanded language $\mathcal{L}_D$, representing all the available information about agent desires. Such a specification may contain quantitative expressions about positive and negative preferences, like $(D^+\varphi, \alpha)$ or $(D^-\psi, \beta)$, as well as qualitative expressions like $D^+\psi \rightarrow_L D^+\varphi$ (resp. $D^-\psi \rightarrow_L D^-\varphi$), expressing that φ is at least as preferred (resp. rejected) as ψ. In particular if $(D^+\phi_i, 1) \in \mathcal{T}$ it means that the agent has maximum preference in ϕ_i and is fully satisfied if it is true, while if there is no $\alpha > 0$ such that $(D^+\phi_j, \alpha) \in \mathcal{T}$, it means that the agent has no positive preference for ϕ_j, i.e. the agent does not have any benefit from ϕ_j becoming true. Analogously, if $(D^-\psi_i, 1) \in \mathcal{T}$ it means that the agent absolutely rejects ϕ_i and thus the states where ψ_i is true are totally unacceptable. If $(D^-\psi_j, \beta) \notin \mathcal{T}$ for any $\beta > 0$ it simply means that ψ_j is not rejected.

Semantics

Many people can argue that, considering the desires as proactive attitudes, reasoning about desires of disjunctions of formulae may have no sense. In most cases we may have plans for achieving φ or ψ individually, or for both $(\varphi \wedge \psi)$ but not for achieving non-deterministically $\varphi \vee \psi$. But since we define the basic language as a propositional language, it is necessary to define the semantics in terms of preferences for disjunctive formulas, and the selection of what formulas are used to reason about desires are left to the definition of a particular theory.

According to the semantics presented in (Benferhat *et al.* 2002b), the degree of positive desire for (or level of satisfaction with) a disjunction of goals $\varphi \vee \psi$ is taken to be the minimum of the degrees for φ and ψ. Intuitively, if an agent desires $\varphi \vee \psi$ then it is ready to accept the situation where the less desired goal becomes true, and hence to accept the minimum satisfaction level produced by one of the two goals. In contrast the satisfaction degree of reaching both φ and ϕ can be strictly greater than reaching one of them separately. These are basically the properties of the *guaranteed possibility* measures (see e.g. (Benferhat *et al.* 2002a)). Analogously, we assume the same model for the degrees of negative desire or rejection, that is, the rejection degree of $\varphi \vee \psi$ is taken to be the minimum of the degrees of rejection for φ and for ψ separately, while nothing prevents the rejection level of $\varphi \wedge \psi$ be greater than both.

The intended models for $\mathcal{L}_D$ are Kripke structures $M = \langle W, e, \pi^+, \pi^- \rangle$, that will be called *Bipolar Desire models* (BD-models for short), where W is a non-empty set of possible worlds and $e : Var \times W \rightarrow \{0,1\}$ is such that, for every world w, $e(\cdot, w)$ is a truth assignment of propositional variables at world w, and is extended to an evaluation $e : \mathcal{L} \times W \rightarrow \{0,1\}$ of arbitrary non-modal formulas in the usual way. Besides, π^+ and π^- are positive and negative preference distributions respectively over worlds, which are used to give semantics to positive and negative desires:

- $\pi^+ : W \rightarrow [0,1]$ is a distribution of positive preferences over the possible worlds. In this context $\pi^+(w) < \pi^+(w')$ means that w' is more preferred than w.

- $\pi^- : W \to [0,1]$ is a distribution of negative preferences over the possible worlds: $\pi^-(w) < \pi^-(w')$ means that w' is more rejected than w.

The truth evaluation of atomic modal formulas $D^-\varphi$ and $D^+\varphi$ is defined as:

- $e(D^+\varphi, w) = \inf\{\pi^+(w') \mid e(\varphi, w') = 1\}$
- $e(D^-\varphi, w) = \inf\{\pi^-(w') \mid e(\varphi, w') = 1\}$

together with the assumption of $\inf \emptyset = 1$. This is extended to compound modal formulas by means of the usual truth-functions for Łukasiewicz connectives (see Annex I). Notice that the evaluation $e(w, \Phi)$ of a modal formula Φ only depends on the formula itself –represented in the preference measure over the worlds where the formula is true– and not on the actual world $w \in W$ where the agent is situated. In such a case, we will also write $e_M(\Phi)$ for $e(w, \Phi)$. This is consistent with the intuition that desires represent ideal preferences of an agent, regardless of the actual world and regardless of the cost of moving to a world where the desire is satisfied.

We will write $M \models \Phi$ when $e(\Phi, w) = 1$ for all $w \in W$. Moreover, let $\mathcal{M}_{\mathcal{BD}}$ be the class of all Kripke structures $M = \langle W, e, \pi^+, \pi^- \rangle$. Then, for each subclass of models $\mathcal{M} \subseteq \mathcal{M}_{\mathcal{BD}}$, given a theory T and a formula Φ, we will write $T \models_{\mathcal{M}} \Phi$ if $M \models \Phi$ for each model $M \in \mathcal{M}$ such that $M \models \Psi$ for all $\Psi \in T$.

Axioms and Rules

To axiomatize the logic with above intended preference-based semantics we need to combine classical logic axioms for non-modal formulae with Rational Pavelka logic axioms for modal formulae. Also, additional axioms characterizing the behavior of the modal operators D^+ and D^- are needed. As already mentioned, a conjunctive combination of one kind of desires (either positive or negative) preferences may be attached a strictly higher preference value, while the preference value of a disjunctive combination of either positive or negative desires is taken as the minimum of the desire degrees, following the intuition that at least the minimum of satisfaction (rejection) is guaranteed. The following axioms and inference rules aim at capturing these combination properties, considering positive or negative desires independently.

We axiomatically define the *Basic Bipolar Desire logic* (BD logic for short) with the following axioms and rules:

Axioms:

(CPC) Axioms of classical logic for non-modal formulas

(RPL) Axioms of Rational Pavelka logic for modal formulas

(BD0$^+$) $D^+(A \vee B) \equiv_L D^+A \wedge_L D^+B$

(BD0$^-$) $D^-(A \vee B) \equiv_L D^-A \wedge_L D^-B$

Inference Rules:

(MP1) modus ponens for $\to$

(MP2) modus ponens for $\to_L$

Introduction of D^+ and D^- for implications:

(ID$^+$) from $\varphi \to \psi$ derive $D^+\psi \to_L D^+\varphi$

(ID$^-$) from $\varphi \to \psi$ derive $D^-\psi \to_L D^-\varphi$.

The notion of proof, denoted $\vdash_{BD}$, is defined as usual from the above axioms and inference rules.

Notice that the two axioms (BD0$^+$) and (BD0$^-$) define the behavior of D^- and D^+ with respect to disjunctions.

The formalization we present for D^- is somewhat different from the approach presented by Benferhat et al. in (Benferhat *et al.* 2002b), where they used a necessity function (i.e., considering $D^-\phi$ as $N(\neg\phi)$). But in their approach the axiomatic is similar since the axiom (BD0$^-$) we present, results from the necessity axiom (i.e., $N(A \wedge B) \equiv_L N(A) \wedge_L N(B)$).

