Agreeing on Institutional Goals for Multi-Agent Societies

D. Gaertner[1,2] J.A. Rodriguez[2] F. Toni[1]

[1]Dept. of Computing
Imperial College London
London SW7 2AZ
United Kingdom

[2]IIIA-CSIC
Campus de la UAB
08193 Bellaterra
Spain

Abstract. We present an argumentation-based approach to the problem of finding a set of institutional goals for multi-agent systems. The behaviour of the autonomous agents we consider in this paper is goal-directed, driven by either individual or common goals. When multiple agents want to set up a collaboration framework (for themselves or others to use), they do so by forming an institution (or organisation). The goals of such institution must be agreed upon by the agents setting up the framework before it can be executed.
We propose to employ argumentation, and in particular assumption-based argumentation, to facilitate the negotiation of institutional goals. We first describe a centralised approach and then provide the rationale for and detail our preliminary efforts at de-centralising the problem. We propose to use the argumentation system CaSAPI as a tool to reason about the collaborative goals of the institution. Our approach mitigates concerns about performance bottlenecks and vulnerability of the system while providing, to some extent, privacy to the individual members of the institution.

1 Introduction

One of the concerns of multi-agent systems (MAS) research is how to achieve certain collective properties despite individual agents' varying behaviour. Thus, there is a wealth of approaches that consider how to get agents (with individual preferences and goals) to interact in such a way that their interactions lead to the desired global properties. Along these lines, economic-based approaches (e.g. coalition formation [25] and mechanism design [8]), cooperation-based approaches (e.g. teamwork [27]), and organisation-based approaches (e.g. organisations [9, 15] and institutions [2]) provide MAS designers with techniques to enact MAS aimed at achieving their global goals. Notice that most approaches share the implicit assumption that there is some designer in charge of choosing the interaction mechanism that agents with individual goals use so that some global goals (or properties) are reached. Consider now that instead of a MAS designer, a group of agents gather together to decide by themselves the interaction

rules of a MAS where they, or other agents, are to operate in. In fact, what we envisage is that a subset of all agents agrees upon the rules and the goals that constitute a regulatory environment (an institution or virtual organisation, see e.g. [23]). Once such an agreement is reached, these agents may or may not be part of the MAS they agreed upon while other agents that were not part of the subset of agents that agreed upon the rules of the society may join said society.

One approach to structuring the interaction between agents recently proposed is the notion of Electronic Institutions [3]. There, speech acts are considered as actions and special institutional agents control what can be uttered. The construction of such an institution and required institutional agents begins with the definition of the goals the institution is meant to achieve. In this paper, we focus on agent institutions defined as software environments composed of autonomous agents that interact according to predefined conventions on language and protocol, guaranteeing that certain norms of behaviour are enforced. An electronic institution is in a sense a natural extension of the social concept of institutions as regulatory systems which shape human interactions.

We will not describe how such an institution is designed or executed (amongst others, Esteva et al. [2] provide information on tools for such purposes and Castelfranchi investigates social power [6]), or how norms and normative positions are handled (see [16] for information on this subject). Instead, we will focus on an earlier stage of the development of such institutions, namely on the question of how multiple agents can join efforts and agree on institutional goals. Such agents would still have individual goals, in addition to their common, institutional goals[1] for a particular collaboration. It is further worth noting that we begin with the supposition that the agents *want* to form an institution. Prior to entering into the argumentation process that we describe in this paper, they will need to explore whether or not they want to collaborate and form an institution at all. We assume that they have answered that question affirmatively.

We will describe two ways of constructing the set of common, institutional goals, both employing the CaSAPI tool [17] for assumption-based argumentation [5]. Firstly, a centralised approach is presented that combines the different goals of all agents and all their individual knowledge bases in the best possible way. Secondly, we detail an approach where each agent expresses its preference or rejection of a goal. A mechanism is then presented where participating agents use arguments based on their individual knowledge to defend their position.

This paper is structured as follows: Section 2 describes assumption-based argumentation as well as the CaSAPI tool and Section 3 introduces an example scenario. We then present the centralised approach in Section 4, show how CaSAPI can realise this approach (in the context of the scenario), discuss the issue of control and describe the disadvantages of the centralised approach before detailing preliminary work on a distributed argumentation mechanism to find common goals for the institution in Section 5. Finally, we look at related work and conclude.

[1] We will assume that the desired collaboration will take place in an institution and hence we will equate collaboration goals and institutional goals in what follows.

