Conflict-free normative agents using assumption-based argumentation

Dorian Gaertner and Francesca Toni

Department of Computing
Imperial College London, UK

Abstract. Argumentation can serve as a useful abstraction for various agent activities and in particular for agent reasoning. In this paper we further support this claim by mapping a form of normative BDI agents onto assumption-based argumentation. By way of this mapping we equip our agents with the capability of resolving conflicts amongst norms, beliefs, desires and intentions. This conflict resolution is achieved by using the agent's preferences, represented in a variety of formats. We illustrate the mapping with examples and use an existing computational tool for assumption-based argumentation, the CaSAPI system, to animate conflict resolution within our agents.

Keywords: norms, BDI agents, conflicts, argumentation

1 Introduction

Normative agents that are governed by social norms (e.g. [3, 5, 23]) may see conflicts arise amongst their individual desires, or beliefs, or intentions. These conflicts may be resolved by rendering information (such as norms, beliefs, desires and intentions) defeasible and by enforcing preferences [25]. In turn, argumentation has proved to be a useful technique for reasoning with defeasible information and preferences (e.g. see [16, 20]) when conflicts may arise.

In this paper we adopt a model for normative agents, whereby agents hold beliefs, desires and intentions, as in a conventional BDI model, but these mental attitudes are seen as contexts and the relationships amongst them are given by means of bridge rules (as in [19]). We understand norms as bridge rules, and adopt a norm representation that builds upon and extends the one given for the BDI+C agent model of [11]. This representation is natural in that norms typically concern different mental attitudes. Bridge rules also lend themselves to be naturally mapped onto argumentation frameworks, as we show in this paper.

The model for normative agents we adopt relies upon preferences over bridge rules being explicitly given, to be used to resolve conflicts, should they arise. We consider three kinds of representations for preferences: (i) by means of total orders over conflicting information; (ii) by means of partial orders over conflicting information; (iii) by dynamic rules that provide partial, domain-dependent definitions of preferences as in e.g. [16, 20].

We will refer to our agents as BDI+N agents.

For the detection and resolution of conflicts arising from choosing to adopt social norms, and for each form of preference representation, we use a specific form of argumentation, known as assumption-based argumentation [2, 7, 9, 12, 18]. This has been proven to be a powerful mechanism to understand commonalities and differences amongst many existing frameworks for non-monotonic reasoning [2], for legal reasoning [18], and for practical and epistemic reasoning [12]. Whereas abstract argumentation [6] focuses on arguments seen as primitive and atomic and attacks as generic relations between arguments, assumption-based argumentation sees arguments as deductions from assumptions in an underlying deductive system and defines attacks against arguments as deductions for the contrary of assumptions supporting those arguments.

Assumption-based argumentation frameworks can be coupled with a number of different semantics, all defined in dialectical terms and borrowed from abstract argumentation. Some of these semantics are credulous and some are sceptical, of various degrees. Different computational mechanisms can be defined to match the semantics, defined in terms of dialectical proof procedures, in particular, GB-dispute derivations [8] (computing the sceptical grounded semantics), AB-dispute derivations [7, 8] (computing the credulous admissible semantics) and IB-dispute derivations [8, 9] (computing the sceptical ideal semantics). All these procedures have been implemented within the CaSAPI system [12].

In this paper we provide a mapping from BDI+N agents onto assumption-based argumentation, and make use of the CaSAPI system to animate the agents and provide conflict-free beliefs, desires and intentions, upon which the commitments of the agents are based. The different procedures that CaSAPI implements provide a useful means to characterise different approaches that BDI+N agents may want to adopt in order to build these commitment stores.

The paper is organised as follows. Section 2 gives some background for and a preliminary definition of our BDI+N agents, focusing on the representation of norms. Section 3 gives some background on the form of argumentation we adopt and show how it can be used to detect and avoid conflicts. Section 4 presents our approach to modelling the agents' preferences (in terms of total orders, partial orders, and dynamic definitions) and using these preferences to resolve conflicts in assumption-based argumentation counterparts of our BDI+N agents. Section 5 concludes.

2 BDI+N Agents: Preliminaries

2.1 Background

Our BDI+N agents are an adaptation and extension of [11], which in turn builds upon [19]. The agent model therein adapts an architecture based on multi-context systems that have first been proposed in [14]. Individual theoretical components of an agent are modelled as separate *contexts*, each of which contains a set of statements in a language L_i together with the axioms A_i and inference rules Δ_i of a (modal) logic. A context i is hence a triple of the form:

$\langle L_i, A_i, \Delta_i \rangle$. Not only can new statements be deduced in each context using the deduction machinery of the associated logic, but these contexts are also interrelated via *bridge rules* that allow the deduction of a formula in one context based on the presence of certain formulae in other, linked contexts. An agent is then defined as a set of context indices $\mathcal{I}$, a function that maps these indices to contexts, another function that maps these indices to theories T_i (providing the initial set of formulae in each context), together with a set of bridge rules BR, namely rules of inference which relate formulae in different contexts. Thus, an agent can be given as follows:

$$Agent = \langle \mathcal{I},\ \mathcal{I} \rightarrow \langle L_i, A_i, \Delta_i \rangle,\ \mathcal{I} \rightarrow T_i,\ BR \rangle$$

The normative agents we are investigating are all extensions of the well-known BDI architecture of Rao and Georgeff [22] and hence the set of context indices $\mathcal{I}$ is {B, D, I}. Bridge rules in BR are inference rules that may be ground, non-ground, or partially instantiated.

