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

Ana Casali¹, Lluís Godo² and Carles Sierra²

Depto. de Sistemas e Informática
Facultad de Cs. Exactas, Ingeniería y Agrimensura
Universidad Nacional de Rosario, Argentina.

Institut d’Investigació en Intel·ligència Artificial (IIIA) - CSIC
Bellaterra, Catalunya, Spain.
Motivations

- Preferences are essential for making choices in complex situations, for mastering large sets of alternatives, and for coordinating a multitude of decisions.
- Negative preferences are also important in modelling different AI problems.
- Following previous works on bipolarity representation of preferences (Benferhat et al.) we contribute:
  - A Basic Logic for graded Desires representation.
  - Different Schemas (extensions) are proposed.
  - An extension to model Intentions is proposed
  - A sound and complete axiomatics are given for each logic.
  - Some operational elements.
María, who lives in busy Buenos Aires, wants to relax for a few days in an Argentinian beautiful destination. She activates a personal agent, 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 1000 km from Buenos Aires \((f)\). She is stressed and would like to get to the destination with a short trip.

*How can the personal agent represent María’s preferences?? How can these desires be used to choose the most adequate tourist package??*
Basic Logic for Desires representation

The Language

- We start from a basic propositional language \( \mathcal{L} \).
- To represent positive and negative desires over formulae of \( \mathcal{L} \), we introduce two modal operators \( D^- \) and \( D^+ \).
- \( D^+ \varphi \) and \( D^- \varphi \) are many-valued formulae, which respectively read as “\( \varphi \) is positively desired” and “\( \varphi \) is negatively desired” (or “\( \varphi \) is rejected”).

The expanded language \( \mathcal{L}_D \) is defined as follows:

- If \( \varphi \in \mathcal{L} \) then \( \varphi \in \mathcal{L}_D \)
- If \( \varphi \in \text{Sat}(\mathcal{L}) \) then \( D^- \varphi, D^+ \varphi \in \mathcal{L}_D \)
- If \( r \in \mathbb{Q} \cap [0, 1] \) then \( r \in \mathcal{L}_D \)
- 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 are definable from \( \neg_L \) and \( \rightarrow_L \))
Basic Logic for Desires representation

The Language

- The notation \((D^+\psi, r)\), with \(r \in [0, 1] \cap \mathbb{Q}\), will be used as a shortcut of the formula \(\tilde{r} \rightarrow_L D^+\psi\), specifying that the level of positive desire of \(\psi\) is at least \(r\).

- The agent’s desires will be represented by a (finite) set of modal formulae or theory \(\mathcal{T} \subset \mathcal{L}_D\) containing:
  - *quantitative expressions* about positive and negative preferences, like \((D^+\varphi, \alpha)\) or \((D^-\psi, \beta)\),
  - *qualitative expressions* like \(D^+\psi \rightarrow_L D^+\varphi\) (resp. \(D^-\psi \rightarrow_L D^-\varphi\)), expressing that \(\varphi\) is at least as preferred (resp. rejected) as \(\psi\).
Semantics - Intuition

- the degree of positive desire for (or level of satisfaction with) a disjunction of goals $\varphi \lor \psi$ is taken to be the minimum of the degrees for $\varphi$ and $\psi$ (Benferhat et al.). *Analogously for negative desires.*

- the satisfaction degree of reaching both $\varphi$ and $\phi$ can be strictly greater than reaching one of them separately. These are basically the properties of the *guaranteed possibility* measures. *The same for negative desires.*
Semantics

The intended models for $L_D$ are Kripke structures $M = \langle W, e, \pi^+, \pi^- \rangle$, Bipolar Desire models.

$\pi^+: W \rightarrow [0, 1]$ and $\pi^-: W \rightarrow [0, 1]$ are positive and negative preference distributions over worlds, which are used to give semantics to positive and negative desires:

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

$e$ is extended to compound modal formulae by means of the usual truth-functions for Łukasiewicz connectives.

The evaluation $e(\Phi, w)$ of a modal formula $\Phi$ only depends on the formula itself and not on the actual world $w \in W$ where the agent is situated. This is consistent with the intuition that desires represent ideal preferences of an agent.
Basic Logic for Desires representation

Axioms and Rules: we need to combine axioms for the different formulae (non-modal and modal) and additional axioms characterizing the behaviour of the modal operators $D^+$ and $D^−$.