2 Assumption-based argumentation

This section provides the basic background on assumption-based argumentation (ABA) and the CaSAPI tool, see [5, 10, 12–14, 17, 18] for details.

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

- $(\mathcal{L}, \mathcal{R})$ is a *deductive system*, consisting of a language $\mathcal{L}$ and a set $\mathcal{R}$ of inference rules,
- $\mathcal{A} \subseteq \mathcal{L}$, referred to as the set of *assumptions*,
- $\overline{\ }$ is a (total) mapping from $\mathcal{A}$ into $\mathcal{L}$, where $\overline{x}$ is referred to as the *contrary* of x.

Intuitively, inference rules may be domain-specific or domain-independent [5]. They may correspond to causal information, to inference rules and axioms in a chosen logic-based language [5] or to laws and regulations [24]. Assumptions are sentences that can be questioned and disputed (as opposed to axioms that are beyond dispute), for example uncertain or unsupported beliefs or decisions. The contrary of an assumption, in general, stands for a reason why the assumption may be undermined (and thus may need to be be dropped).

We will assume that the inference rules in $\mathcal{R}$ have the syntax $l_0 \leftarrow l_1, \ldots l_n$ (for $n \geq 0$) where $l_i \in \mathcal{L}$. We will represent the rule $l \leftarrow$ simply as l. As in [12, 14, 10, 17, 18, 13], we will restrict attention to *flat* ABA frameworks, such that if $l \in \mathcal{A}$, then there exists no inference rule of the form $l \leftarrow l_1, \ldots, l_n \in \mathcal{R}$, for any $n \geq 0$. These frameworks are still quite general and admit many interesting instances [5]. Furthermore, we will adopt a generalisation of ABA frameworks, first given in [17], whereby assumptions allow *multiple contraries*, i.e.

- $\overline{\ }$ is a (total) mapping from $\mathcal{A}$ into $\wp(\mathcal{L})$.

As discussed in [13], multiple contraries are a useful generalisation to ease representation of ABA frameworks, but they do not really extend their expressive power.

Given an ABA framework, an *argument* in favour of a sentence $x \in \mathcal{L}$ supported by a set of assumptions X, denoted $X \vdash x$, is a (backward) deduction from x to X, obtained by applying backwards the rules in $\mathcal{R}$.

Example 1. Let us consider the following ABA framework $\langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \overline{\ } \rangle$ where:
$\mathcal{L} = \{p, a, \neg a, b, \neg b\}$,
$\mathcal{R} = \{p \leftarrow a; \neg a \leftarrow b; \neg b \leftarrow a\}$,
$\mathcal{A} = \{a, b\}$ and
$\overline{a} = \{\neg a\}$, $\overline{b} = \{\neg b\}$.
Then, an argument in favour of p supported by $\{a\}$ may be obtained by applying $p \leftarrow a$ backwards (the argument is $\{a\} \vdash p$).

In order to determine whether a conclusion (set of sentences) should be drawn, a set of assumptions needs to be identified providing an "acceptable" support for the conclusion. Various notions of "acceptable" support can be formalised, using a notion of "attack" amongst sets of assumptions whereby X_1

attacks X_2 iff there is an argument in favour of some $y \in \overline{x}$ supported by (a subset of) X_1 where x is in X_2 (for example, given $\langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \overline{\ } \rangle$ above, $\{b\} \vdash \neg a$ attacks $\{a\} \vdash p$). Then, a set of assumptions is deemed

- *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 (w.r.t. set inclusion) complete;
- *ideal*, iff it is admissible and contained in all maximally (w.r.t. set inclusion) admissible sets.

All these notions are possible formalisations of the notion of "acceptable" support for a conclusion. The first is a *credulous* notions, possibly sanctioning several alternative sets as "acceptable" supports, the latter two are *sceptical* notions, always sanctioning one single set as "acceptable" support. Different computational mechanisms can be defined to match these notions of "acceptable" support for given claims. The CaSAPI system that we propose to use in this paper (CaSAPI version 2 [17]) allows to compute the computational mechanisms of GB-dispute derivations for computing grounded supports [10], AB-dispute derivations for computing admissible supports [12, 10] and IB-dispute derivations for computing sceptical supports [10, 13]. In the case of example 1, there is an AB-dispute derivation for the claim p, computing the admissible support $\{a\}$. However, GB- and IB-dispute derivations fail to find grounded and ideal supports (respectively) for p, since indeed p cannot be sceptically supported (as $\{a\}$ and $\{b\}$ are "equally good" alternative sets of assumptions). If $\langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \overline{\ } \rangle$ in Example 1 is modified so that both c and $\neg c$ are added to $\mathcal{L}$ and the last inference rule in $\mathcal{R}$ is replaced by the two rules:

$\neg b \leftarrow c; \neg c$

with c an additional assumption and $\overline{c} = \{\neg c\}$, then there exist AB-, GB- and IB-dispute derivations for the claim p, all computing the support $\{a\}$ (which is now admissible, grounded and ideal).