An example taken from the Ten Commandments states *"You shall not covet your neighbour's wife"*. This can be expressed as a ground bridge rule as follows: [1]

$$\frac{B\ :\ (correct(bible))}{\neg D\ :\ (have(neighbours\_wife))}$$

An example of a non-ground schema is: for any X and Y, if an agent believes that $X \rightarrow Y$ and it desires Y, then the agent should intend X:

$$\frac{B\ :\ (X \rightarrow Y),\ D\ :\ Y}{I\ :\ X}$$

An example of a partially instantiated schema is that if one believes that Armageddon will strike immediately, then one should not desire anything:

$$\frac{B\ :\ immediately(armageddon)}{\neg D\ :\ X}$$

In the BDI+C agent model of [11] bridge rules in BR may be norms that are meant to feed into a commitment store. Commitments and norms are represented as follows [2], where $\lceil x \rceil$ is the codification of a norm x as a term in Gödel's sense:

[1] In this paper we adopt the following Prolog-like convention: ground terms and predicates start with a lower-case letter and variables start with an upper-case letter.

[2] We ignore here the agent/institution component of *Commit* of [11], which identifies the protagonist and the subject of a commitment.

$$\begin{aligned}
Commitment &::= Commit(\lceil Norm \rceil)\\
Norm &::= \varphi \Rightarrow \psi\\
\varphi &::= ConjLiterals\\
ConjLiterals &::= MLiteral \mid MLiteral \wedge ConjLiterals\\
\psi &::= MLiteral\\
MLiteral &::= MentalAtom \mid \neg MentalAtom\\
MentalAtom &::= \mathtt{B}(term) \mid \mathtt{D}(term) \mid \mathtt{I}(term)
\end{aligned}$$

Example norms are:

$$\mathtt{B}(correct(quran)) \Rightarrow \mathtt{D}(goto(mecca))$$

expressing the fact that if an agent believes that the Qu'ran is correct, then he should desire to go to Mecca; or

$$\mathtt{B}(correct(bible)) \Rightarrow \neg\mathtt{D}(have(neighbours\_wife))$$

stating that an agent believing in the correctness of the Bible, should *not* desire (to have) his neighbour's wife. Note that the representation of norms as inference rules and their representation using ⇒ are equivalent, and indeed the mapping that we will introduce in Section 3 will treat the two representations in exactly the same way.

Similarly to inference rules, norms can be ground, partially instantiated and schemata.

Agents communicate with one another (and potentially sense their environment) and by doing so might need to update their beliefs. New beliefs can trigger a norm (possibly by instantiating a norm schema) and subsequently, a new belief, desire or intention could be adopted by the agent. This new mental attitude however, may be in conflict with existing beliefs, desires or intentions, and thus the commitment store may be inconsistent. BDI+N agents provide a solution to the problem of resolving these conflicts, as we will see later on in the paper.

2.2 Norm Representation in BDI+N Agents

In our BDI+N agents, bridge rules in BR include norms represented in an extension of the norm language given earlier. In particular, we distinguish between two kinds of terms: actions that an agent can execute, such as $goto(mecca)$, are called *action terms*; properties that cannot be executed, such as $correct(bible)$ and $have(neighbours\_wife)$, are called *state terms*. Note that state terms can be brought about by executing actions.

This distinction leads to a refinement of the BNF definition for a mental atom. We modify the commitment and norm representation underlying BDI+C agents [11], so that executable actions are distinguished from properties. Moreover, we allow beliefs to be implications (of the form $B(X \rightarrow Y)$).

$$\begin{aligned} \mathit{MentalAtom} &::= \mathtt{B}(\mathit{stateterm}) \\ &\quad \mid \mathtt{B}(\mathit{Eitherterm} \rightarrow \mathit{Eitherterm}) \\ &\quad \mid \mathtt{D}(\mathit{Eitherterm}) \mid \mathtt{I}(\mathit{actionterm}) \\ \mathit{Eitherterm} &::= \mathit{actionterm} \mid \mathit{stateterm} \end{aligned}$$

We restrict intentions to only concern action terms, since, intuitively, an intention is always about some future behaviour. For example, a human cannot intend to have his neighbour's wife: he can desire it, but this only results in a different intention (e.g. to leave his wife which in turn is an action). Beliefs are restricted to concern state terms, since one cannot believe an action. The only other argument to the belief modality is an implication which can have either state or action terms both as the antecedent and the consequent. Examples of these kinds of beliefs are: $\mathtt{B}(\mathit{sunny} \rightarrow \mathit{stays\_dry}(\mathit{grass}))$ or $\mathtt{B}(\mathit{goto}(\mathit{mecca}) \rightarrow \mathit{goto}(\mathit{heaven}))$.

Finally, note that we do not allow negative terms of either kind. So, for example, we cannot represent directly $\mathtt{B}(\mathit{rainy} \rightarrow \neg \mathit{stays\_dry}(\mathit{grass}))$. However, this belief can be expressed equivalently as $\mathtt{B}(\mathit{raining} \rightarrow \mathit{not\_stays\_dry}(\mathit{grass}))$. This is not restrictive as, in the assumption-based argumentation framework we will adopt, the relationship between $\mathit{not\_stays\_dry}(X)$ and $\mathit{stays\_dry}(X)$ can be expressed by introducing appropriate assumptions a and b intuitively representing $\mathit{not\_stays\_dry}(X)$ and $\mathit{stays\_dry}(X)$, respectively, and by setting the contrary of $\mathit{not\_stays\_dry}(X)$ to b and the contrary of $\mathit{stays\_dry}(X)$ to a.