Finally, the introduction rules for D^+ and D^- state that the degree of desire is monotonically decreasing with respect to logical implication. Moreover, an easy consequence of these rules is that equivalent desires degrees are preserved by classical (Boolean) equivalence.

Lemma 1 *If $\vdash_{CPC}$ denotes deduction in classical propositional calculus, then $\vdash_{CPC} \varphi \equiv \psi$ implies $\vdash_{BD} D^+\varphi \equiv_L D^+\psi$ and $\vdash_{BD} D^-\varphi \equiv_L D^-\psi$.*

The above axiomatization is correct with respect to the defined semantics.

Lemma 2 (soundness) *Let T be a theory and Φ a formula. Then $T \models_{\mathcal{M}_{\mathcal{BD}}} \Phi$ if $T \vdash_{BD} \Phi$.*

Proof: It is a matter of routine to check that the axioms are valid in each BD-model and that the inference rules preserve validity in each BD-model. □

Moreover, the basic BD logic is complete as well for finite theories of closed (modal) formulas.

Theorem 3 (completeness) *Let T be a finite theory of closed formulas and Φ a closed formula. Then $T \models_{\mathcal{M}_{\mathcal{BD}}} \Phi$ iff $T \vdash_{BD} \Phi$.*

Proof: See Annex II. □

Example 1 *María, who lives in busy Buenos Aires, wants to relax for a few days in an Argentinian beautiful destination. She activates a personal agent, based on our BD logical framework, to get an adequate plan, i.e. a tourist package, that satisfies her preferences. She would be very happy going to a mountain place (m), and rather happy practicing rafting (r). In case of going to a mountain place she would like to go climbing (c). On top of this, she wouldn't like to go farther than 1000km from Buenos Aires (f). She is stressed and would like to get to the destination with a short trip. The user interface that helps her express these desires ends up generating a desire theory as follows:*

$$\mathcal{T}_{\mathcal{D}} = \{(D^+m, 0.8), (D^+r, 0.6), D^+m \to_L D^+c, (D^-f, 0.7)\}$$

Once this initial desire theory is generated the tourist advisor personal agent deduces a number of new desires:

$\mathcal{T}_{\mathcal{D}} \vdash_{BD} (D^+(m \wedge r), 0.8)$,
$\mathcal{T}_{\mathcal{D}} \vdash_{BD} (D^+(m \vee r), 0.6)$,

$\mathcal{T}_{\mathcal{D}} \vdash_{BD} (D^+c, 0.8)$

As María would indeed prefer much more to be in a mountain place doing rafting she also expresses the combined desire with a particularly high value: $(D^+(m \wedge r), 0.95)$. *Notice that the extended theory*

$$\mathcal{T}'_{\mathcal{D}} = \mathcal{T}_{\mathcal{D}} \cup \{(D^+(m \wedge r), 0.95)\}$$

remains consistent within BD.

The basic logical schema BD puts almost no constraint on the strengths for the positive and negative desire of a formula φ and its negation $\neg\varphi$. This is in accordance with considering desires as ideal preferences and hence it may be possible for an agent to have contradictory desires. Indeed, the only indirect constraint BD imposes is the following one: if a theory T derives $(D^+\varphi, r)$ and $(D^+\neg\varphi, s)$ then, due to axiom $(BD3^+)$ and rule (ID^+), T also derives both $(D^+\psi, \min(r, s))$ and $(D^+\neg\psi, \min(r, s))$ for any ψ.

In the following section, different properties are added to the preferences as to represent some constraints between the positive and negative desires of a formula and its negation.

Different schemas

The basic schema for preference representation and reasoning provided by the BD logic may be felt too general for some classes of problems and we may want to restrict the allowed assignment of degrees of positive and negative desire (resp. negative) for a formula φ and for its negation $\neg\varphi$. For instance, in the case of considering positive desires as proactive attitudes, it is not an efficient approach to allow to assign non-zero degrees to $D^+\varphi$ and to $D^+\neg\varphi$, since the agent will be looking for plans toward opposite directions, some plans leading to satisfy φ and some others to satisfy $\neg\varphi$.

In the following subsections three different extensions or schemas are proposed to show how different consistency constraints between positive and negative desires can be added to the basic logic, both at the semantical and syntactical level. These different schemas allow us to define different types of agents. Each agent type will accept (respectively restrict) desire formulae in its theory depending on its defined constraints according to the chosen schema.

BD_1 Schema

It may be natural in some domain applications to forbid to simultaneously have positive (in the sense of > 0) desire degrees for $D^+\varphi$ and $D^+\neg\varphi$. This constraint and the corresponding one for negative desires amounts to require that the following additional properties for the truth-evaluations be satisfied in the intended models:

- $\min(e(D^+\varphi, w), e(D^+\neg\varphi, w)) = 0$, and
- $\min(e(D^-\varphi, w), e(D^-\neg\varphi, w)) = 0$.

At the level of Kripke structures, this corresponds to require some extra conditions over π^+ and π^-, namely:

- $\inf_{w\in W} \pi^+(w) = 0$ and
- $\inf_{w\in W} \pi^-(w) = 0$

These are a kind of *anti-normalization* conditions for π^+ and π^-, in the sense that they require the existence of at least one world that is not desired and one world that is not rejected. Let $\mathcal{M}_{\mathcal{BD}_1}$ denote the subclass of models satisfying these conditions.

For instance, following this schema an agent's theory $\mathcal{T}_D$ should not simultaneously contain the formulae $(D^+m, 0.8)$ and $(D^+(\neg m), 0.4)$, or the formulae $(D^-f, 0.7)$ and $(D^-(\neg f), 0.5)$

At the syntactic level, to require these conditions amounts to add to the basic logic BD the following two axioms:

(BD1$^+$) $D^+\varphi \wedge_L D^+(\neg\varphi) \rightarrow_L \bar{0}$
(or equivalently $D^+(\top) \equiv_L \bar{0}$)

(BD1$^-$) $D^-\varphi \wedge_L D^-(\neg\varphi) \rightarrow_L \bar{0}$
(or equivalently $D^-(\top) \equiv_L \bar{0}$)

We will denote by BD_1 the extension of the BD system with the above two axioms (BD1$^+$) and (BD1$^+$), and by $\vdash_{BD_1}$ the corresponding notion of proof.

Theorem 4 (completeness) *Let T be a finite theory of closed formulas and Φ a closed formula. Then $T \models_{\mathcal{M}_{\mathcal{BD}_1}} \Phi$ iff $T \vdash_{BD_1} \Phi$.*

Proof: The proof runs like in Theorem 3 by adding to the theory $\mathcal{BD}$ the instances of axioms (BD1$^+$) and (BD1$^-$) and with the obvious modifications. □

BD_2 Schema

The above logical schema BD_1 does not put any restriction on positive and negative desires for a same goal. According to Benferhat et al. in (Benferhat *et al.* 2006), a coherence condition between positive and negative desires should be considered, namely, an agent cannot desire to be in a world more than the level at which it is tolerated (not rejected). This condition, translated to our framework, amounts to require in the Kripke structures the following constraint between the preference distributions π^+ and π^-:

- $\forall w \in W,\ \pi^+(w) \leq 1 - \pi^-(w)$

To formulate the corresponding axiomatic counterpart that faithfully accounts for the above condition, we consider $\mathcal{M}_{BD_2}$ the subclass of BD-Kripke structures $M = (W, e, \pi^+, \pi^-)$ satisfying the above constraint between $\pi^+$and π^-. Note that $\pi^+(w) \leq 1 - \pi^-(w)$ iff $\pi^+(w) \otimes \pi^-(w) = 0$ [3].