Figure 1 illustrates how the rules, assumptions and contraries of the simple ABA framework modifying Example 1, as given earlier, are entered into CaSAPI using a graphical user interface. The user can enter a claim to be proved (p in what follows), select the type of dispute derivation it requires and control various other features of the computation such as the amount and format of the system's output. Once the input is entered, CaSAPI can be `Run` to determine whether of not the claim admits an "acceptable" support, according to the chosen type of dispute derivation. We will use CaSAPI with GB-dispute derivations only, as this semantics best fits the needs of our application (cf. Section 4 for details).

Note that, compared to conventional abstract argumentation [11], ABA addresses three problems: (i) how to find arguments, (ii) how to determine attacks and (iii) how to exploit the fact that different arguments may share premises. These problems are ignored by abstract argumentation, that sees arguments and attacks as "black-boxes". Instead, in ABA arguments are defined as *backward*

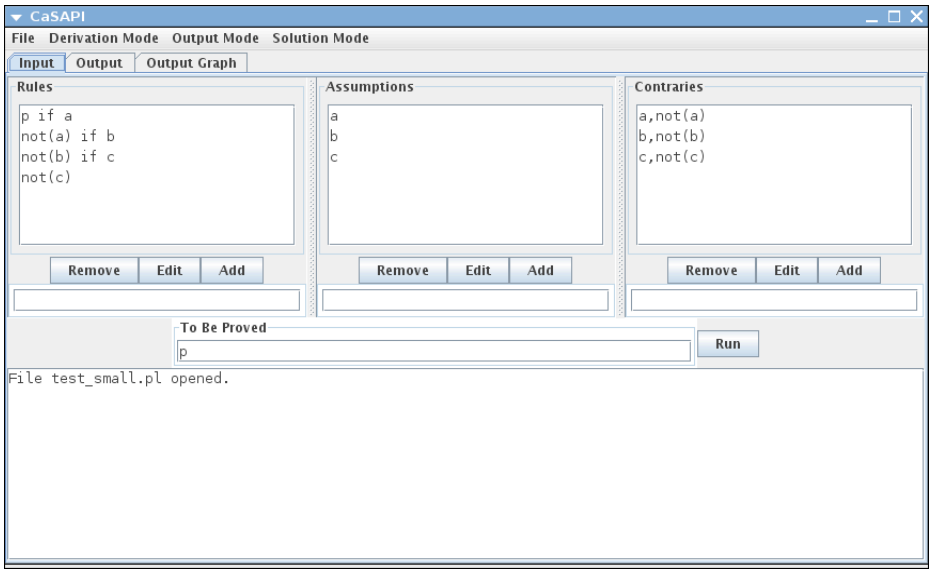

Fig. 1. Screenshot of the input process to CaSAPI

deductions (using sets of *rules* in an underlying logic) supported by sets of *assumptions*, and the notion of attack amongst arguments is reduced to that of *contrary* of assumptions. Moreover, (iii) is addressed by all forms of dispute-derivations (AB-, GB- and IB-) by employing various forms of "filtering steps" (for details, see [12, 10]). ABA can be seen as an instance of abstract argumentation, but is general-purpose nonetheless (in that, e.g., it can be instantiated in many ways to get many different frameworks for non-monotonic reasoning [5]). The relationship between ABA and abstract argumentation is detailed in [10] and exploited by versions 3 and 4 of CaSAPI [18, 13]: we ignore these versions here because they only support AB-dispute derivations and we chose the GB-dispute derivations for our application (cf. Section 4 for details).

3 Scenario

Reaching agreements in a society of self-interested agents is a problem related to the issue of cooperation. The capabilities of negotiation and argumentation are paramount to the ability of agents to reach agreements but as far as we know, neither negotiation-based nor argumentation-based approaches have been explored for the design of institutions. The automated design of (market) institutions has relied so far on mechanism design, namely on the design of protocols for governing multi-agent interactions that have certain desirable properties. However, the use of mechanism design for some (non-financial) institutions where there is a need for either justifying or changing negotiation stances is not sufficient.