We define the Basic Bipolar Desire logic (BD logic) as follows:

Axioms:

(CPC) Axioms of classical logic for non-modal formulae
(RPL) Axioms of Rational Pavelka logic for modal formulae

(BD0$^+$) $D^+(A \lor B) \equiv L D^+ A \land L D^+ B$

(BD0$^-$) $D^−(A \lor B) \equiv L D^− A \land L D^− B$

(BD0$^+$) and (BD0$^-$) define the behaviour of $D^−$ and $D^+$ with respect to disjunctions.
Inference Rules:

(MP1) modus ponens for $\rightarrow$
(MP2) modus ponens for $\rightarrow_L$

Introduction of $D^+$ and $D^-$ for implications:
(ID$^+$) from $\varphi \rightarrow \psi$ derive $D^+\psi \rightarrow_L D^+\varphi$
(ID$^-$) from $\varphi \rightarrow \psi$ derive $D^-\psi \rightarrow_L D^-\varphi$.

- The introduction rules for $D^+$ and $D^-$ state that the degree of desire is monotonically decreasing with respect to logical implication. Then, equivalent desires degrees are preserved by classical (Boolean) equivalence.
- $\vdash_{BD}$ denotes the notion of proof and is defined as usual.
Soundness and Completeness
The above axiomatization is correct with respect to the defined semantics.

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

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

**Theorem (completeness)**
\[ \text{Let } T \text{ be a finite theory of closed formulae and } \Phi \text{ a closed formula. Then } T \models_{\mathcal{M}_{BD}} \Phi \text{ iff } T \vdash_{BD} \Phi. \]
Example 1

María, our tourist, activates a personal agent based on our BD logical framework, to get 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:

\[
\mathcal{T}_D = \{(D^+ m, 0.8), (D^+ r, 0.6), D^+ m \rightarrow_L D^+ c, (D^- f, 0.7)\}
\]
Example 1

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

- $\mathcal{T}_D \vdash_{BD} (D^+(m \land r), 0.8)$,
- $\mathcal{T}_D \vdash_{BD} (D^+(m \lor r), 0.6)$,
- $\mathcal{T}_D \vdash_{BD} (D^+c, 0.8)$

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: $(D^+(m \land r), 0.95)$.

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

The extended theory $T'_D$ remains consistent within BD.
Different Schemas

The basic logical schema $BD$ puts almost no constraint on the strengths for the positive and negative desire of a formula $\varphi$ 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.

- This may be too general for some classes of problems and we may want to restrict the allowed assignment of degrees of positive and negative desires.

- Three different extensions are proposed to show how different consistency constraints can be added to the $BD$ logic, both at the semantical and syntactical level.

- These different schemas allow us to define different types of agents.
It may be natural in some domain applications to forbid to simultaneously have $D^+\varphi$ and $D^+-\varphi$.
This constraint and the corresponding one for negative desires amounts to require:

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

At the level of Kripke structures, this corresponds to:

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

Let $\mathcal{M}_{BD_1}$ denote the subclass of models satisfying these conditions.

At the syntactic level, to require:

- $(BD_1^+)\ D^+\varphi \land_L D^+(\neg\varphi) \rightarrow_L \bar{0}$
- $(BD_1^-)\ D^-\varphi \land_L D^-(-\varphi) \rightarrow_L \bar{0}$
$BD_2$ Schema

An agent cannot desire to be in a world more than the level at which it is tolerated (not rejected). Translated to our framework, amounts to require:

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

To capture at the syntactical level this class of structures, we consider:

$$(BD_2) (D^+ \varphi \otimes D^- \varphi) \rightarrow_L \tilde{0}$$
**BD$_3$ Schema**

If a world is rejected (desired) to some extent, it cannot be positively desired (rejected) at all. At the semantical level, this amounts to:

\[
\min(\pi^+(w), \pi^-(w)) = 0
\]

At the syntactic level:

\[
(BD3) \ (D^+\varphi \land \neg D^-\varphi) \rightarrow \neg 0
\]

**Theorem (completeness)**

*Let T be a finite theory of closed formulae and $\Phi$ a closed formula. Then*

- $T \models_{BD_1} \Phi$ ****iff**** $T \vdash_{BD_1} \Phi$
- $T \models_{BD_2} \Phi$ ****iff**** $T \vdash_{BD_2} \Phi$
- $T \models_{BD_3} \Phi$ ****iff**** $T \vdash_{BD_3} \Phi$. 
Example 2

(Example 1 continuation)
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$_1$, as this schema allows for opposite desires the agent simply adds the formula ($D^−c, 1$) into the former desire theory $\mathcal{T}_D'$, yielding the new theory

\[
\mathcal{T}_D'' = \{(D^+m, 0.8), (D^+r, 0.6), (D^+(m \land r), 0.95), (D^+c, 0.85), (D^−f, 0.7), (D^−c, 1)\},
\]
Example 2

If María selects $BD_2$, the formulae $D^+ c$ and $D^- c$ are not allowed to have degrees summing up more than 1, and hence the above theory $T_D''$ becomes inconsistent. Actually, $T_D''$ 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

How these positive and negative desires may be used for the agent to generate intentions?