2.3 Example

For illustrative purposes, throughout the paper we use an example employing agents from the ballroom scenario described in [10]. We consider a single dancer agent at a traditional ballroom. This dancer can be represented as an agent

$$\langle \mathcal{I} = \{B, D, I\},\ \mathcal{I} \rightarrow \langle L_i, A_i, \Delta_i \rangle,\ \mathcal{I} \rightarrow T_i,\ BR \rangle$$

with bridge rules in BR including

$$\frac{B\ :\ (X \rightarrow Y), D\ :\ (Y)}{I\ :\ (X)} \quad (\mathit{if\ X\ is\ an\ actionterm}) \tag{1}$$

$$\frac{B\ :\ (X \rightarrow Y), D\ :\ (Y)}{D\ :\ (X)} \quad (\mathit{if\ X\ is\ a\ stateterm}) \tag{2}$$

$$\frac{D\ :\ (X)}{I\ :\ (X)} \quad (\mathit{if\ X\ is\ an\ actionterm}) \tag{3}$$

and inference rules in Δ_B:

$$\frac{B(X \rightarrow Y) \wedge B(X)}{B(Y)} \quad (\mathit{modus\ ponens\ for\ B}) \tag{4}$$

Note that (4) corresponds to the modal logic schema K for beliefs, but is not present for desires and intentions since implications can be believed but neither desired nor intended. Furthermore, we do not have positive or negative introspection (modal logic schemata 4 and 5) since in this paper we ignore nested beliefs, desires and intentions for the sake of simplicity.

The bridge rules BR include also norms in the domain language of the ballroom. Examples of these norms are:

$$B(attractive(X)) \Rightarrow D(danceWith(X)) \tag{5}$$

$$B(sameSex(X, self)) \Rightarrow \neg I(danceWith(X)) \tag{6}$$

$$B(thirsty(self)) \Rightarrow I(goto(bar)) \tag{7}$$

Finally, one needs to define the initial theories T_i of the agent, detailing its initial beliefs, desires and intentions. Let us assume we are modeling a dancer which is male, not thirsty and considers its friend and fellow dancer Bob to be attractive. Hence T_B includes $B(attractive(bob))$ and $B(sameSex(bob, self))$. Then, from the first belief, norm (5) and an instance of bridge rule (3), one can derive that our dancer should intend to dance with Bob. However, from the second belief and norm (6) one can derive the exact opposite, namely that our dancer should not intend to dance with Bob. We believe that this conflict is undesirable and intend to address the problem of resolving it.

3 Conflict Avoidance

In this Section we show how assumption-based argumentation can help to avoid conflicts, in the absence of any additional (preference) information that might help to resolve them.

3.1 Background

An assumption-based framework is a tuple $\langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \overline{} \rangle$ where

- $(\mathcal{L}, \mathcal{R})$ is a deductive system, with a language $\mathcal{L}$ and a set $\mathcal{R}$ of inference rules,
- $\mathcal{A} \subseteq \mathcal{L}$, is the set of candidate *assumptions*,
- $\overline{}$ is a (total) mapping from $\mathcal{A}$ into $\mathcal{L}$, where $\overline{\alpha}$ is the *contrary* of α.

We will assume that the inference rules have the syntax $c_0 \leftarrow c_1, \ldots c_n$ $(n > 0)$ or c_0, where $c_i \in \mathcal{L}$. As in [7], we will restrict attention to *flat* assumption-based frameworks, such that if $c \in \mathcal{A}$, then there exists no inference rule of the form $c \leftarrow c_1, \ldots, c_n \in \mathcal{R}$.

Example 1. $\mathcal{L} = \{p, a, \neg a, b, \neg b\}$, $\mathcal{R} = \{p \leftarrow a; \quad \neg a \leftarrow b; \quad \neg b \leftarrow a\}$, $\mathcal{A} = \{a, b\}$ and $\overline{a} = \neg a$, $\overline{b} = \neg b$.

An *argument* in favour of a sentence x in $\mathcal{L}$ supported by a set of assumptions X is a backward deduction from x to X, obtained by applying backwards the rules in $\mathcal{R}$. For the simple assumption-based framework above, an argument in favour of p supported by $\{a\}$ may be obtained by applying $p \leftarrow a$ backwards.

In order to determine whether a conclusion is to be sanctioned, a set of assumptions need to be identified that would provide an "acceptable" support for the belief, namely a "consistent" set of assumptions including a "core" support as well as assumptions that defend it. This informal definition can be formalised in many ways, using a notion of "attack" amongst sets of assumptions whereby X *attacks* Y iff there is an argument in favour of some $\overline{x}$ supported by (a subset of) X where x is in Y. In example 1 above, $\{b\}$ attacks $\{a\}$.

Possible formalisations of "acceptable" support are: a set of assumptions is

- *admissible*, iff it does not attack itself and it counter-attacks every set of assumptions attacking it;
- *complete*, iff it is admissible and it contains all assumptions it can defend, by counter-attacking all attacks against them;
- *grounded*, iff it is minimally complete;
- *ideal*, iff it is admissible and contained in all maximally admissible sets.