Nobel laureate Ronald Coase in 1937 noted that in order to "make things", a collaboration is required and setting up such collaboration is costly. He used this insight to justify the existence of big companies which otherwise would be replaced by individuals operating in free and unregulated markets. In his recent book [28], Don Tapscott describes how in an increasingly connected world, the cost of collaborating is evaporating and the raison d'être of huge corporations ceases to exist. He provides the following example[2] to illustrate his claim:

> Take the Chinese motorcycle industry, which has tripled its output to 15m bikes per year over the past decade. There aren't really any Chinese equivalents of the big Japanese and American firms - Honda or Harley. Instead, there are hundreds of small firms [...]. Their representatives meet in tea-houses, or collaborate online, each sharing knowledge, and contributing the parts or services they do best. The companies that assemble the finished products don't hire the other companies; assembling the finished product is just another service. A "self-organised system of design and production" has emerged [...].

One can easily envisage that such ad-hoc collaborations will have a number of generic goals such as good communication between collaborators, sustainability of the collaboration (or the goal of achieving the objectives by a certain deadline), quality leadership and a high degree of connectedness and unity between the different participants. On a less generic dimension, business metrics can serve as goals for the institution, too. For example, profit increases, reduction in risk and rise of business value of the collaboration as a whole and/or of its participants can be considered collaboration goals. Finally, domain (i.e. collaboration) specific goals need to be considered such as the production of a certain amount of goods, the maintenance of a certain level of employee satisfaction, a certain throughput, a certain penalty on delivery delays and the return on capital employed.

Not all of these goals will be explicit goals of any given institution and the collaborators can decide which ones they value most and should strive to embed in their operational processes. If all collaborators have the same goals and they are all achievable (i.e. the set of goals is consistent and acceptable) then this set of goals will be the basis for the construction of the institution. However, in most cases, individual collaborators have conflicting ideas of the goals of the collaboration and as these goals must be shared by all collaborators, a mechanism is needed that reaches an agreement. We present a solution to this problem based on ABA.

In order to demonstrate our approach in this paper, we use a scenario of three agents named Adrian, Betty and Carles, working in the Chinese motorcycle industry and intending to institutionalise a collaboration for themselves to operate. They share a common language $\mathcal{L}$ in order to avoid ontological misunderstandings [3]. These three agents have some shared knowledge between them (stored

[2] The example is quoted from a review of the book which appeared in the Guardian newspaper on September 5th, 2007.

[3] In the future, ontology mapping methods (see e.g. [22]) can be used to align different languages.

in the shared knowledge base SKB) and some individual knowledge (stored in IKB_A, IKB_B and IKB_C, respectively). Each agent has some goals stored in its personal goal base (in this example, $GB_A = \{g_1, g_2, g_3\}$, $GB_B = \{g_2, g_4, g_5\}$ and $GB_C = \{g_1, g_6\}$) that it would like to see as the goals of the collaboration, and an individual vision (represented as rules in the corresponding IKB) of how these goals can be achieved (i.e. what is required for them to be reached). Each one wants its own goals to become common goals of the institution.

g_1: *to produce 100 motorbikes this week*
g_2: *to always clear the assembly line at the end of each day*
g_3: *to produce 100 sidecars this week*
g_4: *to improve/foster relations between the three collaborators*
g_5: *to make the institution sustainable/repeatable*
g_6: *to make Carles as the leader of this collaboration*

Whether or not a goal is achievable depends on a number of facts to hold true and a number of assumptions to be "assumable". The individual visions of how to achieve (relevant) goals are given in Table 1. Here the IKBs are represented as sets of rules of ABA frameworks. To ease understanding, the rules for each agent are partitioned into rules for achieving goals and rules for propositions and beliefs needed to support goals. The CaSAPI tool that we propose to employ treats all rules in the same way. Furthermore, it is a coincidence that each agent i has goal rules for exactly the goals in its corresponding GB_i as in general agents may hold information about how to achieve goals they do not hold themselves.

