A graded BDI agent model to represent and reason about preferences

Ana Casali[a], Lluís Godo[b], Carles Sierra[b]

[a] *Facultad de Cs. Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario*
Centro Internacional Franco Argentino de Ciencias de la Información y de Sistemas (CIFASIS)
Av. Pellegrini 250, 2000 Rosario, Argentine
[b] *Artificial Intelligence Research Institute (IIIA - CSIC)*
Campus de la Universitat Autònoma de Barcelona s/n, 08193 Bellaterra, Spain

Abstract

In this research note, we introduce a graded BDI agent development framework, *g-BDI* for short, that allows to build agents as multi-context systems that reason about three fundamental and *graded* mental attitudes (i.e. beliefs, desires and intentions). We propose a sound and complete logical framework for them and some logical extensions to accommodate slightly different views on desires.

Key words: BDI agents, multi-context systems, uncertainty, bipolar preferences, fuzzy logic.

1. Introduction

Consider the following scenario: *María, who lives in busy Buenos Aires, wants to relax for a few days in an Argentinian beautiful destination. She would be rather happy of practicing rafting and very happy going to a mountain place. She would like more to go climbing than to go trekking. On top of this, she is stressed and would like to get to the destination with a short trip.* Taking into account her preferences and constraints, the task of María's personal agent is to get, using domain knowledge, an adequate tourist package satisfying her preferences.

Preferences are the proactive attitude of intentional agents, the motor that make the agent act, building suitable plans that try to satisfy the most preferred goals, while satisfying a given set of constraints. Constraints are also a key modeling aspect that account for restrictions or rejections over the possible states the agent can reach.

In BDI agent architectures [14, 21, 23], desires represent the *ideal* agent preferences regardless of the current agent perception of the environment and regardless of the cost involved in actually achieving them. We consider important for an agent to distinguish what is positively desired from what is not rejected. By doing so, positive desires then represent what the agent would like to be the case while negative desires will correspond to what the agent rejects or does not want to occur. Furthermore, if the agent needs to represent different levels of preference or rejection, the notions of positive and negative desires become naturally graded.

To have a powerful and flexible representation of an agent preferences is thus a fundamental issue to be addressed in any agent model. With this aim, in this work we present a general framework to define graded BDI agent architectures (*g-BDI* agents for short), based on multi-context systems [21]. In particular, we introduce a graded logical framework (i.e. languages, axioms and inference rules) to represent and reason, not only about the agent positive and negative desires, but also about other mental attitudes as beliefs and intentions. Namely, in a g-BDI agent, belief degrees represent to what extent the agent believes a formula is true, degrees of positive or negative desire allow the agent to set different levels of preference or rejection respectively and intention degrees represent also a preference level but, in this case, modeling the cost/benefit trade-off of reaching a goal.

Email addresses: acasali@fceia.unr.edu.ar (Ana Casali), godo@iiia.csic.es (Lluís Godo), sierra@iiia.csic.es (Carles Sierra)

Computational aspects of this agent framework are out of the scope of this research note due to lack of space. First results on giving operational semantics to g-BDI agents, using process calculus have been presented in [7]. As a case study of our g-BDI agent framework we have designed and implemented a prototype of a Tourism Recommender agent (*T-Agent*) [10] and have obtained encouraging initial experimental results [9]. This paper builds on authors' previous work [8] where the logical representation of graded preferences and intentions was expressed in a unique "flat" logical framework. Instead, here we present a multi-context agent framework where a different context is used for each mental attitude.

The structure of this work is as follows. In Section 2, we first present the basic ingredients of the g-BDI agent model with its multi-context specification, and in the next sections the different components of the agent model are defined. In Section 3 we formalize the desire context to represent the agent's positive and negative preferences, inspired in the bipolar representation of preferences in the framework of possibilisitic logic proposed in [2, 3]. Next, in Section 4 we formalize the intention context and in Section 5, we briefly outline the belief context to represent the agent uncertain beliefs. The necessary functional contexts for planning and communication, and the bridge rules to transfer formulae between theories for an illustrative agent model are briefly described in Section 6. Finally, a very short technical appendix with main facts of Rational Pavelka logic, that is used throughout the paper, has been included.

2. The Graded BDI Agent Model

Several previous works have proposed agent theories and architectures to provide multiagent systems with a strong formal basis. Among them, one of the most widely recognized is the BDI agent architecture presented by Rao and Georgeff [23]. We consider that an extension of this architecture to incorporate *degrees* in the different attitudes would not only make the model semantics richer, but it would also help the agents in taking better decisions. With that aim we have revisited the classical BDI agent architecture to represent and reason under uncertain beliefs and graded motivations. In this section we introduce the basic ingredients of a general model for graded BDI agents (g-BDI). The g-BDI model we consider extends the multi-context specification of agents proposed by Parsons et al. in [21] with the ability to represent graded mental attitudes.

2.1. Multi-context specification

Multi-context systems (MCS) were introduced in [15] to allow different formal (logic) components to be defined separately and then interrelated. A MCS specification contains three basic components: units or contexts, logics, and bridge rules, which channel the propagation of formulae among theories. Thus, an agent programmed as a MCS is defined as a group of interconnected units: $\langle \{C_i\}_{i\in I}, \Delta_{br} \rangle$, where each context $C_i \in \{C_i\}_{i\in I}$ defines a logic specified by a tuple $C_i = \langle L_i, A_i, \Delta_i \rangle$ with L_i, A_i and Δ_i being the language, axioms, and inference rules respectively. Δ_{br} is a set of bridge rules, that can be understood as rules of inference with premises and conclusions in different contexts. For instance:

$$\frac{C_1 : \psi, C_2 : \varphi}{C_3 : \theta}$$

means that if formula ψ is deduced in context C_1 and formula φ is deduced in context C_2 then the theory of context C_3 is extended with formula θ. When a theory $T_i \subset L_i$ is associated with each unit, the specification of a particular agent is complete. The deduction mechanism of these systems is based on two kinds of inference rules, internal rules Δ_i inside each unit, and bridge rules Δ_{br} outside. Internal rules allow to draw consequences within a theory, while bridge rules allow to relate the results within one or several theories with extensions of another theory. Any reasonable implementation of a multi-context system needs some kind of control strategy to synchronize and co-ordinate both kinds of inferences (i..e. by internal rules and by bridge rules).