To capture at the syntactical level this class of structures, we consider the extension of the system BC with the following axiom:

(BD$_2$) $(D^+\varphi \otimes D^-\varphi) \rightarrow_L \bar{0}$

We will denote by BD$_2$ the extension of BD with the axiom (BD$_2$)[4].

[3] Here we use the same symbol as the Lukasiewicz connective $\otimes$ to denote its corresponding the truth-function on $[0, 1]$ of , i.e. $x \otimes y = \max(x + y - 1, 0)$ for any $x, y \in [0, 1]$, see Annex I.

[4] An equivalent presentation of axiom (BD2) is $D^+\varphi \rightarrow_L \neg_L D^-\varphi$.

Notice that this axiom is valid in every BD-structure $M = (W, e, \pi^+, \pi^-) \in \mathcal{M}_{BD_2}$. Indeed, for any non modal φ, we have:

$e_M(D^+\varphi \otimes D^-\varphi) =$
$\inf\{\pi^+(w) \mid e(w,\varphi) = 1\} \otimes \inf\{\pi^-(w) \mid e(w,\varphi) = 1\} =$
$\inf\{\pi^+(w) \otimes \pi^-(w') \mid e(w,\varphi) = e(w',\varphi) = 1\} \leq$
$\inf\{\pi^+(w) \otimes \pi^-(w) \mid e(w,\varphi) = 1\} = 0.$

that is, for any φ, the evaluations of $D^+\varphi$ and $D^-\varphi$ are such that $e_M(D^+\varphi) \leq 1 - e_M(D^-\varphi)$.

Conversely, if the (BD_2) axiom is valid in a BD-structure $M = (W, e, \pi^+, \pi^-)$ then necessarily it must satisfy the condition $\pi^+(w) \leq 1 - \pi^-(w)$ for any $w \in W$, i.e. $M \in \mathcal{M}_{BD_3}$.

Proof: W.l.o.g., we can assume that W is such that $e(w,\cdot) = e(w',\cdot)$ iff $w = w'$. Then for each $w \in W$ consider the formula $\varphi_w = (\bigwedge_{p_i \in l^+} p_i) \wedge (\bigwedge_{p_i \in l^-} \neg p_i)$, where $l^+ = \{p \in Var \mid e(w,p) = 1\}$ and $l^- = \{p \in Var \mid e(w,p) = 0\}$. It is clear that $e(v,\varphi_w) = 1$ iff $v = w$, and hence $e(w, D^+\varphi_w) = \pi^+(w)$ and $e(w, D^-\varphi_w) = \pi^-(w)$. Therefore, $e_M(D^+\varphi_w \otimes D^-\varphi_w) = 0$ iff $\pi^+(w) \otimes \pi^-(w) = 0$. □

These properties, together with a suitable adaptation of the proof of Theorem 3, leads to the following completeness result for BD_2.

Theorem 5 (completeness) *Let T be a finite theory of closed formulas and Φ a closed formula. . Then $T \models_{\mathcal{M}_{\mathcal{BD}2}} \Phi$ iff $T \vdash_{BD_2} \Phi$.*

BD_3 Schema

An stronger consistency condition between positive and negative preferences was considered in (Casali *et al.* 2005), requiring that if a world is rejected to some extent, it cannot be positively desired at all. And conversely, if a goal (any classically satisfiable formula) is somewhat desired it cannot be rejected. Indeed, at the semantical level, this amounts to require the intended BD-models $M = (W, e, \pi^+, \pi^-)$ to satisfy the following condition for any $w \in W$:

- $\pi^-(w) > 0$ implies $\pi^+(w) = 0$
  (or equivalently, $\min(\pi^+(w), \pi^-(w)) = 0$)

We will denote by $\mathcal{M}_{BD_3}$ the subclass of BD-Kripke structures satisfying this latter condition.

At the syntactic level, the corresponding axiom that faithfully represents this consistency condition is the following one:

(BD3) $(D^+\varphi \wedge_L D^-\varphi) \rightarrow_L \bar{0}$

We will denote by BD_3 the extension of the BD system by the above axiom (BD3), and by $\vdash_{BD_3}$ the corresponding notion of proof.

Theorem 6 (completeness) *Let T be a finite theory of closed formulas and Φ a closed formula. Then $T \models_{\mathcal{M}_{\mathcal{BD}3}} \Phi$ iff $T \vdash_{BD_3} \Phi$.*

Proof: Again it is an easy adaptation of the proof of Theorem 3. □

Example 2 *(Example 1 continued)*

María, a few days later, breaks her ankle. She activates the recommender agent to reject the possibility of going climbing (c). If María selects for the agent the schema BD_1, the agent simply adds the formula $(D^-c, 1)$ into the former desire theory $\mathcal{T}'_{\mathcal{D}}$, yielding the new theory

$$\mathcal{T}''_{\mathcal{D}} = \{(D^+m, 0.8), (D^+r, 0.6), (D^+(m \wedge r), 0.95), (D^+c, 0.85), (D^-f, 0.7), (D^-c, 1)\},$$

as the schema allows for opposite desires [5].

If María selects BD_2, the formulas D^+c and D^-c are not allowed to have degrees summing up more than 1, and hence the above theory $\mathcal{T}''_{\mathcal{D}}$ becomes inconsistent. Actually, $\mathcal{T}'_{\mathcal{D}}$ becomes also inconsistent under BD_3, BD_3 is stronger than BD_2 (it does not even allow to have non-zero degrees for D^+c and D^-c). In these cases, the agent applies a revision mechanism, for instance to cancel $(D^+c, 0.85)$ from theory.

From Desires to Intentions

After analyzing in the previous section different schemas to model desires in an agent architecture, in this section our aim is to show how these positive and negative desires may be used for the agent to generate intentions.

Intentions, as well as desires, represent another type of agent preferences. In (Rao and Georgeff 1991)'s BDI model of agent, intentions are considered as fundamental pro-attitudes. However, in our approach, intentions results from the agent beliefs and desires and then, we do not consider them as a basic attitude. Indeed, we consider that intentions cannot depend just on the benefit, or satisfaction, of reaching a goal φ –represented in $D^+\varphi$, but also on the world state w and the cost of transforming it into a world w_i where the formula φ is true. Then, by allowing a graded representation of the strength of intentions we are able to attach to intentions a measure of the cost/benefit relation involved in the feasible actions the agent can take toward the intended goal. Note that a similar semantics for intentions is used in (Schut *et al.* 2001), where the net value of an intention is defined as the difference between the value of the intention outcome and the cost of the intention. In (Rao and Georgeff 1991), this relation is summarized in the payoff function over the different paths.