Table 1. Example scenario as ABA framework

IKB_A	IKB_B	IKB_C
$g_1 \leftarrow a_1, x_1$ $g_2 \leftarrow p_1, x_2$ $g_3 \leftarrow p_2, a_2, x_3$ $g_3 \leftarrow a_3, x_4$	$g_2 \leftarrow q_1, x_5$ $g_4 \leftarrow b_1, q_2, x_6$ $g_5 \leftarrow q_3, b_2, x_7$	$g_1 \leftarrow r_1, c_1, x_8$ $g_6 \leftarrow r_2, x_9$
p_1 $p_2 \leftarrow a_1$	$q_1 \leftarrow b_3$ q_2 q_3 $\neg b_2$	r_3 $\neg a_3 \leftarrow r_3$ $\neg b_1 \leftarrow r_4$ $r_4 \leftarrow r_5, c_2$ $r_5 \leftarrow c_3$

For example, goal g_1 can be achieved in two ways — it can be achieved assuming *sufficiently many spare parts are in stock* (assumption a_1) or it can be achieved if *a reliable third part producer exists* (proposition r_1) and it can be assumed that *outsourcing is acceptable to all collaborators* (assumption c_1). Another example is that goal g_4 can be achieved when b_1 can be assumed (*everyone is happy*) and proposition q_2 holds (*the collaboration is profitable as a whole*). Finally[4], goal g_5 is achieved if *all collaborators are trustworthy* (proposition q_3) and assuming

[4] We refrain from detailing how to achieve the remaining goals and leave it to the reader's imagination to fill the other elements of $\mathcal{L}$ (e.g. p_2 and a_3) with meaning.

there is *at least constant demand for motorcycles* (assumption b_2). Betty's individual knowledge base IKB_B would thus contain the rule $g_4 \leftarrow b_1, q_2$ as well as two other goal rules. In what follows, we treat all p_i, q_i and r_i as propositions (i.e. elements of a propositional language $\mathcal{L}$ that are not assumptions) and all a_i, b_i and c_i as assumptions for Adrian, Betty and Carles, respectively. Assumptions can represent beliefs one cannot prove or disprove or actions to be performed by an agent. The choice is left to the designer of the agents. In addition to these a_i, b_i and c_i's, there are so-called applicability assumptions that we denote with x_i which are attached to each goal rule. Hence a rule $g_6 \leftarrow r_2, x_9$ can be read as "goal g_6 (about *Carles as leader*) can (usually) be achieved if proposition r_2 (*Carles is the most senior agent*) holds" and this rule is applicable if x_9 is the case. An agent can dispute the applicability of this rule (and thereby argue that g_6 can not be achieved in this way) by showing the contrary of x_9 (e.g. that seniority is irrelevant when determining the leader of a collaboration).

In addition to the individual knowledge bases, there is a shared knowledge base SKB containing the fact that *not sufficiently many spare parts are in stock* ($\neg a_1$) and the rule $r_1 \leftarrow r_3$, stating that *a reliable third party producer exists* if *MotoTaiwanInc has been reliable in the past.*

When defining an ABA framework one must also specify the contraries of all assumptions. For simplicity, we take the notion of contrary to be logical negation. Therefore, we have $\overline{x} = \neg x$ for any assumption x. We can now define an ABA framework for Betty where the inference rules are $\mathcal{R}_B = IKB_B \cup SKB$, the set of assumptions is $\mathcal{A}_B = \{b_1, b_2, b_3, x_5, x_6, x_7\}$ and the contrary of each assumption is its logical negation. The language $\mathcal{L}_B$ is made up of all literals that feature in the rules in IKB_B.

Using this ABA framework, Betty on her own can see that goal g_2 is achievable, if b_3 can be assumed (since b_3 *'the deal with the external distributor is fair'* allows to deduce q_1 *'a working distribution channel exists'* and that proposition is needed for goal g_2). Furthermore, she can achieve g_4 if b_1 can be assumed. Since $\neg b_2$ is a fact that Betty knows about, she can already see on her own that g_5 is not an acceptable goal for the collaboration and will thus never put it forward. We can similarly construct ABAs $\langle \mathcal{L}_A, \mathcal{R}_A, \mathcal{A}_A, \overline{\ } \rangle$ and $\langle \mathcal{L}_C, \mathcal{R}_C, \mathcal{A}_C, \overline{\ } \rangle$ for Adrian and Carles, respectively. Note, for each $i \in \{A, B, C\}$, $\mathcal{L}_i \subseteq \mathcal{L}$.

We now look at two approaches to determine the set of common goals between the three agents that can then be used to construct an institution from them.