In the running illustrative architecture we have *mental* contexts to represent beliefs (BC), desires (DC) and intentions (IC). We also consider two *functional* contexts, for planning (PC)

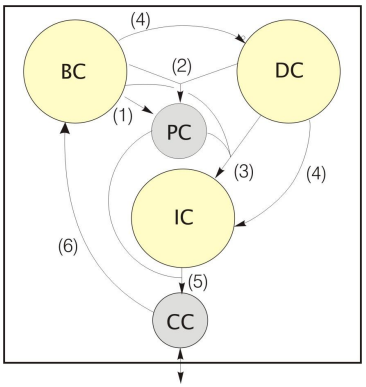

Figure 1: A multi-context architecture of a graded BDI agent

and communication (CC). The planner context is in charge of finding plans to change the current world into another world, where some desire is satisfied, and of computing the cost associated to the plans. The communication context is the agent door to the external world, receiving and sending messages. In summary, the illustrative g-BDI agent architecture is thus defined as:

$$A_g = (\{BC, DC, IC, PC, CC\}, \Delta_{br})$$

We could certainly program agents with several belief contexts for instance. The architecture is a minimum skeleton containing the three modalities. The different contexts of the architecture will be described in some detail in the following sections. In Figure 1 we present a schema of the architecture used to illustrate the g-BDI model with the mentioned set of mental contexts (BC, DC and IC) and functional ones (PC and CC) and some bridge rules ((1) to (6)) interrelating them. The contexts and concrete architecture we present will thus serve as a blueprint to design different kinds of agents.

2.2. Logical Framework: a many-valued modal approach

During the last two decades, the Artificial Intelligence community has undertaken the problem of knowledge representation and reasoning under uncertain and incomplete knowledge, see e.g. [17]. Among the approaches proposed in the literature, Hájek and colleagues have developed an alternative approach (see e.g. [16]) where uncertainty reasoning is formalized as suitable modal theories over suitable $[0, 1]$-valued fuzzy logics. The basic idea of this approach is to consider *the belief degree of a (classical) proposition as the truth-degree of a fuzzy modal proposition.* For instance, in the case where belief degrees are modelled as probabilities, for each classical (two-valued) formula φ, they consider a graded modal formula $B\varphi$ which is interpreted as "φ is probable", whose truth-degree can be set as the probability of φ. Moreover, using Łukasiewicz logic one can express the governing axioms of probability theory as logical axioms involving modal formulae of the kind $B\varphi$. In this way, many-valued models of these modal axioms faithfully correspond to probability measures over classical (non-modal) formulae. This approach can be also applied to other graded mental attitudes, such as desires or intentions, as it will be done in the next two sections. [1]

3. Desire Context (DC)

The Desire context is in charge of dealing with the agent's desires. We define next a many-valued modal logic to represent and reason about the agent bipolar preferences, i.e. a language, a

[1]Indeed, for simplicity reasons, we will use in this research note the same underlying fuzzy logic for beliefs, desires and intentions, the so-called Rational Pavelka logic (RPL), an expansion of Łukasiewicz logic with rational truth-constants [16], which is briefly described in Appendix I.

semantics and a set of axioms. Adopting the semantics based on guaranteed possibility measures, a precise meaning of *desire degrees* is given.

As desires are ideal preferences, we consider that it may be somewhat controversial, and domain dependent, to set (normative) general restrictions about positive (respectively negative) desires both on a formula and its negation, and also between the positive and negative desires on a given formula. Hence, we will present a basic axiomatics for bipolar desires representation, and then, we will consider several axiomatic extensions to cope with some meaningful additional constraints.

3.1. DC Language

The language $\mathcal{L}_{DC}$ is defined over a classical propositional language $\mathcal{L}$ (built from a countable set of propositional variables *Var* with connectives $\rightarrow$ and $\neg$), which is expanded with two (fuzzy) modal operators D^+ and D^-. $D^+\varphi$ reads as "φ is positively desired" and its truth degree represents the agent's level of satisfaction would φ become true. $D^-\varphi$ reads as "φ is negatively desired" (or "φ is rejected") and its truth degree represents the agent's level of disgust on φ becoming true. We will use a fuzzy modal logic to formalize graded desires and we select Rational Pavelka logic as the underlying fuzzy logic (see Appendix I). More precisely, formulae of the expanded language $\mathcal{L}_{DC}$ are defined as follows, where $\mathrm{Sat}(\mathcal{L})$ denotes the set of satisfiable formulae of $\mathcal{L}$:

- If $\varphi \in \mathcal{L}$ then $\varphi \in \mathcal{L}_{DC}$
- If $\varphi \in Sat(\mathcal{L})$ then $D^-\varphi, D^+\varphi \in \mathcal{L}_{DC}$ [2]
- If $r \in Q \cap [0,1]$ then $\overline{r} \in \mathcal{L}_{DC}$
- If $\Phi, \Psi \in \mathcal{L}_{DC}$ then $\Phi \rightarrow_L \Psi \in \mathcal{L}_{DC}$ and $\neg_L\Phi \in \mathcal{L}_{DC}$ (other Łukasiewicz logic connectives, like $\otimes, \oplus, \wedge_L, \vee_L, \equiv_L$ are definable from $\neg_L$ and $\rightarrow_L$)

We will call a modal formula *closed*, or D-formula, 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 $\overline{r} \rightarrow_L D^+\psi$, and reads as: the level of positive desire of ψ is at least r. Analogously for $(D^-\psi, r)$ and $\overline{r} \rightarrow_L D^-\psi$.