The formalization of the intention's semantics is difficult, because it does not depends only in the formula intended, but also in which *feasible* plan the agent may execute to achieve a state where the formula is valid. Some preliminary work adressing its formalization in another context can be seen in (Casali *et al.* 2005).

[5] The fact of having both positive and negative desires may be handled in different ways depending on the kind of agent's behavior. For instance, if the agent follows (Benferhat *et al.* 2002b)'s approach, where negative desires are used as strong constraints, the agent would then first discard those packages including mountain climbing (that is, D^+c would be ignored), and among the remaining ones it would then look for packages satisfying at least some positive preferences.

Logic for intention representation

We need to extend the BD logical framework introduced in the previous Section to include new modal-like operators to represent graded intentions. We assume the agent has a finite set of actions or plans Π^0 at her disposal to achieve the goals. Then, for each action $\alpha \in \Pi^0$, we introduce an operator I_α, such that the truth-degree of a formula $I_\alpha\varphi$ will represent the strength the agent intends φ by means of the execution of the particular action α. Besides, we introduce another operator I with the idea that $I\varphi$ represents the intention degree of φ through the best plan in Π^0. We also need to introduce in the language special atoms C_α for each $\alpha \in \Pi^0$, whose semantics will represent the (bounded and normalized) cost of executing each action α at each world.

Therefore, for a given (finite) set of actions $\Pi^0 \subset \Pi$, the extended language $\mathcal{L}_{DI}$ will expand the former language $\mathcal{L}_D$ by including a set of propositional variables $Var_{cost} = \{C_\alpha\}_{\alpha\in\Pi^0}$ and elementary modal formulae $I_\alpha\varphi$ and $I\varphi$, where $\varphi \in Sat(\mathcal{L})$. We will also include for technical reasons new unary connectives δ_n, for each natural n, whose semantics will become clear shortly[6].

The semantics defined next shows that the value of the intentions depends on the formula intended to bring about and on the benefit the agent gets with it. It also depends on the agent knowledge on possible plans that may change the world into one where the goal is true, and their associated cost. Formally, the intended models will be enlarged Kripke structures $M = \langle W, e, \pi^+, \pi^-, \{\pi_\alpha\}_{\alpha\in\Pi^0}\rangle$ where now $\pi_\alpha : W \times W \to [0, 1]$ is a *utility* distribution corresponding to action α: $\pi_\alpha(w, w')$ is the utility of applying α to transform world w into world w'. It is implicitly assumed here that a value $\pi_\alpha(w, w') = 0$ represents that action α is not feasible at world w. Further, $e : W \times (Var \cup Var_{cost}) \to [0, 1]$ evaluates in each world propositional variables in such a way that variables from Var are evaluated into $\{0, 1\}$ while propositional variables from Var_{cost} into $[0, 1]$ (so variables from var are two-valued while variables from Var_{cost} are many-valued). Then e is extended to Boolean formulas as usual and to atomic modal formulas as in BD logic together with these additional clauses:

- $e(w, I_\alpha\varphi) = \inf\{\pi_\alpha(w, w') \mid w' \in W, e(w', \varphi) = 1\}$
- $e(w, I\varphi) = \max\{e(w, I_\alpha\varphi) \mid \alpha \in \Pi^0\}$

and to compound modal formulas using still the truth functions of Lukasiewicz logic together with interpretation of the δ_n connectives: $e(w, \delta_n\Phi) = e(w, \Phi)/n$.

As usual, we will write $M \models \Phi$ when $e(\Phi, w) = 1$ for all $w \in W$ and will denote by $\mathcal{M}_{\mathcal{DI}}$ be the class of all Kripke structures $M = \langle W, e, \pi^+, \pi^-, \{\pi_\alpha\}_{\alpha\in\Pi^0}\rangle$. Then, for each subclass of models $\mathcal{M} \subseteq \mathcal{M}_{\mathcal{DI}}$, given a theory T and a formula Φ, we will write $T \models_{\mathcal{M}} \Phi$ if $M \models \Phi$ for each model $M \in \mathcal{M}$ such that $M \models \Psi$ for all $\Psi \in T$.

At this point we introduce the Desires-Intentions logic, DI for short, as the expansion of BD over the language $\mathcal{L}_{DI}$ with these additional axioms and rules for modal formulas:

1. Axioms for the δ_n's:
   $\delta_n\Phi \oplus \overset{n}{\ldots} \oplus \delta_n\Phi \equiv_L \Phi$
   $\delta_n\Phi \otimes (\delta_n\Phi \oplus \overset{n-1}{\ldots} \oplus \delta_n\Phi) \to_L \overline{0}$
2. (BD0) axiom for I_α modalities
   $I_\alpha(\varphi \vee \psi) \equiv_L I_\alpha\varphi \wedge_L I_\alpha\psi$
3. Definitional Axiom for I:
   $I\varphi \equiv_L \bigvee_{\alpha\in\Pi^0} I_\alpha\varphi$
4. Inference Rules:
   introduction of I and I_α for implications: from $\varphi \to \psi$ derive $I\psi \to_L I\varphi$ and $I_\alpha\psi \to_L I_\alpha\varphi$ for each $\alpha \in \Pi^0$

The notion of proof for DI, denoted $\vdash_{DI}$, is defined as always from the above axioms and inference rules. The presented axiomatics is obviously sound and one can prove completeness in an analogous way as for BD logic and we omit the proof.

Theorem 7 *Let T be a theory of closed modal formulas and Φ a closed modal formula. Then $T \vdash_{DI} \Phi$ iff $T \models_{\mathcal{M}_{DI}} \Phi$.*

So defined, the semantics and axioms for the I_α operators is very general and probably it is not evident how to capture the idea above expressed that the truth-degree of a formula $I_\alpha\varphi$ should take into account not only how much φ is desired but also other information as for instance how much α is costly. Of course, one can think of many possible ways to do this but, only as way of example, the possibly simplest way is to consider the value of $I_\alpha\varphi$ as the arithmetic mean of the value of $D^+\varphi$ with 1 minus the cost value of C_α. Indeed, consider the following expression:

$$I_\alpha\varphi \equiv_L \delta_2 D^+\varphi \oplus \delta_2\neg_L C_\alpha \qquad (I_\alpha\text{-Def})$$

Then one can easily show that (I_α-Def) is consistent in DI. In fact, this formula is valid in all DI-models $M = (W, e, \pi^+, \pi^-, \{\pi_\alpha\}_{\alpha\in\Pi^0})$ such that $\pi_\alpha(w, w') = (\pi^+(w') + 1 - e(w, C_\alpha))/2$ and in such models it holds that, for any $w \in W$, $e(w, I_\alpha\varphi) = e(w, \delta_2 D\varphi \oplus \delta_2\neg_L C_\alpha)$. Therefore the logic DI can capture such a notion of intention strength of reaching a goal φ through an action α as an average of the positive desire degree of φ and the cost of performing α.