Most of these formalisations are matched by computational mechanisms [7, 9, 8], defined as disputes between two fictional players within an agent mind: the proponent and the opponent, trying to establish the "acceptability" of a given statement/belief with respect to the chosen formalisation of support. The three mechanisms are GB-dispute derivations, for grounded support, AB-dispute derivations, for admissible support, and IB-derivations, for ideal support. Like the formalisations they implement, these mechanisms differ in the level of scepticism of the proponent player:

- in GB-dispute derivations the proponent is prepared to take no chance and is completely sceptical in the presence of seemingly equivalent alternatives;
- in AB-dispute derivations the proponent would adopt any alternative that is capable of counter-attacking all attacks without attacking itself;
- in IB-dispute derivations, the proponent is wary of alternatives, but is prepared to accept common ground between them.

The three procedures are implemented within the CaSAPI system for argumentation [12].

In order to employ assumption-based argumentation to avoid (and resolve, in Section 4) conflicts, one has to provide a mapping from the agent representation introduced in Section 2 onto the assumption-based argumentation framework and choose a suitable semantics. Given such a mapping, one can then run CaSAPI, the argumentation tool, and hence *reason on demand* about a given conclusion.

3.2 Naive Translation into Assumption-Based Argumentation

In our proposed translation, we see all bridge rules BR, theories T_i, axioms A_i and inference rules Δ_i as inference rules in an appropriate assumption-based framework. For instance, consider the ballroom example. This can be translated into the following assumption-based framework:

$$\mathcal{L} = L_B \cup L_D \cup L_I$$

$$\mathcal{A} = \emptyset$$

$$\mathcal{R} = \left\{ \begin{array}{l} I(X) \leftarrow B(X \rightarrow Y), D(Y), actionterm(X) \\ D(X) \leftarrow B(X \rightarrow Y), D(Y), stateterm(X) \\ B(Y) \leftarrow B(X \rightarrow Y), B(X) \\ I(X) \leftarrow D(X) \\ D(danceWith(X)) \leftarrow B(attractive(X)) \\ \neg I(danceWith(X)) \leftarrow B(sameSex(X, self)) \\ I(goto(bar)) \leftarrow B(thirsty(self)) \\ B(attactive(bob)) \\ B(sameSex(bob, self) \\ actionterm(danceWith(X)) \\ stateterm(attractive(X)) \\ stateterm(sameSex(X, Y)) \end{array} \right\}$$

where a definition for $\overline{}$ is not required, since the set of assumptions $\mathcal{A}$ is empty. Note that in assumption-based argumentation all inference rules in $\mathcal{R}$ are ground [3] and indeed the non-ground inference rules above are a short-hand for the set of all their ground instances.

Having constructed an instance of assumption-based argumentation in this way, one can now use the mechanisms described in [8] and implemented in the CaSAPI system [12] to determine whether a given statement/belief is supported, and, if so, by which assumptions/arguments.

This translation is naive in that the resulting framework justifies (and CaSAPI succeeds with) the conflicting conclusions

$$I(danceWith(bob)) \text{ and } \neg I(danceWith(bob))$$

simultaneously. In the absence of any additional information, these conflicts can be avoided by introducing assumptions in $\mathcal{A}$, as we will see below. [4]

3.3 Avoiding Conflicts using Assumption-Based Argumentation

The conflict between $I(danceWith(bob))$ and $\neg I(danceWith(bob))$ as the conclusions of two rules can be avoided by rendering the application of the two rules mutually exclusive. This can be achieved by attaching assumptions to these rules

[3] However, the computational mechanisms in [7, 9, 8] and CaSAPI [12] can also handle variables in inference rules.

[4] Note that assumption-based argumentation frameworks with an empty set of assumptions $\mathcal{A}$, such as the one given earlier, only allows the construction of arguments that are straight deductions from $\mathcal{R}$ alone.

and setting the contrary of the assumption associated to any rule to be the conclusion of the other rule. This would correspond to rendering the corresponding norms/inference rules defeasible. Assumptions attached in such a manner can be considered as *rule applicability* predicates.

In the ballroom example, the fourth and sixth rules of the naive translation above are replaced by the following two rules

$$I(X) \leftarrow D(X), \alpha(X)$$

$$\neg I(danceWith(X)) \leftarrow B(sameSex(X, self)), \beta(danceWith(X))$$

where $\mathcal{A} = \{\alpha(t), \beta(t) \mid t \text{ is ground}\}$ and $\overline{\alpha(t)} = \neg I(t)$ and $\overline{\beta(t)} = I(t)$.

Within the revised assumption-based framework, the conflicting conclusions $I(danceWith(bob))$ and $\neg I(danceWith(bob))$ can no longer be supported simultaneously. However, adopting the notion of admissible support (implemented as AB-derivations), $I(danceWith(bob))$ and $\neg I(danceWith(bob))$ can be supported separately, in a credulous manner. On the other hand, adopting the notions of grounded or ideal support (and GB- or IB-derivations), neither $I(danceWith(bob))$ nor $\neg I(danceWith(bob))$ can be supported, sceptically.

Overall, the translation into assumption-based argumentation given in this Section allows to avoid conflicts, but not resolve them. Below, we show how to resolve conflicts in the presence of additional information, in the form of preferences over norms/inference rules.