In this context, the agent's preferences will be expressed as a theory $\mathcal{T}_\mathcal{D}$ (a set of D-formulae) containing 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 $(D^+\phi_i, 1) \in \mathcal{T}_\mathcal{D}$ means that the agent has maximum preference in ϕ_i and is fully satisfied if it is true. While $(D^+\phi_j, \alpha) \notin \mathcal{T}_\mathcal{D}$ for any $\alpha > 0$ means that the agent is indifferent to ϕ_j and the agent does not benefit from ϕ_j becoming true. Analogously, $(D^-\psi_i, 1) \in \mathcal{T}_\mathcal{D}$ means that the agent absolutely rejects ψ_i and thus the states where ψ_i is true are totally unacceptable. If $(D^-\psi_j, \beta) \notin \mathcal{T}_\mathcal{D}$ for any $\beta > 0$ it simply means that ψ_j is not rejected.

3.2. DC Semantics

Some people would argue that if we consider desires as a proactive attitude, reasoning about desires on *disjunctions* of formulae is not very intuitive. In most cases an agent may have plans to achieve φ or ψ individually, or to achieve both $(\varphi \wedge \psi)$ but not for achieving one of them, that is $\varphi \vee \psi$. But in some cases, expressing desires for disjunctions may lead to more succinct specifications. According to the semantics presented in [2], the degree of positive desire for (or level of satisfaction with) a disjunction of desires $\varphi \vee \psi$ 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 desires. In contrast, the satisfaction degree of reaching both φ and ϕ can be strictly

[2] We define the modal formulae excluding the possibility of having positive and negative desires on a contradiction, $\bot \notin Sat(\mathcal{L})$.

greater than reaching one of them separately. These are basically the properties of the *guaranteed possibility* measures (see e.g. [2]). Analogously 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 DC models are Kripke structures $M = \langle W, e, \pi^+, \pi^- \rangle$ where W and e are defined as usual and π^+ and π^- are preference distributions over worlds, which are used to give semantics to positive and negative desires:

- $\pi^+ : W \to [0,1]$ 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 for non-modal formulae $e : \mathcal{L} \times W \to \{0,1\}$ is defined in the usual (classical) way. It is extended to atomic modal formulae $D^-\varphi$ and $D^+\varphi$ by:

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

together with the assumption that $\inf \emptyset = 1$. This is extended to compound modal formulae by means of the usual truth-functions of Łukasiewicz connectives.[3]

Notice that the evaluation $e(w, \Phi)$ of a modal formula Φ only depends on the formula itself and not on the actual world $w \in W$ where the agent is situated, so we will also write $e_M(\Phi)$ for $e(w, \Phi)$. This is consistent with the intuition that desires represent ideal preferences, 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 $w \in W$. Let $\mathcal{M}_{\mathcal{DC}}$ be the class of all Kripke structures $M = \langle W, e, \pi^+, \pi^- \rangle$. Then, for each subclass of models $\mathcal{M} \subseteq \mathcal{M}_{\mathcal{DC}}$, given a theory $\mathcal{T}$ and a formula Φ, we will write $\mathcal{T} \models_{\mathcal{M}} \Phi$ if $M \models \Phi$ for each model $M \in \mathcal{M}$ such that $M \models \Psi$ for all $\Psi \in \mathcal{T}$.

3.3. DC Axioms and Rules

To axiomatize the above 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 behaviour of the modal operators D^+ and D^- are needed. The following are the axioms and rules of the DC logic:

Axioms:

(CPC) Axioms of classical logic for non-modal formulae
(RPL) Axioms of Rational Pavelka logic for modal formulae
(DC0$^+$) $D^+(\varphi \vee \psi) \equiv_L D^+\varphi \wedge_L D^+\psi$ [4]
(DC0$^-$) $D^-(\varphi \vee \psi) \equiv_L D^-\varphi \wedge_L D^-\psi$

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 $D^-\psi \to_L D^-\varphi$.

[3] Standard truth functions on $[0,1]$ of primitive Łukasiewicz connectives are as follows: $x \to_L y = \min(1, 1 - x + y), \neg_L x = 1 - x$. Truth functions of main derived connectives are: $x \otimes y = \max(0, x + y - 1), x \oplus y = \min(1, x + y), x \wedge_L y = \min(x, y), x \vee_L y = \max(x, y), x \equiv_L y = 1 - |x - y|$.

[4] Notice that $\wedge_L$ is interpreted by the minimum, namely $e(\Phi \wedge_L \Psi, w) = \min(e(\Phi, w), e(\Psi, w))$.

The formalization we present for D^- is somewhat different from the approach presented by Benferhat et al. in [3], where they use a necessity function, i.e. they consider $D^-\phi$ as $N(\neg\phi)$. Nonetheless, their axiomatic approach is equivalent to ours, since the axiom (DC0$^-$) corresponds to the necessity axiom $N(\varphi \wedge \psi) \equiv N(\varphi) \wedge_L N(\psi)$. The introduction rules for D^+ and D^- state that the degree of desire is monotonically decreasing with respect to logical implication. A straightforward consequence of these rules is that degrees of desire preserve Boolean logical equivalence, i.e. if φ and ψ are classically equivalent, $D^+\varphi$ and $D^+\psi$, as well as $D^-\varphi$ and $D^-\psi$, are many-valued equivalent.

The notion of proof, denoted by $\vdash_{DC}$, is defined as usual from the above axioms and inference rules. It is a matter of routine to check that the axioms are valid in each DC-model and that the inference rules preserve validity in each DC-model, hence the above axiomatization is sound with respect to the defined semantics. Moreover, the basic DC logic is complete for finite theories of closed (modal) formulae.