Therefore this axiom (or similar ones leading to different definitions for the I_α operators) can be included in a specialized theory over DI to specify particular behaviors of the Intention modality. One can also specify some particular semantics for the action cost variables C_α. For instance, if we assume that the set Π^0 may contain compound actions expressed in a similar way as compound programs in Dynamic logic, one could consider the following natural axioms

(C1) $C_\gamma \equiv_L C_\alpha \vee C_\beta$, if $\gamma, \alpha, \beta \in \Pi^0$ and $\gamma = \alpha \cup \beta$

(C2) $C_\gamma \equiv_L C_\alpha \oplus C_\beta$, if $\gamma, \alpha, \beta \in \Pi^0$ and $\gamma = \alpha; \beta$

governing the costs of a nondeterministic union and of a concatenation of actions respectively. Axiom (C1) represents a kind of conservative attitude since it assigns to the nondeterministic action $\alpha \cup \beta$ the maximum of the costs of α and β. Axiom (C2) establishes the (bounded) sum of costs of α and β as the cost of the composed action $\alpha; \beta$. If we denote by DI_2 the extension of DI logic with these two axioms, then

[6] Actually for our purposes below we would only need to introduce δ_2.

one can easily prove some consequences for the behavior of the I_α's operators in the theory defined by the (I_α-Def) formulas:

Lemma 8
(i) If $\alpha, \beta, \alpha \cup \beta \in \Pi^0$, *then* $\{(I_\gamma{-}\mathrm{Def})\}_{\gamma\in\Pi^0} \vdash_{DI_2} I_{\alpha\cup\beta}\varphi \equiv_L I_\alpha\varphi \wedge I_\beta\varphi$
(ii) If $\alpha, \beta, \alpha;\beta \in \Pi^0$, *then* $\{(I_\gamma{-}\mathrm{Def})\}_{\gamma\in\Pi^0} \vdash_{DI_2} I_{\alpha;\beta}\varphi \rightarrow_L I_\alpha\varphi \wedge I_\beta\varphi$

Proof: (i) comes from the fact that in Rational Lukasiewicz logic[7] one can prove the following equivalences: $\neg_L(\Phi \vee \Psi) \equiv_L \neg_L\Phi \wedge \neg_L\Psi$, $\delta_n(\Phi \wedge \psi) \equiv_L \delta_n\Phi \wedge \delta_n\Psi$, and $\Gamma \oplus (\Phi \wedge \Psi) \equiv_L (\Gamma \wedge \Phi) \oplus (\Gamma \wedge \Psi)$. On the other hand (ii) is a consequence of the following implications provable in RLL: $(\Phi \rightarrow_L \Psi) \rightarrow_L (\delta_n\Phi \rightarrow_L \delta_n\Psi)$ and $(\Phi \rightarrow_L \Psi) \rightarrow_L (\Gamma \oplus \neg_L\Psi \rightarrow_L \Gamma \oplus \neg_L\Phi)$. □

Operational elements

Up to this point we have proposed an expressive logical framework to represent and reason about an agent's desires and intentions. But to evaluate the different factors involved, the agent needs to represent, besides desires and intentions, her beliefs about the world and in particular those related to the possible actions and the changes they cause in the environment where the agent is situated. Moreover, she needs a planner to look for feasible plans that from the current state of the world, permits the agent achieve her goals. Different processes are related to the agent's intentions selection that eventually lead to the action she undertakes.

The agent's belief representation and reasoning can be done, in a similar way we have done for desires and intentions, by introducing into the logical framework graded belief formulas $B\varphi$, where φ is a Boolean formula. Indeed, one can take Dynamic logic as the base Boolean language to define on top of it modal-like formulas like $(B[\alpha]\varphi, r)$ with the intended meaning that the belief (e.g. probability) degree of having φ after α's execution is at least r for some rational $r \in [01]$. Details on this kind of graded belief representation are out of the scope of this paper, see (Casali *et al.* 2005) for a preliminary formalization.

In order to make all the described logical ingredients operational in a deliberative agent architecture, they should be complemented with some functional elements which somehow go beyond a purely (flat) logical formalization. To fix ideas, we consider an intentional agent architecture that is based on the logical framework(s) defined in the previous sections.

We describe next its main components, which are shown in Figure 1 where boxes represent tasks, cylinders represent the logic structures or theories that support the different graded attitudes and plans, and where arrows illustrate the information flow and the sequences between the processes[8].

- *a set of current Beliefs* , representing the uncertain information the agent has about its environment;

[7]That is, the extension of Lukasiewicz logic with the δ_n's operators (Gerla 2001).

[8]Some of these processes are out of the scope of the paper but are included for the sake of completeness.

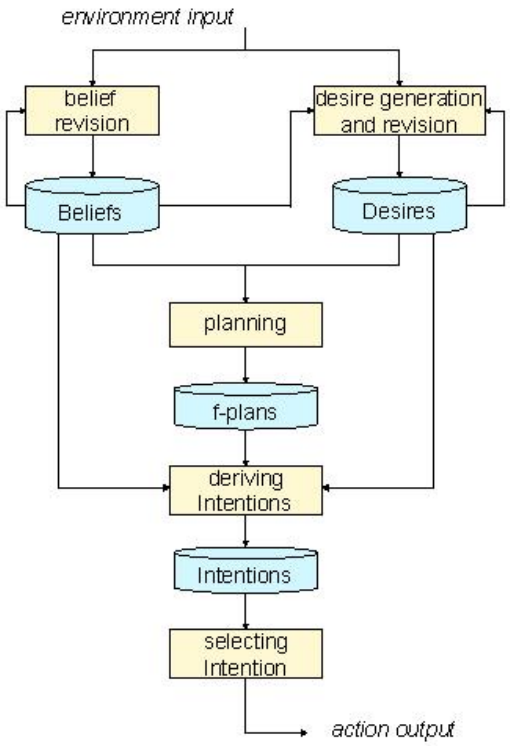

Figure 1: Agent architecture

- *a set of current Desires* , representing the graded positive and negative agent desires, making up the agent desire theory;
- *a belief revision process*, which takes new inputs and the agent's current beliefs and determines a new set of graded beliefs;
- *a desire generation and revision function*, which determines and/or revises the agent's graded positive and negative desires on the basis of her current beliefs and her previous desires. The revision process used depends on the logic schema selected for the agent desire representation;
- *a planning process*, which is in charge of first building plans from atomic actions (these possible actions are part of the agent beliefs) and then of looking for *feasible plans*. Feasible plans are plans which fulfill (to some degree) positive desires, satisfy some preconditions and avoid undesired postconditions. Filtering plans to identify which ones are feasible can be modeled in a logic style by an inference rule like this:

$$\frac{(D^+\varphi, r), (D^-\psi, s), plan(\alpha, \chi, A, c), (B([\alpha]\varphi), b_0), (B(A \rightarrow \neg\psi), 1), B(\chi, 1)}{fplan(\varphi, \alpha, \chi, A, c)} \quad (1)$$

which generates a predicate *fplan(φ, α, P, A, c_α)*, standing for α is plan that achieves φ with precondition χ, postcondition A and cost c, whenever: (1) it is believed

(above some threshold level b_0) that plan α leads to satisfy a positively desired goal φ (encoded as $(D^+\varphi, r)$ and $(B([\alpha]\varphi), b_0)$), (2) α's precondition χ is satisfied ($(B\chi, 1)$), and (3) α's postconditions A avoid negative desires ψ (encoded as $(B(A \rightarrow \neg\psi), 1)$ and $(D^-\psi, s)$).