4 Conflict Resolution using Preferences

In this Section we show how to use assumption-based argumentation in order to reason normatively and resolve, with the help of preferences, conflicts that come about by accepting or committing to certain norms, beliefs, desires or intentions. As in the previous Section, we will need for norms and mental atoms to be defeasible, and will achieve this defeasibility by using assumptions that prevent an agent from both intending and not intending the same thing simultaneously. Similarly, we will not allow any agent to simultaneously believing and not believing or desiring and not desiring the same thing.

Using preferences, we can, for example, prioritise certain beliefs over a norm or certain norms over desires. Thus, one can think of preferences as the *normative personality* of an agent. For instance, in the ballroom example, an agent who values norms (3) and (5) more than norm (6) will indeed intend to dance with Bob, whereas another agent, who values social conformance or norm (6) more highly, will not have such an intention.

We will adopt the following revised agent model:

$$Agent = \langle \mathcal{I},\ \mathcal{I} \rightarrow \langle L_i, A_i, \Delta_i \rangle,\ \mathcal{I} \rightarrow T_i,\ BR,\ \mathcal{P} \rangle$$

where the new component $\mathcal{P}$ expresses the agent's preferences. We will consider various representations for $\mathcal{P}$ below, and provide a way to use them to resolve conflicts by means of assumption-based argumentation. Concretely, we will start with a total ordering and a cluster-based translation for conflict-resolution. Then we add more flexibility by allowing the order to be partial. Finally, we suggest a way of defining preferences using meta-rules, e.g. as done by [16, 20], and following the approach proposed in [18].

4.1 Preferences as a Total Ordering

The preference information $\mathcal{P}$ can be expressed as a function that provides a mapping from bridge rules and elements of theories/axioms/inference rules to rational numbers. For now, let us assume that $\mathcal{P}$ provides a total ordering and that the type of $\mathcal{P}$ is

$$BR \cup A_B \cup A_D \cup A_I \cup \varDelta_B \cup \varDelta_D \cup \varDelta_I \cup T_B \cup T_D \cup T_I \to \mathbb{Q}.$$

We will assume that lower numbers indicate a higher preference for the piece of information in question.

In order to translate the preferences $\mathcal{P}$ of an agent into a form that assumption-based argumentation can handle, we propose the following mechanism. First, we assume a translation into $\mathcal{R}$ as given in Section 3.3. Then, all rules in $\mathcal{R}$ are clustered according to their conclusion. Rules in the same cluster all have the same mental atom in their conclusion literal (so that fellow cluster members have either exactly the same or exactly the opposite conclusion). Next, each cluster of rules is considered in turn. All elements of one cluster are sorted in ascending order $\pi_1, \ldots, \pi_n$ by the preference of their corresponding norm, belief etc. Without loss of generality, let us assume that each rule π_i is of the form $l_i \leftarrow r_i$ where l_1 is the left-hand side of the most important rule and r_n is the right-hand side of the least important rule:

$$l_1 \leftarrow r_1 \qquad l_2 \leftarrow r_2 \qquad l_3 \leftarrow r_3 \qquad \ldots \qquad l_n \leftarrow r_n$$

Remember, that each r_i can be a singleton, a conjunction of mental atoms or the empty set and that a rule π_j is more important than rule π_k if and only if $\mathcal{P}(\pi_j) < \mathcal{P}(\pi_k)$.

Then, we add a new assumption p_i to the right-hand side of each rule (with the exception of the most important rule). By introducing assumptions into rules we make these rules defeasible. We do not add an assumption to the first rule since this is not defeasible. The resulting set of inference rules is:

$$l_1 \leftarrow r_1 \qquad l_2 \leftarrow r_2, p_2 \qquad l_3 \leftarrow r_3, p_3 \qquad \ldots \qquad l_n \leftarrow r_n, p_n$$

where each p_i is a new assumption. Then, by appropriately defining contraries, we can render conflicts impossible. Concretely, we further add rules for new sentences q_i of the form:

$$\begin{array}{lllll}
q_2 \leftarrow r_1 & q_3 \leftarrow r_2, p_2 & q_4 \leftarrow r_3, p_3 & \cdots & q_n \leftarrow r_{n-1}, p_{n-1} \\
 & q_3 \leftarrow q_2 & q_4 \leftarrow q_3 & \cdots & q_n \leftarrow q_{n-1} \\
 & & q_4 \leftarrow q_2 & \cdots & q_n \leftarrow q_{n-2} \\
 & & & \cdots & \cdots \\
 & & & & q_n \leftarrow q_2
\end{array}$$

Intuitively, q_{i+1} holds if π_i is "selected" (by assuming p_i) and applicable (by r_i holding) or any of the other more important (earlier) rules is selected and applicable. One can now define the contraries of each of the assumptions p_i in such a way as to allow norms with a smaller subscript to override norms with higher subscripts. If we set $\overline{p_i} = q_i$ for all $i > 1$, assumption p_i can be made (and thus rule π_i is applicable) only if q_i cannot be shown. The only way for q_i to hold is when both r_{i-1} and p_{i-1} hold (this would also make rule π_{i-1} applicable) or any of the other more important rules is applicable. Hence π_i is only applicable if π_j is not applicable for any j smaller than i.