Theorem 1 (completeness). *Let $\mathcal{T}$ be a finite theory of modal formulae and Φ a modal formula. Then $\mathcal{T} \models_{\mathcal{M}_{DC}} \Phi$ iff $\mathcal{T} \vdash_{DC} \Phi$.*

Proof: We basically follow the type of proof of [16, Th 8.4.9] with some adaptations, for details the reader is referred to [8]. □

The basic logical schema DC puts almost no constraint on the strengths for the positive and negative desires 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 supported by different arguments.[5]

Example 1. *Recall the scenario described in Section 1. María activates a personal agent, based on our g-BDI agent model, to get an adequate tourist package that satisfies her preferences: she would be rather happy of practicing rafting (r) and very happy going to a mountain place (m), she would like more to go climbing (c) than to go trekking, and she wouldn't like to go farther than 1000km from Buenos Aires (f). The user interface that helps her express these desires ends up generating a desire theory as follows:*

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

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

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

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 $\mathcal{T}'_{\mathcal{D}}$ remains consistent within DC:

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

Notice that if we consider a compound goal such as going to a far mountain place, represented by the conjunction $(m \wedge f)$, using the DC axioms, the theory $\mathcal{T}_{\mathcal{D}}$ proves the following lower bounds for the positive and negative desire degrees of $m \wedge f$:

$$(D^+(m \wedge f), 0.8), \quad (D^-(m \wedge f), 0.7)$$

So, $m \wedge f$ has both a high positive and negative desire degree. This has not to be seen as a shortcoming of the model, it actually reflects that the compound goal has one subgoal which is highly preferred but also another subgoal which is highly rejected.

[5] The only indirect constraint DC imposes is the following one: if a theory $\mathcal{T}$ derives $(D^+\varphi, r)$ and $(D^+\neg\varphi, s)$ then, due to axiom $(DC0^+)$ and rule (ID^+), $\mathcal{T}$ also derives both $(D^+\psi, \min(r, s))$ for any ψ.

3.4. Some Additional Consistency Schemas

The basic schema for preference representation and reasoning provided by the DC logic may be felt too general for some classes of problems and we may want to restrict the possible assignments of degrees to positive and negative desires for a formula φ and its negation $\neg\varphi$. For instance, as positive desires are proactive attitudes, it may not be efficient to assign non-zero degrees to $D^+\varphi$ and to $D^+\neg\varphi$, since the agent will be searching for necessarily conflicting plans, some aiming to satisfy φ and some to satisfy $\neg\varphi$.

We propose three different extensions, or schemes, that impose upon the basic logic three different consistency constraints between positive and negative desires, both at semantic and syntactic levels. The completeness proofs for the different extensions proposed run like Theorem 1 with the necessary modifications. These different schemes underpin different types of agents, as the constraints limit in different ways what formulae the agent will accept as desires.

3.4.1. DC_1 Schema

It may be unnatural in some domains to simultaneously have positive (in the sense of > 0) desire degrees for $D^+\varphi$ and $D^+\neg\varphi$. This constraint and its dual for negative desires amounts to require the following additional properties of the truth-evaluations 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 $\inf_{w\in W} \pi^-(w) = 0$.

At the syntactic level these conditions are equivalent to add to the basic axiomatic for DC the following two axioms:

(DC1$^+$) $D^+\varphi \wedge_L D^+(\neg\varphi) \rightarrow_L \bar{0}$ (or equivalently $\neg_L D^+(\top)$)
(DC1$^-$) $D^-\varphi \wedge_L D^-(\neg\varphi) \rightarrow_L \bar{0}$ (or equivalently $\neg_L D^-(\top)$)

Indeed, one can prove completeness with respect to $\mathcal{M}_{\mathcal{DC}1}$ for deductions from finite theories over the schematic extension of the DC logic with the above two axioms.

We would like to point out that under this schema one can consistently assign positive (and negative) desires degrees to compound goals starting from the degrees for the atomic ones (e.g. literals). Indeed, assume we have n atomic goals $p_1, \ldots p_n$, and consider a theory $T = \{D^+ l_i \equiv \alpha_i\}$, where l_i is p_i or $\neg p_i$, representing an assignment of positive desire degrees to literals under the above schema DC_1, that is, if $D^+ l_i \equiv \alpha_i \in T$, with $\alpha_i > 0$, then $D^+\neg l_i \equiv 0 \in T$. It is easy to show that this theory is consistent with assigning to conjunctions of different literals $l_1 \wedge \ldots \wedge l_j$ the compound degree

$$\alpha_1 \oplus \ldots \oplus \alpha_j$$

where e.g. $\oplus$ is the associative operation $x \oplus y = x + y - x \cdot y$. In this way one can consistently assign desire degrees to any compound goal.

3.4.2. DC_2 Schema

The above logical schema DC_1 does not put any restriction on positive and negative desires for the same goal (any classically satisfiable formula). According to Benferhat et al. in [3], 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 axiomatic counterpart that faithfully accounts for the above condition, we consider $\mathcal{M}_{DC_2}$ the subclass of DC-Kripke structures $M = (W, e, \pi^+, \pi^-)$ satisfying the above constraint between $\pi^+$and π^-. Note that $\pi^+(w) \leq 1 - \pi^-(w)$ iff $\pi^+(w) \otimes \pi^-(w) = 0$.[6]

To capture at the syntactical level this class of structures, we consider the extension DC_2 of the DC system with the following axiom: [7]

(DC2) $(D^+\varphi \otimes D^-\varphi) \rightarrow_L \bar{0}$

As in the previous subsection, DC_2 can be proved to be complete for finite theories with respect to the subclass $\mathcal{M}_{DC_2}$ of DC-structures.