- *a set of current feasible plans (fplans)*, representing those feasible plans the planner has found for the current situation (agent beliefs and desires).
- *a process for deriving intentions*, which for each feasible plan α that allows to achieve a goal φ, an intention formula $I_\alpha\varphi$ is derived with its corresponding degree. According to the notion adopted in axiom (I_α-Def), the intention degree is taken as a trade-off between the benefit of reaching the goal and the cost of the plan α. Axiom (I_α-Def), or a similar one, can be made operational by means of an inference rule like the following one which derives this value from the degree d of $D^+\varphi$ and the cost c of the plan α:

$$\frac{(D^+\varphi, d), fplan(\varphi, \alpha, \chi, A, c)}{(I_\alpha\varphi, f(d, c))} \quad (2)$$

  Different choices for the function f allow us to model different agent behaviors. For instance, if we consider an *equilibrated agent* where the factors involved are equally important, the function might be defined as the average among these factors, i.e. $f(d, c) = (d + (1 - c))/2$. This function is indeed the one modelled by axiom (I_α-Def).
- *a set of current Intentions*, representing those goals (together with their intention degrees) that the agent is committed to try to bring about by the execution of suitable plans.
- *an Intention - Action selection process*, which determines which action to perform on the basis of each selected intention. To look for the best plan α to achieve a goal φ. From the set of current intentions and feasible plans, the agent selects for a given goal φ the plan α which leads to a maximum intention degree for $I_\alpha\varphi$, represented by the degree of the formula $I\varphi$ (see the definitional axiom for I)

Example 3 *(Example 1 continued).*

The recommender agent takes all desires expressed by María, our stressed tourist, and follows the steps explained in this section:

- ***Beliefs:*** *the agent updates her current beliefs about the tourism plans offered, the tourism domain (structured using destinations ontologies) and the beliefs about how these packages can satisfy the user's preferences.*
- ***Desire generation:*** *exactly what was generated in Example 1 above, i.e.* $T'_D = \{(D^+m, 0.8), (D^+r, 0.6), (D^+(m \wedge r), 0.95), (D^-f, 0.7)\}$.
- ***Planning, to look for feasible packages:*** *from this set of positive and negative desires (T'_D) and knowledge about the tourist packages the agent can offer and the benefits they bring, and using a Planner, the agent looks for feasible plans, that are believed to achieve positive desires (i.e in this case* m*,* r *or* $m \wedge r$*) by their execution but avoiding the negative desire (i.e.* f*) as post-condition (see rule (1)).*
- ***Current feasible packages:*** *the agent finds that the plans* Mendoza (Me) *and* SanRafael (Sr) *are feasible plans for the combined goal* $m \wedge r$*, while* Cumbrecita (Cu) *is feasible only for* m*. The Planner also computes the normalized cost (*$c \in [0, 1]$*) of these plans being respectively:* $c_{Me} = 0.60$ *and* $c_{Sr} = 0.70$ *and* $c_{Cu} = 0.55$.
- ***Deriving the Intention formulae*** $I_\alpha\varphi$***, for each feasible plan*** α ***toward a desire*** φ***.*** *The intention degrees for satisfying each desire* m*,* r *and* $m \wedge r$ *by the different feasible plans are computed by a rule that trades off the cost and benefit of satisfying a desire by following a plan. The Agent uses rule (2) and the function* $f(d, c) = (d + (1 - c))/2$*, where* d *is the desire degree and* c *is the normalized cost of the plan to computes the intentions degrees toward* $m \wedge r$*,* m *and* r *by executing the feasible plans* Mendoza (Me) *and* SanRafael (Sr)*.*
- ***Current Intentions:*** *as a result of the previous process, the set of intentions contains the following formulas:*
  $(I_{Me}(m \wedge r), 0.675)$, $(I_{Sr}(m \wedge r), 0.625)$,
  $(I_{Me}(m), 0.60)$, $(I_{Me}(r), 0.50)$,
  $(I_{Sr}(m), 0.55)$, $(I_{Sr}(r), 0.45)$,
  $(I_{Cu}(m), 0.625)$.
- ***Selecting Intention-plan:*** *the agent decides the tourist's recommendation. From this set of current intentions, the Agent decides to recommend the plan* $Mendoza\ (Me)$ *since it brings the best cost/benefit relation (represented by the intention degree 0.675) to achieve* $m \wedge r$*, satisfying also the tourist's negative desire.*

Related Work and Discussion

In this work we have formalized a logical framework to represent and reason about graded agent desires and intentions. Inspired in (Benferhat *et al.* 2002b; 2006), a bipolar representation of preferences has been used to represent agent desires, since we deem important to separately represent positive and negative desires. This is related but somewhat different from the logic defined in (Lang *et al.* 2002) to represent (conditional) desires with a semantics based on utility losses and gains. These utilities are added up to a single measure which, together with domain knowledge, induces a preference relation over worlds. As for intentions, we have followed (Cohen and Levesque 1990) and proposed a model where intentions depend not only on the benefit of reaching a goal (that is modelled as a positively desired formula), but also on the state of the current world and the cost of transforming it into a world where the goal is satisfied. Our graded representation of intentions (see (Casali *et al.* 2005) for preliminary ideas within a multicontext agent architecture) allows us to provide as intended semantics a measure

of the cost/benefit relation involved in the agent actions toward the intended goal. A similar semantics for intentions is used in (Schut *et al.* 2001), where the net value of an intention is defined as the difference between the value of the intention outcome and the cost of the intention and in (Rao and Georgeff 1991), this relation is encoded in the payoff function over the different paths.

Bipolar preferences can be used for decision making or problem solving. For instance, in (Benferhat *et al.* 2006) and in (Bistarelli *et al.* 2007), they show how to combine bipolar preferences and other types of information (i.e. domain restrictions) to find optimal solutions in a logical framework and in soft constraint problems respectively. In this paper, we have provided some operational rules showing how the agent desires, together with her beliefs about the environment, plans and the transformation they produce, may be used in deciding the agent intention and the best plan to follow, mimicking the role of a suitable set of bridge rules (i.e. rules relating formulas from different contexts) used in a multicontext specification of agency in (Casali *et al.* 2005). Recently in (Rahwan and Amgoud 2006) and in (Rotstein *et al.* 2007) the problem of desire, intention and plan generation in BDI agents is approached in argumentation frameworks. In these works rules to generate desires are also included and a revision process for the different agent attitudes based on argumentation is presented. We consider important as a future line of work, to include in our logical framework a revision process for desires and intentions in order to keep these attitudes consistent for agents living in dynamic environments.

Acknowledgments

The authors thank the anonymous referees and Pilar Dellunde for their interesting comments that have helped a lot to improve this paper. Godo and Sierra also acknowledge partial financial support by the Spanish projects MULOG2 (TIN2007-68005-C04) and *Agreement Technologies* (CONSOLIDER, CSD2007-0022).