- in our approach intentions cannot depend just on the satisfaction, of reaching a goal $\varphi$ (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 $\varphi$ is true.

- by allowing a graded representation of the strength of intentions we are able to represent a measure of the cost/benefit relation involved in the feasible actions the agent can take toward the intended goal.
Logic for Intention representation

Language
- A set of propositional variables $\text{Var}_{\text{cost}} = \{C_\alpha\}_{\alpha \in \Pi^0}$
- Elementary modal formulae $I_\alpha \varphi$ and $I \varphi$, where $\alpha \in \Pi^0$

Semantics
The intended models will be enlarged Kripke structures $M = \langle W, e, \pi^+, \pi^-, \{\pi_\alpha\}_{\alpha \in \Pi^0} \rangle$ where $\pi_\alpha : W \times W \to [0, 1]$ is a utility distribution corresponding to $\alpha$

Additional axioms and rules
1. (BD0) axiom for $I_\alpha$ modalities: $I_\alpha (\varphi \lor \psi) \equiv_L I_\alpha \varphi \land_L I_\alpha \psi$
2. Definitional Axiom for $I$: $I \varphi \equiv_L \bigvee_{\alpha \in \Pi^0} I_\alpha \varphi$
3. introduction of $I$ and $I_\alpha$ for implications: from $\varphi \rightarrow \psi$ derive $I_\psi \rightarrow_L I \varphi$ and $I_\alpha \psi \rightarrow_L I_\alpha \varphi$ for each $\alpha \in \Pi^0$

Theorem

Let $T$ be a theory of closed modal formulae and $\Phi$ a closed modal formula. Then $T \vdash_{\text{DI}} \Phi$ iff $T \models_{\mathcal{M}_{\text{DI}}} \Phi$. 
Agent architecture

environment input

- belief revision
- desire generation and revision

Beliefs

- Desires

- planning
- deriving Intentions

- f-plans

- Intentions

- selecting Intention

action output
Example 3

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

- **Beliefs:** the agent updates its current beliefs about the tourism plans offered, the tourism domain (destinations ontologies) and the beliefs about how these packages can satisfy the user’s preferences.

- **Desire generation:** were generated in Example 1:  
  \[ T'_D = \{(D^+ m, 0.8), (D^+ r, 0.6), (D^+ (m \land r), 0.95), (D^- f, 0.7)\} \]

- **Planning, to look for feasible packages:** from this set of positive and negative desires (\(T'_D\)) and knowledge about the offered tourist packages and the benefits they bring, and using a Planner, the agent looks for feasible plans, that are believed to achieve positive desires (\(m, r\) or \(m \land r\)) but avoiding the negative desire (\(f\)) as post-condition.
Example 3

- **Current feasible packages:** the agent finds that the plans *Mendoza* (*Me*) and *SanRafael* (*Sr*) are feasible plans for the combined goal $m \land r$, while *Cumbrecita* (*Cu*) is feasible only for $m$. The Planner also computes the normalized cost of these plans: $c_{Me} = 0.60$ and $c_{Sr} = 0.70$ and $c_{Cu} = 0.55$.

- **Deriving the Intention formulae $l_\alpha \varphi$, for each feasible plan $\alpha$ toward a desire $\varphi$:** The intention degrees for satisfying each desire $m$, $r$ and $m \land r$ by the different feasible plans are computed by the following rule that trades off the cost and benefit of satisfying a desire by following a plan:

  $$(D^+ \varphi, d), \ fplan(\varphi, \alpha, \chi, A, c)$$

  $$(l_\alpha \varphi, f(d, c))$$
Example 3

The Agent uses 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 compute the different intentions degrees.

- **Current Intentions:** as a result of the previous process, the set of intentions contains the following formulae:
  
  - $(I_{Me}(m \land r), 0.675)$,
  - $(I_{Sr}(m \land r), 0.625)$,
  - $(I_{Me}(m), 0.60)$,
  - $(I_{Me}(r), 0.50)$,
  - $(I_{Sr}(m), 0.55)$,
  - $(I_{Sr}(r), 0.45)$,
  - $(I_{Cu}(m), 0.625)$.

- **Selecting Intention-plan:** the agent decides to recommend the plan *Mendoza (Me)* since it brings the best cost/benefit relation (represented by the intention degree 0.675) to achieve $m \land r$, satisfying also the tourist’s negative desire.
We have formalized a logical framework to represent and reason about graded agent desires and intentions.

A bipolar representation of preferences has been used to represent the agent positive and negative desires.

We have provided some operational rules showing how the agent desires, together with other information may be used in deciding the agent intention and the best plan to follow.

We plan to include a revision process for desires and intentions in order to keep these attitudes consistent for agents living in dynamic environments.