3.4.3. DC_3 Schema

A stronger consistency condition between positive and negative preferences was considered in [6], requiring that if a world is rejected to some extent, it cannot be positively desired at all. And conversely, if a goal is somewhat desired it cannot be rejected. Indeed, at the semantic level, this amounts to require the intended DC-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, $\min(\pi^+(w), \pi^-(w)) = 0$)

This is a stronger condition than the one presented in the DC_2 schema. We will denote by $\mathcal{M}_{DC_3}$ the subclass of DC-Kripke structures satisfying it. At the syntactic level, the axiom that faithfully represents this consistency condition is the following one:

(DC3) $(D^+\varphi \wedge_L D^-\varphi) \rightarrow_L \bar{0}$

Again, the extension of DC logic with the $(DC3)$ axiom can be proven to be complete for finite theories with respect to the subclass $\mathcal{M}_{DC_3}$ of DC-structures.

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 DC_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.

If María selects DC_2, the formulae D^+c and D^-c are not allowed to have degrees adding up to more than 1, and hence the above theory $\mathcal{T}''_{\mathcal{D}}$ becomes inconsistent. Actually, $\mathcal{T}''_{\mathcal{D}}$ becomes also inconsistent under DC_3, DC_3 is stronger than DC_2 (it does not even allow to have non-zero degrees for D^+c and D^-c). In these cases, the agent should apply a revision mechanism, for instance to remove $(D^+c, 0.85)$ from the theory.

4. Intention Context (IC)

This context represents the agent intentions. We follow the model introduced by Rao and Georgeff [23], in which an intention is considered a fundamental pro-attitude with an explicit representation. However, as in the work of Cohen and Levesque [11], in our approach, intentions result from the agent's beliefs and desires and then, we do not consider them as a basic attitude. Intentions, as well as desires, represent a sort of the agent preferences. We consider that intentions cannot depend just on the benefit, or satisfaction, of reaching a (positive) desire φ —represented in $D^+\varphi$, but also on the world's state w and the cost of transforming it into a world w' where the

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

[7] An equivalent presentation of axiom (DC2) is $D^+\varphi \rightarrow_L \neg_L D^-\varphi$.

formula φ is true. By allowing degrees in intentions we represent a measure of the cost/benefit relation involved in the agent's actions towards the desired goal. The formalization of the intention semantics alone is difficult, because generally it will not only depend on the formula intended, but also on the plan that the agent executes to achieve a state where the formula is valid. Therefore, as it will be shown below, we have chosen to a very general framework to axiomatize intentions, able to be specialized to more particular semantics if needed.

We represent in this context two kinds of graded intentions, intention of a formula φ considering the execution of a particularly plan α, noted $I_\alpha\varphi$, and the final intention to φ, noted $I\varphi$, which takes into account the best path to reach φ. As in the other contexts, if the degree of $I\varphi$ is δ, it may be considered that the truth degree of the expression "φ is intended" is δ. The intention to make φ true must be the consequence of finding a feasible plan α, that permits to achieve a state of the world where φ holds.

The language $\mathcal{L}_{IC}$ to represent the agent intentions is defined in a similar way as we did in DC, starting with the same basic propositional language $\mathcal{L}$ and incorporating a family of many-valued modal operators. We assume the agent has a finite set of actions or plans Π^f (a finite subset of the potentially infinite set of actions Π) at her disposal to achieve the desires. Then, for each $\alpha \in \Pi^f$ we introduce a modal 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.[8] We also introduce another modal operator I with the idea that the truth-degree $I\varphi$ will represent the intention degree with which the agent intends φ by means of the best plan in Π^f.

Models for IC are Kripke structures $M = \langle W, e, \{\nu_\alpha\}_{\alpha\in\Pi^f}\rangle$ where, as in the DC structures, W is a set of worlds and $e : W \times Var \to \{0, 1\}$ is such that, for each world $w \in W$, $e(w, \cdot)$ is a Boolean evaluation of propositional variables, which is extended to propositional formulae as usual. Here, for each $\alpha \in \Pi^f$, $\nu_\alpha : \mathcal{U} \to [0, 1]$ is a mapping (where $\mathcal{U} \subseteq 2^W$ is such that the sets $\{w \in W \mid e(w, \varphi) = 1\}$ are ν-measurable for each proposition φ) which provides a measure of the degree of intention for (the set of models of) a goal by means of the plan α. Intuitively, without any other information about how much the goal is desired, how costly is the application of α or how probable is that the goal will become true by executing α, the only property which is required to ν_α is that a disjunction of goals has to be intended at least to the degree of the least intended goal, i.e. for each $A, B \in \mathcal{U}$:

$$\min(\nu_\alpha(A), \nu_\alpha(B)) \leq \nu_\alpha(A \cup B)$$

Actually, this is a very general condition, which is compatible with different particular semantics an agent may consider. Indeed, as one would expect, this is a weaker condition than the one required for desires since in case degrees of intentions become independent of the cost of the actions (because e.g. all actions have the same cost), their properties should be basically the same as those of desires. This condition is compatible as well with the semantics of intention degrees used in step 5 of Section 6 as a form of expected global benefit, where benefit is understood as the utility of reaching a goal (as a function of its desire degree) minus the cost of the action used to reach that goal[9] Indeed, let us denote by $[\varphi]$ the set of worlds where φ is true, $P_\alpha([\varphi])$ the probability of making φ true after α, $D([\varphi])$ the positive desire degree of φ, $u : [0, 1] \to \mathbb{R}$ a non-increasing mapping transforming desire degrees into negative costs, c_α the (real-valued) cost of the action α, and $h : \mathbb{R} \to [0, 1]$ a non-decreasing map interpreting benefits into normailzed utility degrees, and define $\nu_\alpha([\varphi]) = h(P_\alpha([\varphi]) \cdot (u(D([\varphi])) - c_\alpha))$. Then one can indeed check that the inequality $\nu_\alpha([\varphi] \cup [\psi]) \geq \min(\nu_\alpha([\varphi]), \nu_\alpha([\psi]))$ holds.