Annex I: About Rational Pavelka Logic

Rational Pavelka's Logic RPL is an extension of Łukasiewicz's infinitely-valued logic admitting to explicitly reason about partial degrees of truth. Introduced by Pavelka in the late seventies, it is described in a simple formalization in (Hájek 1998). Since the approach described in this paper strongly relies on this logic, here we follow the latter and present its main notions and properties.

Formulas are built from propositional variables $p_1, p_2, \ldots$ and truth constants $\overline{r}$ for each *rational* $r \in [0,1]$ using connectives $\to_L$ and $\neg_L$. Other connectives can defined from these ones. In particular, among others, one can define two conjunctions and two disjunctions exactly as in Łukasiewicz's logic, i.e.

$$\begin{array}{lll} \varphi \otimes \psi & \text{stands for} & \neg_L(\varphi \to_L \neg_L \psi) \\ \varphi \oplus \psi & \text{stands for} & \neg_L\varphi \to_L \psi \\ \varphi \vee_L \psi & \text{stands for} & (\varphi \to_L \psi) \to_L \psi \\ \varphi \wedge_L \psi & \text{stands for} & \neg_L(\neg_L\varphi \vee_L \neg_L\psi) \\ \varphi \equiv_L \psi & \text{stands for} & (\varphi \to_L \psi) \wedge_L (\psi \to_L \varphi) \end{array}$$

Łukasiewicz's truth functions for the connectives $\to_L$ and $\neg_L$ are (we use the same symbol than for the connectives):

$$\begin{array}{rcl} x \to_L y & = & \min(1, 1-x+y) \\ \neg_L x & = & 1-x \end{array}$$

Taking into account this definitions, it is easy to check that the truth functions for the above definable connectives are the following ones:

$$\begin{array}{rcl} x \otimes y & = & \max(0, x+y-1) \\ x \oplus y & = & \min(x+y, 1) \\ x \vee_L y & = & \max(x, y) \\ x \wedge_L y & = & \min(x, y) \\ x \equiv_L y & = & 1-|x-y| \end{array}$$

An *evaluation* e is a mapping of propositional variables into $[0,1]$. Such a mapping uniquely extends to an evaluation of all formulas respecting the above truth functions and further assuming that $e(\overline{r}) = r$ for each rational $r \in [0,1]$. An evaluation is a model of a set of formulas T whenever $e(\varphi) = 1$ for all $\varphi \in T$. We write $T \models RPL\varphi$ to denote that $e(\varphi) = 1$ for every evaluation e model of T.

Logical axioms of RPL are:

(i) axioms of Łukasiewicz's logic

$$\begin{array}{l} \varphi \to_L (\psi \to_L \varphi) \\ (\varphi \to_L \psi) \to_L ((\psi \to_L \chi) \to_L (\varphi \to_L \chi)) \\ (\neg_L\varphi \to_L \neg_L\psi) \to_L (\psi \to_L \varphi) \\ ((\varphi \to_L \psi) \to_L \psi) \to_L ((\psi \to_L \varphi) \to_L \varphi) \end{array}$$

(ii) bookkeeping axioms: (for arbitrary rational $r, s \in [0,1]$):

$$\begin{array}{l} \neg_L\overline{r} \equiv_L \overline{1-r} \\ \overline{r} \to_L \overline{s} \equiv_L \overline{\min(1, 1-r+s)} \end{array}$$

The only *deduction rule* is modus ponens for $\to_L$. The notion of proof in RPL, denoted $\vdash_{RPL}$, is defined as usual from the above axioms and rule.

RPL enjoys two kinds of completeness. The so-called Pavelka completeness reads as follows. Let T be an arbitrary set of formulas (theory) and φ a formula. The *provability degree* of φ in T is defined as $|\varphi|_T = \sup\{r \mid T \vdash_{RPL} \overline{r} \to_L \varphi\}$. The *truth degree* of φ in T is defined as $\|\varphi\|_T = \inf\{e(\varphi) \mid e \text{ evaluation}, e \text{ model of } T\}$. Notice that both $\|\varphi\|_T$ and $|\varphi|_T$ may be irrational.

Pavelka-style completeness theorem for RPL: For each T and φ,

$$|\varphi|_T = \|\varphi\|_T.$$

i.e. the provability degree equals to the truth degree.

Besides, RPL enjoys a classical completeness property but only for deductions from finite theories, which will be used in the proof in Annex II.

Finite strong completeness theorem for RPL: For each *finite* T and φ, $T \vdash_{RPL} \varphi$ iff $T \models_{RPL} \varphi$.

Annex II: proof of Theorem 3

Theorem 3 (completeness) *Let T be a finite theory of closed formulas and Φ a closed formula. Then $T \models_{\mathcal{M}_{\mathcal{BD}}} \Phi$ iff $T \vdash_{BD} \Phi$.*

Proof: We basically follow the proof of Theorem 8.4.9 in (Hájek 1998), with some adaptations.

Assume $p_1, \ldots, p_n$ contain at least all the propositional variables involved in T and Φ, and let Nor be the set of 2^{2^n} non logically equivalent Boolean formulas in DNF built from the p_i's. For each non-modal φ built from the p_i's, let $\varphi_{NF} \in Nor$ denote its corresponding normal form. Then for each modal Φ let us denote by Φ_{NF} the result of replacing each atomic modal component of the form $D^+\varphi$ or $D^-\varphi$ by $D^+\varphi_{NF}$ or $D^-\varphi_{NF}$ respectively. Finally, for each modal theory S let us denote by S_{NF} the result of replacing each $\Phi \in S$ by Φ_{NF}.

The idea is that the modal theory T can be represented as a (finite) theory over the propositional logic RPL. For each modal formula $D^+\varphi$ introduce a propositional variable p^+_ϕ, and for each $D^-\varphi$ another propositional variable p^-_φ. Then define a mapping $*$ from closed modal formulas to RPL formulas as follows:

- $(D^+\varphi)^* = p^+_\varphi$,
- $(D^-\varphi)^* = p^-_\varphi$,
- $(\overline{r})^* = \overline{r}$, for each rational $r \in [0,1]$,
- $(\Phi \&_L \Psi)^* = \Phi^* \&_L \Psi^*$
- $(\Phi \rightarrow_L \Psi)^* = \Phi^* \rightarrow_L \Psi^*$.

If S is a set of modal formulas, let $S^* = \{\Phi^* \mid \Phi \in S\}$.

