Picking Up the Best Goal
An Analytical Study in Defeasible Logic

Guido Governatori\textsuperscript{1,4}, Francesco Olivieri\textsuperscript{1,2,4}, Antonino Rotolo\textsuperscript{3}, Simone Scannapieco\textsuperscript{1,2,4}, and Matteo Cristani\textsuperscript{2}

\textsuperscript{1} NICTA, Queensland Research Laboratory, Australia
\textsuperscript{2} Department of Computer Science, University of Verona, Italy
\textsuperscript{3} CIRSFID and DSG, University of Bologna, Italy
\textsuperscript{4} Institute for Integrated and Intelligent Systems, Griffith University, Australia

Abstract. In this paper we analyse different notions of the concept of goal starting from the idea of sequences of “alternative acceptable outcomes”. We study the relationships between goals and concepts like agent’s beliefs, norms, and desires, and we propose a formalisation using Defeasible Logic that will be able to provide a computationally feasible approach. The resulting system captures various nuances of the notion of goal against different normative domains, for which the right decision is not only context-dependent, but it must be chosen from a pool of alternatives as wide as possible.

1 Motivation and Basic Intuitions

The BDI architecture is a prominent approach to model rational agents \cite{1,5}. BDI agents are means-ends reasoners equipped with: (i.) Desires, Goals, Intentions (or Tasks); (ii.) a description of the current state of the environment (Beliefs); (iii.) Actions. The key tenet of this architecture is that the agent’s behaviour is the outcome of a rational balance among different mental states. Previous seminal works on the BDI paradigm \cite{2,3}, or implementing the BDI architecture in Defeasible Logic \cite{5,7} have assumed that mental states are primitive and independent from each other, even though some mutual influences are considered (e.g., intentions are seen as desires satisfied up to commitment).

We work here on a different perspective to provide a fresh and efficient rule-based framework that considers goals, desires, and intentions as facets of the same phenomenon (all of them being goal-like attitudes): the notion of outcome, which is simply something an agent would like or is expected to achieve. An advantage of the proposed framework is that it allows agents to compute different degrees of motivational attitudes, and degrees of commitment that take into account other factors, such as beliefs and norms.

While different schemas for generating and filtering agents’ outcomes are possible, we will restrict ourself to schemas where we adopt the following principles:

– When an agent faces alternative outcomes in a given context, it is natural to rank them in a preference order;
– Beliefs prevail over conflicting motivational attitudes, thus avoiding various cases of wishful thinking \cite{8,9};
– Norms and obligations are used to filter social motivational states (social intentions) and compliant agents [9, 6];
– Goal-like attitudes can be derived via conversion using other mental states, such as beliefs (e.g., believing that Madrid is in Spain may imply that the goal to go to Madrid implies the goal to go to Spain) [6].

Consider the following example, Alice, during her holidays, plans to pay a visit to her friend John, who lives close to her parents. The plan can be described by the sentence

I shall come over to John’s place to visit him on Monday, but if he is not home or the visit is not possible, I am going to visit my parents. If this is not possible as well, I shall take some rest at home.

This idea can be easily implemented by building for each alternative a sequence of other alternatives $A_1, \ldots, A_n$ that are preferred when the first choice is no longer feasible. Normally, each set of alternatives is the result of a specific context $C$. This scenario can be represented as $C \Rightarrow A_1, \ldots, A_n$ which is closely related of contrary-to-duty obligations [10], where a norm is represented by an Obligation rule of the type:

$$r_1 : \text{drive}_\text{car} \Rightarrow \neg \text{damage} \odot \text{compensate} \odot \text{foreclosure}.$$  

Rule $r_1$ states that, if an agent drives a car, she has the obligation not to cause any damage to others; if this happens, she is obliged to compensate; if she fails to compensate, there is an obligation of foreclosure. The previous setting can be rewritten as:

$$r_2 : \text{holiday} \Rightarrow \cup \text{visit}_\text{friend} \odot \text{visit}_\text{parents} \odot \text{stay}_\text{home}.$$  

where $r_2$ is a rule introducing the oUtcome mode. In both examples, the sequences express a preference ordering among alternatives, which means that also stay_home and foreclosure, though not the best options, still correspond to acceptable situations.

Besides rules for outcomes and obligations, we also have rules for beliefs such as

$$r_3 : \text{friend}_\text{away} \Rightarrow B \neg \text{visit}_\text{friend}$$  

for which we assume there is no preference ordering, since they do not express expected outcomes but simply describe how the world is.

These building blocks allow us to introduce different types of goal-like attitudes and degrees of commitment to outcomes: desires, goals, intentions, and social intentions.

**Desires as Acceptable Outcomes.** Suppose an agent is equipped with the following outcome rules expressing two preference orderings:

$$r : a_1, \ldots, a_n \Rightarrow \cup b_1 \odot \cdots \odot b_m$$  
$$s : a'_1, \ldots, a'_n \Rightarrow \cup b'_1 \odot \cdots \odot b'_k$$  

and that the situation described by $a_1, \ldots, a_n$ and $a'_1, \ldots, a'_n$ are mutually compatible but $b_1$ and $b'_1$ are not, namely $b_1 = \neg b'_1$. In this case $b_1, \ldots, b_m, b'_1, \ldots, b'_k$ are anyway all acceptable outcomes, including the incompatible outcomes $b_1$ and $b'_1$. Desires are expected or acceptable outcomes, independently of whether they are compatible with other expected or acceptable outcomes.