Finally, let us mention that the above condition on ν_α says nothing about whether the intention degree is monotonically increasing or decreasing with respect to inclusion (implication). In fact, if a goal φ has a relatively high intention degree, then the joint intention degree of φ with another goal ψ, i.e. the degree of the conjunction $\varphi \wedge \psi$, can be higher when ψ is also highly desired, but

[8] In the IC context we are not concerned about the question of whether a given desire can be reached by the execution of a particular action, this is left for the Planner context, see Section 6.

[9] A similar semantics for intentions is used in [24], 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.

can also be lower if ψ is rejected.

Then, an evaluation e in each world is extended to atomic modal formulae by stipulating

- $e(w, I_\alpha\varphi) = \nu_\alpha(\{w \in W \mid e(w, \varphi) = 1\})$,
- $e(w, I\varphi) = \max\{e(w, I_\alpha\varphi) \mid \alpha \in \Pi^f\}$

and to compound modal formulae using the truth functions of Rational Pavelka logic.

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

A complete axiomatics for the IC logic with respect to the class of structures $\mathcal{M}_{\mathcal{IC}}$ is the following:

1. Axioms of classical logic for the non-modal formulae
2. Axioms of Rational Pavelka logic for the modal formulae
3. Axiom for I_α modalities: $(I_\alpha\varphi \wedge_L I_\alpha\psi) \rightarrow_L I_\alpha(\varphi \vee \psi)$
4. Definitional Axiom for I: $I\varphi \equiv_L \bigvee_{\alpha\in\Pi^f} I_\alpha\varphi$
5. Inference Rules: modus ponens for $\rightarrow$ and for $\rightarrow_L$, and introduction of I_α for equivalences: from $\varphi \equiv \psi$ derive $I_\alpha\psi \equiv_L I_\alpha\varphi$ for each $\alpha \in \Pi^f$.

The axiom for the I_α is indeed weaker than the $(DC0^+)$ and $(DC0^-)$ axioms for desires, while axiom 4 syntactically captures the idea that the intention degree of φ is the highest degree with which the agent intends φ by means of some plan in Π^f. The rule of introduction of the I_α operators for equivalences is needed to guarantee the syntactic irrelevance of these operators, and it is obviously weaker than the rules of introduction of the positive and negative desires for implications in the DC.

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

Theorem 2. *Let $\mathcal{T}$ be a finite modal theory and Φ a modal formula. Then $\mathcal{T} \vdash_{IC} \Phi$ iff $\mathcal{T} \models_{\mathcal{M}_{IC}} \Phi$.*

5. Belief Context (BC)

In this context the agent represents her uncertain beliefs about the world where she lives. Since situated agents need to reason about their possible actions and the changes they may cause to the environment, the potential consequences of actions must be part of any situated agent's beliefs set. We use propositional Dynamic logic (PDL) [19] to describe statements related to action execution, and to represent probabilistic beliefs on them we adopt a similar fuzzy modal approach to the one used in the other mental contexts. Here, atomic modal formulas of the BC logic are of the form $B\varphi$, where φ is a PDL formula. In particular, $B[\alpha]\varphi$ is meant to denote that it is *likely*, or *probable*, that after executing action α, φ becomes true. The probabilistic logic for BC is then axiomatized as a theory in Rational Pavelka logic (RPL), again in a similar way than as we do in the DC and IC contexts. Due to space limitations, details on the language definition, axioms and rules for the BC context are not included here but can be checked in [6]. We would like to note that other uncertainty models (like the possibilistic necessity model) might be used as well.

6. Functional contexts and Bridge rules

In this section we present the remaining necessary components of our multi-context agent architecture example: the Planner context (PC), the Communication context (CC) and the Bridge rules (BR). The Planner context (PC) builds plans to satisfy the user's desires, where plans have an associated cost according to the actions involved. The theory of planning in PC includes special predicates to represent actions, plans, beliefs and desires, such as the predicate *fplan*:

- *fplan(φ, α, preC, postC, c_α)* is generated within PC when a plan instance is found to be feasible, that is, if and only if: (i) α makes φ true, (ii) φ is such that $(D^+\varphi, d_0)$ belongs to the DC theory of the agent over a given threshold d_0, (iii) the preconditions hold to some degree, i.e. $(BpreC, s_0)$ must be in the BC theory for a given threshold s_0, and (iv) avoids negative desires as post-conditions, i.e. if $(\neg_L D^- postC, 1 - e_0)$ belongs to the agent DC theory then e_0 must be very small. $c_\alpha \in \mathbb{R}$ is the cost of the plan.

The communication context (CC) makes it possible to encapsulate the agent's internal structure by having a unique and well-defined interface with the environment. The theory inside this context will take care of the sending and receiving of messages to and from other agents in the multiagent society where our g-BDI agents live. The communication context perceives changes in the environment and adds them as beliefs into the belief context BC via a bridge rule (see step 1 below). Also, it declares the preferred actions to execute via another bridge rule (see step 7 below).

For our running illustrative example of g-BDI agent architecture (see Figure 1), we define a set of basic bridge rules and we show the information flow from perception to action.