After applying this procedure to all clusters, none of the clusters of rules can give rise to conflicts and since rules in different clusters have different conclusions, there cannot be any inter-cluster conflicts either. Hence, the resulting assumption-based framework $\langle \mathcal{L}', \mathcal{R}', \mathcal{A}', \overline{}' \rangle$ is conflict-free, where:

$$\begin{aligned}
&\mathcal{L}' = \mathcal{L} \ \cup\ \textstyle\bigcup_{i=1\ldots n}\{p_i, q_i\} \\
&\mathcal{R}' = \{l_i \leftarrow r_i, p_i \mid (l_i \leftarrow r_i) \in \mathcal{R}\} \ \cup\ \{q_{i+1} \leftarrow r_i, p_i \mid (l_i \leftarrow r_i) \in \mathcal{R}\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \cup\ \{q_i \leftarrow q_j \mid 1 < j < i\} \\
&\mathcal{A}' = \mathcal{A} \ \cup\ \textstyle\bigcup_{i=2\ldots n}\{p_i\} \\
&\forall p_i \in \mathcal{A}' : \overline{p_i} = q_i
\end{aligned}$$

Let us consider the ballroom example from Section 2.3 again. Assume that the most important norm is (5) - $B(attractive(X)) \Rightarrow D(danceWith(X))$ followed by norm (6) - $B(sameSex(X, self)) \Rightarrow \neg I(danceWith(X))$ and norm schema (4) - $D(X) \Rightarrow I(X)$. Assume further that the premises of both norms (5) and (6) are fulfilled, unifying X with *bob*[5]. Using norm (5) we derive $D(danceWith(bob))$. Now, only norm (6) and norm schema (4) have conflicting conclusions and are grouped together for the purpose of conflict resolution. In this example, we assumed that norm (6) is more important than norm schema (4) and hence we get a cluster:

$$\begin{aligned}
&\neg I(danceWith(bob)) \leftarrow B(sameSex(bob, self)) \\
&I(danceWith(bob)) \leftarrow D(danceWith(bob)), p_2 \\
&q_2 \leftarrow B(sameSex(bob, self))
\end{aligned}$$

and since p_2 is the only assumption, we only have to set one contrary, namely:

$$\overline{p_2} = q_2$$

Now the mental literal $\neg I(danceWith(X))$ will be supported, but its complementary literal will not. Note that norm (7) that stated that thirsty dancers should go to the bar does not play a part in resolving the present conflict. One may therefore argue that the requirement of having a total preference order of rules is an unnatural one. Indeed, one may not need to be able to express a pref-

[5] Norm schemata are instantiated at this stage.

erence between certain rules that are unrelated (i.e. concerned with different, non-conflicting conclusions).

Note further, that we are adopting the *last-link* principle [20] in using preferences for resolving conflicts, which uses the strength of the last rule used to derive the argument's claim for comparison. According to this principle, the fact that norm schema (4) is based on a desire derived using the most important norm is irrelevant.

Once the mapping has been formulated, reasoning with the original framework is mapped onto reasoning with an assumption-based argumentation framework. Alternative semantics are available (and implemented in CaSAPI) to compute whether a given claim is supported.

4.2 Preferences as a Partial Ordering

We propose a different representation for preferences if the ordering of rules is not total. We replace the function $\mathcal{P}$ with a binary relation $\mathcal{P}$ which holds facts of the form $pref(\pi_i, \pi_j)$ that express the agent's preference of rule π_i over rule π_j [6]. Intuitively, $pref(\pi_i, \pi_j)$ means that rule π_i is more preferred to π_j. In this section, we further stipulate that $\mathcal{P}$ contains only facts about pairs of rules whose conclusions are conflicting. We deem it unnecessary to express preferences between rules that do not conflict since they will never be part of the same cluster. We will assume that this relation $\mathcal{P}$ is irreflexive and asymmetric. It may or may not be transitive depending on the intended use (there are convincing arguments in favour and against transitivity in the literature and we leave this choice open). It may also be appropriate to assume that $pref$ is not cyclic. The asymmetry and irreflexivity requirements can be expressed as follows:

$$\begin{array}{lcl} \bot & \leftarrow & pref(\pi_1, \pi_2) \wedge pref(\pi_2, \pi_1) \wedge \pi_1 \neq \pi_2 \\ \bot & \leftarrow & pref(\pi_1, \pi_1) \end{array}$$

for an appropriate notion of inconsistency $\bot$. We refrain from axiomatising the $pref$ relation since the notion of $\bot$ does not exist in our framework. An appropriate extension is part of our future research.

Let us assume again that rules in $\mathcal{R}$ can be clustered, as in the previous Section. If no preference is specified for some conflicting rules in the same cluster, namely they have the same importance, then any of these rules can be applied, i.e. the agent has no preference of one over the other. For example, let us consider two rules π_1 and π_2 in the same cluster, of the form $l_1 \leftarrow r_1$ and $l_2 \leftarrow r_2$ respectively, where l_1 and l_2 are in conflict (i.e. opposite mental literals), but neither is preferred. We follow the same mechanism as before and add two assumptions to the rules, yielding:

$$l_1 \leftarrow r_1, p_1 \qquad l_2 \leftarrow r_2, p_2$$

[6] Here, we assume a naming convention for rules whereby π_i is the name of rule $l_i \leftarrow r_i$.