Now, let $\mathcal{BD} = \{\Phi \mid \Phi$ instance of modal axioms (BD3$^+$) and (BD3$^-$)$\}\cup\{D^+\varphi \rightarrow_L D^+\psi, D^-\varphi \rightarrow_L D^-\psi \mid \varphi \rightarrow \psi$ theorem of CPC $\}$. We next show that the following statements are equivalent:

1) $T \models_{BD} \Phi$
2) $T^* \cup \mathcal{BD}^* \models_{RPL} \Phi^*$
3) $T^*_{NF} \cup (\mathcal{BD}_{NF})^* \models_{RPL} \Phi^*_{NF}$
4) $T^*_{NF} \cup (\mathcal{BD}_{NF})^* \vdash_{RPL} \Phi^*_{NF}$
5) $T^* \cup \mathcal{BD}^* \vdash_{RPL} \Phi^*$
6) $T \vdash_{BD} \Phi$

- (1 $\Rightarrow$ 2): Let us assume $T^* \cup \mathcal{BD}^* \not\models_{RPL} \Phi^*$. This means there is an RPL-evaluation v model of $T^* \cup \mathcal{BD}^*$ and $v(\Phi^*) < 1$. We build then a model $M_v = \langle W, e, \pi^+, \pi^- \rangle$ as follows:

  - W is the set of Boolean evaluations of the propositional variables $q_1, ..., q_n$;
  - $e(w, q) = w(q)$, for each propositional variable q and $e(w, \cdot)$ is extended to Boolean formulas as usual;
  - $e(w, \overline{r}) = r$ for each rational $r \in [0,1]$;
  - $e(w, D^+\varphi) = v(p^+_\varphi)$ and $e(w, D^-\varphi) = v(p^-_\varphi)$, and $e(w, \cdot)$ is extended to compound modal formulas using RPL connectives;
  - $\pi^+(w) = v(p^+_{A_w})$ and $\pi^-(w) = v(p^-_{A_w})$, where A_w is the elementary conjunction built with literals from the propositional variables $q_1, \ldots, q_n$ such that $e(w, A_w) = 1$ and $e(w', A_w) = 0$ if $w' \neq w$;

  Since $M_v \models \varphi \equiv \vee_{w \in W} A_w$, it is easy to check that, so defined, $e(w, D^+\varphi) = \inf\{\pi^+(w') \mid e(w', \varphi) = 1\}$ and $e(w, D^-\varphi) = \inf\{\pi^-(w') \mid e(w', \varphi) = 1\}$. Therefore M_v is BD model, and since by construction $e(w, \Psi) = v(\Psi^*)$ for all modal formula Ψ and world $w \in W$, we also have in particular $e(w, \Psi) = v(\Psi^*) = 1$ for all $\Psi \in T^*$ and $e(w, \Phi) = v(\Phi^*) < 1$ and hence $T \not\models_{BD} \Phi$.
- (2 $\Rightarrow$ 3): Assume e is a RPL-evaluation of the propositional variables $p_{\varphi_{NF}}$ which is a model of $T^*_{NF} \cup (\mathcal{BD}_{NF})^*$ but $e(\Phi^*_{NF}) < 1$. Then extend e to propositional variables p^+_φ and p^-_φ by putting $e'(p^+_\varphi) = e(p^+_{\varphi_{NF}})$ and $e'(p^-_\varphi) = e(p^-_{\varphi_{NF}})$. It is easy to check that so defined e' is such that $e'(\Phi^*) = e((\Phi_{NF})^*)$ for any modal formula Φ, and hence e' is a model of $T^* \cup \mathcal{BD}^*$ and $e'(\Phi^*) = e(\Phi^*_{NF}) < 1$.
- (3 $\Rightarrow$ 4): Since $T^*_{NF} \cup (\mathcal{BD}_{NF})^*$ is a finite theory (recall that there are finitely-many formulas in Nor), then 4) follows from 3) by the finite strong standard completeness of RPL.
- (4 $\Rightarrow$ 5): Easy since, using Lemma 1, if $\vdash \varphi \equiv \psi$ (in classical propositional logic) then $T^* \cup \mathcal{BD}^*$ proves in RPL both $p^+_\varphi \equiv_L p^+_\psi$ and $p^-_\varphi \equiv_L p^-_\psi$, and hence $T^* \cup \mathcal{BD}^* \vdash_{RPL} \Phi^* \equiv_L (\Phi_{NF})^*$ for each modal formula Φ.
- (5 $\Rightarrow$ 6): Let $\Psi^*_1, \ldots, \Psi^*_n$ be a BD-proof of Φ^* from $T^* \cup \mathcal{BD}^*$. This is converted into a BD-proof of Φ from T by adding for each Φ^*_i which is of the form $p^+_\varphi \rightarrow_L p^+_\psi$ (resp. $p^-_\varphi \rightarrow_L p^-_\psi$) with $\varphi \rightarrow \psi$ being a theorem of CPC, a proof of $\varphi \rightarrow \psi$ in CPC and then applying the rule of introduction of D^+ (resp. D^-) for implications.
- (6 $\Rightarrow$ 1): This is soundness.

□

References

Benferhat S., Dubois D., Kaci S. and Prade, H. Bipolar Possibilistic Representations. *Proceedings of UAI 2002*: pages 45-52. Morgan Kaufmann, 2002.

Benferhat S., Dubois D., Kaci S. and Prade, H. Bipolar representation and fusion of preferences in the possilistic Logic framework. *In Proceedings of KR-2002*, 421-448, 2002.

Benferhat, S., Dubois, D., Kaci, S. and Prade, H., Bipolar possibility theory in preference modeling: Representation, fusion and optimal solutions. *Information Fusion*, Elsevier, 7, 135-150, 2006.

Bistarelli S., Pini M.S., Rossi F. and K. Brent Venable, Bipolar preference problems: framework, properties and solving techniques, Springer LNAI 4651, 78-92, 2007.

Casali A., Godo L. and Sierra C. Graded BDI Models For Agent Architectures. Leite J. and Torroni P. (Eds.) *CLIMA V, LNAI 3487*, pp. 126-143, Springer-Verlag, 2005.

Cohen, P. R. and Levesque, H. J. Intention is choice with commitment. *Artificial Intelligence*, 42:213-261, 1990.

Dennet, D. C. *The Intentional Stance*. MIT Press, Cambridge, MA, 1987.

Gerla B., Rational Łukasiewicz logic and DMV-algebras. *Neural Networks World*, vol 11 (2001), 579-584.

Hájek P., Metamatematics of Fuzzy Logic, *Trends in Logic*, 4, Kluwer Academic Publishers.

Lang J., van der Torre, L. and Weydert E. Utilitarian Desires. *Autonomous Agents and Multi-Agent Systems* 5(3): 329-363 (2002).

Rahwan I. and Amgoud L., An argumentation based approach for practical reasoning, in *Proceedings of AAMAS '06*, 347–354, Hakodate, Japan, ACM, NY, USA, 2006.

Rao, A. And Georgeff M. Modeling Rational Agents within a BDI-Architecture. *In proceedings of KR-92*, pp. 473-484 (ed R. Fikes and E. Sandewall), Morgan Kaufmann, San Mateo, CA, 1991.

Rao, A. And Georgeff M. Deliberation and Intentions. In *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, 1991.

Rotstein N., García A. and Simari G. Reasoning from Desires to Intentions: A Dialectical Framework. *In Proceedings AAAI-07*. Vancouver, BC, Canada. 136-141. 2007.

Schut, M., Wooldridge, M. and Parsons S. Reasoning About Intentions in Uncertain Domains in *6th ECSQARU 2001, Proceedings*, pp. 84-95, Toulouse, France, 2001.