1. *CC perceives the environment:* The agent perceives the environment and generates graded beliefs (where the degree r depends on the perception) by means of the following bridge rule:
$$\frac{CC : \varphi}{BC : (B\varphi, r)}$$
2. *DC gets the user's graded positive and negative desires:* The agent receives the user preferences as formulae in the DC context in a similar way: from user's expressed desires, and using a bridge rule, to graded positive and negative desires.
3. *Desires and beliefs are passed from DC and BC to PC:* From positive and negative desires, and beliefs about plans and domain knowledge, bridge rules generate corresponding predicate instances (quoting using $\lceil . \rceil$)[10] in the PC context:
$$\frac{DC : (D^+\varphi, d)}{PC : posdesire(\lceil (D^+\varphi, d) \rceil)} \quad \frac{DC : (D^-\psi, d)}{PC : negdesire(\lceil (D^-\psi, d) \rceil)} \quad \frac{BC : (B\Phi, r)}{PC : belief(\lceil (B\Phi, r) \rceil)}$$
4. *PC looks for feasible plans*, as mentioned above, feasible plans fulfill (to some degree) positive desires, satisfy some preconditions and avoid undesired postconditions. Thus, PC generates predicate instances of *fplan(φ, α, PreC, PostC, c_α)*.
5. *A process for deriving intentions:* Here we assume actions (and plans) do not fail[11] and that a belief degree r in a formula $(B[\alpha]\varphi, r)$ is interpreted as the probability that φ satisfies the user by executing α. Then the intention degree to reach a desire φ by means of a plan α is taken as a trade-off between the benefit of reaching this desire and the cost of the plan, weighted by the belief degree r. This is implemented by the following bridge rule:
$$\frac{DC : (D^+\varphi, d), BC : (B[\alpha]\varphi, r), PC : fplan(\varphi, \alpha, preC, postC, c_\alpha)}{IC : (I_\alpha\varphi, h(r \cdot (u(d) - c_\alpha)))} \quad (1)$$
where $u : [0, 1] \to \mathbb{R}$ is a non-decreasing mapping that transforms desire degrees into negative costs (benefits), i.e. $u(d)$ can be interpreted as how much the user accepts to pay to achieve a goal desired to the degree d, and $h : \mathbb{R} \to [0, 1]$ is a non-decreasing transformation that maps global benefits back to normalized utility degrees. Indeed, the value $h(r \cdot (u(d) - c_\alpha))$ can be read as a monotone transformation of the *expected benefit* of intending φ through plan α.
6. *Action selection process:* The information supplied by the above bridge rule to the IC context allows this context to derive $(I\varphi, i_\varphi)$, a single intention formula for each desire φ. That is, i_φ is the intention degree of the best feasible plan for φ (see the definitional axiom for I in Section 4).

[10] A quoting mechanism such as $\lceil . \rceil$ allows to transform modal formulae into first order logic terms.
[11] Otherwise, we would need to consider richer stochastic processes like MDPs or POMDPs.

7. *The PC and IC inform CC of the best plans for each desire:* This is done by the following rule:

$$\frac{PC : fplan(\varphi, \alpha, preC, postC, c_\alpha), IC : (I_\alpha\varphi, i_\varphi), IC : (I\varphi, i_\varphi)}{CC : do(\alpha, i_\varphi)}$$

The agent interacts with the environment through the Communication Context CC by declaring which plan α the agent will finally execute. To do so, the CC context selects the action with the highest degree among the formulae $do(\alpha, i_\varphi)$ received via the previous bridge rule.

7. Discussion and Related Work

In this research note we have reported on the development of a graded intentional agent scheme for practical reasoning, defined as a multi-context system, where logical contexts provide a formal representation (complete axiomatizations) of the different mental attitudes (belief, desires and intentions) and an interaction framework for them.

The preliminary idea of extending BDI architectures with graded mental attitudes goes back to Parsons and Giorgini [20], where, also within a multi-context system approach, they introduce graded beliefs by means of mass assignments in the sense of the Dempster-Shafer theory. This multi-context model of BDI agents in a bi-valued approach has been previously proposed in [21]. Blee et al. [3] also introduce grades in all the mental notions of BDI. They use a common syntax for the agent's mental attitudes while in our multi-context system approach we define in a separate way a suitable logic for each attitude.

The problem of preference representation has been tackled in the literature with quite a different number of approaches. Inspired in [2, 3], a possibilistic bipolar representation of preferences has been used in our agent model. This is related but somewhat different from the logic defined by Lang et al. in [18] 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 (qualitative) partial preference ordering over worlds. In contrast with Boutiller [5] they also differentiate between factual background knowledge, that tells which worlds are physically impossible, and contingent knowledge, expressing which of the physical possible worlds can be the actual state of affairs. Then, the agent should aim at the best feasible world by performing a suitable action but their work is not focused on action theories. We find some differences with respect to our agent mode. First, it does not include an explicit representation of the agent rejections (negative preferences); second desires are only qualitatively represented (preference order over worlds) while in our model can also represent the strength of them in a numerical way.

An argumentation approach to practical reasoning is proposed in [22, 1] where they provide a rich argumentation-based framework for uncertain beliefs, consistent desires, and for generating consistent plans for achieving these desires. These plans, called intentions, are generated via some ad-hoc rules, that could be mapped into particular bridge rules in our approach. In contrast to our approach, the authors do not present a, strictly speaking, formal system (a sound and complete logical system) to represent and reason with these graded attitudes according to a suitable uncertainty model.

Finally, in [12, 13] the authors also propose a very similar possibilistic approach to deal with graded beliefs and desires, that are used afterwards to determine the agent's goals. A set of desire generation rules (similar to the ones in [22]) determine the graded positive desires and they refer to the representation of negative desires we use in our model as future work. In particular, in [13] they propose different belief change operators to deal with trust and distrust. Belief change may induce changes in the justification degrees of some desires/goals. Then, in [12] they explore the belief-goal consistency and incompleteness in the proposed formalism for BDI agents. These relations and the ones called realisms [21], as well as the desire generation rules, can also be modeled in the g-BDI agent architecture by defining appropriate bridge rules.

It is clearly a matter for further research to include in our logical framework a revision process for beliefs, desires and intentions in order to keep these attitudes consistent for agents living in dynamic environments.

Acknowledgments The authors thank the anonymous reviewers for many useful comments that have helped to significantly improve the paper and acknowledge partial support of the Spanish project AT (Consolider CSD2007-022, Ingenio 2010).

References

[1] Amgoud L. and Prade H., Formalizing Practical Reasoning under Uncertainty: An Argumentation-Based Approach. In *Proc. of IAT 2007*, 189-195, 2007.

[2] Benferhat S., Dubois D. and Prade, H., Towards a possibilistic logic handling of preferences. *Applied Intelligence* 14(3):303-317, 2001.