However, instead of introducing new q_i terms in order to define the contraries of the assumptions p_i, we directly set: $\overline{p_1} = l_2$ and $\overline{p_2} = l_1$. In this way, each rule is only applicable if the other one is not. This is analogous to what we have done for avoiding conflicts in Section 3.3. Below, we will focus on rules for which $pref$ facts are specified in the clauster.

Elements of one cluster are no longer sorted by their quantitative preference. Instead, each rule π of the form $l \leftarrow r$ is rewritten as:

$l \leftarrow r, p$
$q \leftarrow r_1, pref(\pi_1, \pi)$
$\ldots$
$q \leftarrow r_n, pref(\pi_n, \pi)$

where p is a new assumption, and $\pi_1 : l' \leftarrow r_1 \ldots \pi_n : l' \leftarrow r_n$ are all the rules in the cluster where l' is the complement of l. Finally, we set $\overline{p} = q$.

In order to illustrate this mapping, consider again the ballroom example, where rules are named $\pi_1, \ldots \pi_{12}$ following the order in which they are presented in Section 3.2. If $pref(\pi_6, \pi_4) \in \mathcal{P}$ then in the resulting assumption-based framework, a subset of the set of inference rules is:

$I(X) \leftarrow D(X), p_4(X).$
$q_4(X) \leftarrow B(sameSex(X, self)), pref(\pi_6, \pi_4).$

$\neg I(danceWith(X)) \leftarrow B(sameSex(X, self)), p_6(X).$
$q_6(X) \leftarrow D(X), pref(\pi_4, \pi_6).$

$pref(\pi_6, \pi_4).$

The first rule applies only if $D(X)$ and $p_4(X)$ both hold. However, it is defeated by the fact that the contrary of $p_4(X)$ holds. This contrary $(q_4(X))$ is dependent on $pref(\pi_6, \pi_4)$, which is true in this example. Similarly, the rule with the conclusion $\neg I(danceWith(X))$ applies, only if both $B(sameSex(X, self))$ and $p_6(X)$ hold. In our example, this rule is not defeated, since the contrary of $p_6(X)$ cannot be shown. This contrary depends on $pref(\pi_4, \pi_6)$, which does not hold. It can hence be seen how the content of $\mathcal{P}$ influences the applicability of rules.

4.3 Defining Dynamic Preferences via Meta-rules

The relation $\mathcal{P}$ described in the previous subsection holds simple facts. One can easily extend these facts into rules [7] by adding extra conditions. For example, one could replace the fact $pref(\pi_1, \pi_2)$ with two meta-rules one stating

[7] Note that these are meta-rules only concerning preferences predicates and should not be confused with the rules that act as arguments to these preference predicates.

$pref(\pi_1, \pi_2) \leftarrow sunny$ and another one stating $pref(\pi_2, \pi_1) \leftarrow rainy$. This allows the agent to change the preference between two norms depending on the weather.

The addition of conditions to determine preferences makes the applicability of a certain norm dependent on the fulfilment of the conditions and hence allows more fine-grained control over arguments. The transformation defined for preferences as partial orders still applies here.

Note that one can view these meta-rules themselves as norms along the lines of "one should prefer norm 1 over norm 2 whenever the sun shines".

So far we have consider "logical" conflicts, obtained by deriving a literal and its complement. However, assumption-based argumentation allows to consider kinds of conflict, for example allowing to contrasts $goto(bar)$ with $danceWith(X)$, on the basis that nobody can go to the bar and be on the dance-floor at the same time. Imagine the possibility of such a conflict. Then, norm (7), referring to thirsty dancers, conflicts with an instance of norm schemata (4), that refers to dance intentions. A dancer that considers himself a "gentleman" then might prefer π_4 over π_7, resisting the temptation to go for a drink. A selfish dancer on the other hand may prefer π_7 over π_4. An agent considering itself a "gentleman" can be a dynamic notion, that can change, e.g. once the dancer has been to the bar a few times. Considering the meta-rules themselves as norms opens up many potential future investigations that we are looking forward to conduct.

5 Conclusions

In this paper we have proposed to use assumption-based argumentation to solve conflicts that a normative agent can encounter, arising from applying conflicting norms but also due to conflicting beliefs, desires and intentions. We have employed qualitative preferences over an agent's beliefs, desires and intentions and over the norms it is subjected to in order to resolve conflicts.

We have provided a translation from the agent definition to assumption-based argumentation, thus paving the way to deploying the working implementation of the argumentation system CaSAPI for concretely resolving the conflicts. Indeed, after manually applying the translation described in this paper, one can execute CaSAPI and obtain a supporting argument supporting the given claim, if successul. From these, one can derive which rules (norms or mental atoms) have been relied upon during the argumentation process. It would be useful to embed the implementation of this translation into the CaSAPI system.

Normative conflicts have previously been addressed from a legal reasoning perspective by Sartor [25] and from a practical reasoning point of view by Kollingbaum and Norman [17]. It is traditional in the legal domain to order laws hierarchically, using criteria such as source, chronology and speciality. One such system by Garcia-Camino et al. [13] uses these criteria and a meta-order over them to solve conflicts in compound activities. As far as we know, argumentation and in particular assumption-based argumentation, has received little attention in the agent community with respect to normative conflicts.

Argumentation-based negotiation (see for example [21]) is a field of artificial intelligence that concerns itself with resolving conflicts in a multi-agent society. However, to the best of our knowledge, the only architecture for individual agents that uses argumentation is the KGP model [15] that follows the approach of [16] to support its control component and its goal decision capability. The KGP model has been extended to support normative reasoning [24] but no conflict resolution amongst the outcomes of norm enforcement and beliefs is performed in this extension.