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

[4] Blee J., Billington D., and Sattar A., Reasoning with levels of modalities in BDI logic. In *Proc. of PRIMA07*, LNAI 5044, 410-415, 2009.

[5] Boutilier C., Towards a logic for qualitative decision theory. In *Proc. of KR'94*, 75-86, 1994.

[6] Casali A., Godo L. and Sierra C., Graded BDI Models For Agent Architectures. In *Proc. of CLIMA V*, LNAI 3487, 126-143, Springer-Verlag, 2005.

[7] Casali A., Godo L. and Sierra C., A Language for the Execution of Graded BDI Agents. To appear in *L. J. of the IGPL*. A preliminary version appears in *Proc. of FAMAS'007*, 65-82, Durham, UK, 2007.

[8] Casali A., Godo L. and Sierra C., A Logical Framework to Represent and Reason about Graded Preferences and Intentions. In *Proc. of KR 2008*, 27-37, 2008.

[9] Casali A., Godo L., Sierra C., Validation and Experimentation of a Tourism Recommender Agent based on a Graded BDI Model. In: *Artificial Intelligence Research and Developement*, Proc. of CCIA 2008, T. Alsinet et al. (Eds.), IOS Press, pp. 41-50, 2008.

[10] Casali, A., Godo, L., Sierra, C., A Tourism Recommender Agent: From theory to practice. *Inteligencia Artificial* 40, 23-38, 2008.

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

[12] da Costa Pereira C., Tettamanzi A., Goal Generation from Possibilistic Beliefs Based on Trust and Distrust. In *Proc. of DALT 2009*, LNAI 5948, pp. 35-50, 2010.

[13] da Costa Pereira C., Tettamanzi A., Belief-Goal Relationships in Possibilistic Goal Generation. In *Proc. of ECAI 2010*, 641-646, 2010.

[14] Georgeff M., Pell B., Pollack M., Tambe M. and Wooldridge M., The Belief-Desire-Intention Model of Agency. In J.P. Muller et al. (Eds.), *Proc. of ATAL'98*, LNAI 1555, 1-10, Springer, 1999.

[15] Giunchiglia F. and Serafini L., Multilanguage Hierarchical Logics (or: How we can do without modal logics). *Artificial Intelligence* 65, pp. 29-70, 1994.

[16] Hájek P., *Metamathematics of Fuzzy Logic*, Trends in Logic 4, Kluwer Academic Pub., 1998.

[17] Halpern J. Y., *Reasoning about Uncertainty*. The MIT Press. Cambridge Massachusetts, 2003.

[18] Lang J., van der Torre, L. and Weydert E., Utilitarian Desires. *Autonomous Agents and Multi Agent systems*, vol. 5:3, 329-363, 2002.

[19] Meyer, J.-J. Ch., Dynamic Logic for Reasoning about Actions and Agents. In (J. Minker ed.) *Logic-Based Artificial Intelligence*, Kluwer Academic Publishers, 281-311, 2000.

[20] Parsons S. and and Giorgini P., An approach to using degrees of belief in BDI agents. In (B. Bouchon-Meunier et al., eds.) *Information, Uncertainty, Fusion*, Kluwer, 81-92, 1999.

[21] Parsons S., Sierra C. and Jennings N. R., Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, 8(3): 261-292, 1998.

[22] Rahwan, I., and Amgoud, L., An Argumentation based Approach for Practical Reasoning. In *Proc. of AAMAS 2006*, 347-354, 2006.

[23] Rao, A. and Georgeff M., BDI agents: From theory to practice. In *Proc. of the 1st International Conference on Multi-Agents Systems*, pp 312-319, 1995.

[24] Schut, M., Wooldridge, M. and Parsons S.,Reasoning About Intentions in Uncertain Domains. In *Proc. of ECSQARU 2001*, LNAI 2143, 84-95, Springer, 2001.

Appendix I: About Rational Pavelka logic

Rational Pavelka logic RPL is an extension of Łukasiewicz's infinitely-valued logic by expanding its language with rational truth-constants to explicitly reason about degrees of truth. It was introduced by Pavelka in the late seventies and Hájek provided in [16] a simpler formalization.

Formulae are built from propositional variables $p_1, p_2, \ldots$ and truth constants $\overline{r}$ for each *rational* $r \in [0,1]$ using two connectives, an implication $\to_L$ and a negation $\neg_L$. Other connectives can defined from these ones, in particular two conjunctions, two disjunctions and an equivalence:

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

Łukasiewicz's truth functions for the connectives $\to_L$ and $\neg_L$ are (using the same symbols as for the connectives):

$$x \to_L y = \min(1, 1 - x + y) \qquad \neg_L x = 1 - x$$

Taking them into account, the corresponding truth functions for the above definable connectives are:

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

An RPL *evaluation* e is a mapping of propositional variables into $[0,1]$. Such a mapping uniquely extends to an evaluation of all formulae using the above truth functions and defining $e(\overline{r}) = r$ for each rational $r \in [0,1]$. Note that $e(\overline{r} \to_L \varphi) = 1$ iff $r \leq e(\varphi)$. An evaluation is a model of a set of formulae (theory) $\mathcal{T}$ whenever $e(\varphi) = 1$ for all $\varphi \in \mathcal{T}$. A formula ψ is a logical consequence of a theory $\mathcal{T}$, written $\mathcal{T} \models_{RPL} \psi$, whenever $e(\psi) = 1$ for every evaluation e that is a model of $\mathcal{T}$.

Logical axioms of RPL are:

(i) axioms of Łukasiewicz's logic

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

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

$$\neg_L\overline{r} \equiv_L \overline{1 - r} \qquad \overline{r} \to_L \overline{s} \equiv_L \overline{\min(1, 1 - r + s)}$$

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 has been shown to be complete for deductions from finite theories: for each *finite* $\mathcal{T}$ and φ, $T \vdash_{RPL} \varphi$ iff $T \models_{RPL} \varphi$.