We have adopted a last-link approach to dealing with preferences in deriving conflicting conclusions [20]. An alternative from the standard literature is the principle of the weakest link [1] which compares the minimum strength of the sentences used in each argument. Furthermore, we plan to research the effects of splitting the preference function into four separate ones for beliefs, desires, intentions and norms. One may be able to draw conclusions about the kind of normative personality an agent possesses depending on how these individual preference functions relate. Such relationships have been used quantitatively by Casali et al. [4] in their work on graded BDI agents.

Acknowledgements

This research was partially funded by the Sixth Framework IST programme of the EC, under the 035200 ARGUGRID project. The second author has also been supported by a UK Royal Academy of Engineering/Leverhulme Trust senior fellowship.

References

1. L. Amgoud and C. Cayrol. Inferring from inconsistency in preference-based argumentation frameworks. *J. Autom. Reason.*, 29(2):125–169, 2002.
2. A. Bondarenko, P. Dung, R. Kowalski, and F. Toni. An abstract, argumentation-theoretic framework for default reasoning. *Art. Intelligence*, 93(1-2):63–101, 1997.
3. J. Broersen, M. Dastani, J. Hulstijn, Z. Huang, and L. van der Torre. The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In *Proceedings of AGENTS '01*, pages 9–16, New York, NY, USA, 2001. ACM Press.
4. A. Casali, L. Godo, and C. Sierra. Graded bdi models for agent architectures. *Lecture Notes In Artificial Inteligence*, 3847:126–143, 205.
5. F. Dignum, D. Morley, E. Sonenberg, and L. Cavendon. Towards socially sophisticated BDI agents. In *Proceedings of ICMAS '00*, pages 111–118, Washington, DC, USA, 2000. IEEE Computer Society.
6. P. Dung. The acceptability of arguments and its fundamental role in non-monotonic reasoning and logic programming and n-person game. *Artif. Intell.*, 77, 1995.
7. P. Dung, R. Kowalski, and F. Toni. Dialectic proof procedures for assumption-based, admissible argumentation. *Artificial Intelligence*, 170:114–159, 2006.
8. P. Dung, P. Mancarella, and F. Toni. Computing ideal sceptical argumentation. Technical report, Imperial College London, 2006.
9. P. Dung, P. Mancarella, and F. Toni. A dialectic procedure for sceptical, assumption-based argumentation. In *COMMA 06*, 2006.

10. D. Gaertner, K. Clark, and M. Sergot. Ballroom etiquette: a case study for norm-governed multi-agent systems. In *Proceedings of the 1st International Workshop on Coordination, Organisation, Institutions and Norms*, 2006.
11. D. Gaertner, P. Noriega, and C. Sierra. Extending the BDI architecture with commitments. In *Proceedings of the 9th International Conference of the Catalan Association of Artificial Intelligence*, 2006.
12. D. Gaertner and F. Toni. A credulous and sceptical argumentation system (`www.doc.ic.ac.uk/~dg00/casapi.html`). In *Proceedings of ArgNMR*, 2007.
13. A. García, P. Noriega, and J.-A. Rodríguez-Aguilar. An Algorithm for Conflict Resolution in Regulated Compound Activities. In *ESAW workshop*, 2006.
14. F. Giunchiglia and L. Serafini. Multi-language hierarchical logics or: How we can do without modal logics. *Artificial Intelligence*, 65(1):29–70, 1994.
15. A. Kakas, P. Mancarella, F. Sadri, K. Stathis, and F. Toni. The KGP model of agency. In *Proceedings of the European Conference on Artificial Intelligence*, pages 33 – 37, August 2004.
16. A. Kakas and P. Moraitis. Argumentation based decision making for autonomous agents. In *Proceedings of AAMAS '03*, pages 883–890, 2003.
17. M. Kollingbaum and T. Norman. Strategies for resolving norm conflict in practical reasoning. In *ECAI Workshop Coordination in Emergent Agent Societies*, 2004.
18. R. A. Kowalski and F. Toni. Abstract argumentation. *Journal of AI and Law, Special Issue on Logical Models of Argumentation*, 4(3-4):275–296, 1996.
19. S. Parsons, C. Sierra, and N. Jennings. Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, 8(3):261–292, 1998.
20. H. Prakken and G. Sartor. Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-Classical Logics*, 7(1):25–75, 1997.
21. I. Rahwan, S. Ramchurn, N. Jennings, P. McBurney, S. Parsons, and L. Sonenberg. Argumentation-based negotiation. *Knowledge Engineering Review*, 2004.
22. A. S. Rao and M. P. Georgeff. BDI-agents: from theory to practice. In *Proceedings of the First Intl. Conference on Multiagent Systems*, San Francisco, 1995.
23. F. Sadri, K. Stathis, and F. Toni. Normative KGP agents. *Computational and Mathematical Organization Theory*, 12(2/3):101–126, 2006.
24. F. Sadri, K. Stathis, and F. Toni. Normative kgp agents. *Computational & Mathematical Organization Theory*, 12(2-3), October 2006.
25. G. Sartor. Normative conflicts in legal reasoning. *Artificial Intelligence and Law*, 1(2-3):209–235, June 1992.