4 Centralised approach

Any agent-based institution requires a set of institutional goals which are used to create the structure and elements of the institution. When human designers dictate these goals, the institution can be constructed from them. However, when several autonomous agents come together to form an institution, they must agree on a set of institutional goals that are acceptable to all of them. Note that within ABA a goal is "acceptable" with respect to a given semantics if there exists an acceptable set of assumptions according to that semantics supporting the goal.

Considering the semantics described in Section 2, we propose to use the sceptical grounded semantics for our scenario. As argued in [14], some domains such as legal reasoning (where a guilty verdict must be obtained via sceptical reasoning) and multi-agent systems (where decisions must be made unanimously) require sceptical semantics. In this paper we propose that agents argue about goals for their future collaboration and such goals must be agreed upon scepically — the set of acceptable goals must be unique. As justification consider the case where two goals depend upon two conflicting assumptions. Credulous semantics would allow to find support for either goal individually while sceptical reasoning excludes both goals[5]. Using the former, the acceptance of goals would hence be dependent on the order in which they are considered — which is a complication we will leave for future work.

Assume n agents want to collaborate. An intuitive approach to find a set of institutional goals consists of combining all the individual knowledge bases IKB_i of the n agents with the shared knowledge base SKB and reason with this combined knowledge. We would have the following assumption-based argumentation framework:

$\mathcal{R} = SKB' = SKB \cup \bigcup_{k=0}^{n} IKB_k$
$\mathcal{A} = \bigcup_{k=0}^{n} \mathcal{A}_k$ where $\mathcal{A}_k$ are the assumptions of agent k including its applicability assumptions

The contrary of an element x of $\mathcal{A}$ can be constructed by using the fact that the individual sets of assumptions are disjoint, as follows: find the agent i that has x as an assumption and use the contrary function of agent i. If the requirement that two sets $\mathcal{A}_k$ are pair-wise disjoint is dropped, the combined contrary function $^{-\prime}$ is constructed as follows:

$\overline{x}' = \{y \mid y = \overline{x}^i \text{ and } i \text{ is an agent } \}$.

This combined contrary function returns sets of elements. For example, if Betty and Carles have the assumption α that the sky is blue and Betty thinks that $\overline{\alpha} = \{sky\_is\_grey\}$ while Carles thinks that $\overline{\alpha} = \{sky\_is\_red\}$, then the combined system would return as a set of contraries for the assumption α the set $\{sky\_is\_grey, sky\_is\_red\}$. Showing that any one of these contraries holds, is sufficient to disprove α. Whether or not the sets of assumptions are disjoint depends on the kind of institution required.

Having constructed such an ABA framework, we can now run the argumentation system CaSAPI and query it about one goal at a time. Those goals that are acceptable according to the chosen semantics (see Section 2) are selected as institutional goals. In the example scenario with six individual goals, only goals g_1 and g_2 will become institutional goals. Adrian's rule $g_1 \leftarrow a_1$ is not helpful, since $\neg a_1$ is part of the shared knowledge base, but Carles has a way of showing g_1 provided c_1 can be assumed. CaSAPI attempts (and fails) to use the rule

[5] CaSAPI also supports the *ideal* semantics, which is less sceptical than the grounded one. It also has a unique extension but the set of goals it accepts is a superset of those accepted using grounded semantics. In future work, we will evaluate the differences.

$g_1 \leftarrow a_1$ before backtracking and succeeding by using Carles' way of showing g_1. Goal g_2 is acceptable to Betty (as discussed in Section 3.1) and since nobody can attack her argument (that from assumption b_3 the goal is possible), it becomes an institutional goal. Note, however, that Betty's goal g_4, which was acceptable to her before, given that b_1 can be assumed, is not an institutional goal. The reason is that Carles has a way of showing $\neg b_1$ (*not everyone is happy*) provided he can assume c_2 and c_3. Since between the three agents nobody can attack either c_2 or c_3, Carles' attack succeeds and Betty's goal g_4 is not acceptable.

4.1 Control issues

Rather than using CaSAPI to query one potential goal at a time, we can also use a meta-interpreter which will attempt to show that all goals are acceptable at once. If this fails, the meta-interpreter will remove one goal at a time from the set of all goals (and possibly backtrack) in order to find the biggest acceptable subset of goals. If a reason is put forward to use a different semantics (e.g. admissibility), then the meta-interpreter will need to make use of additional machinery to control the order of querying. Finally, agents could express a preference of the goals they would like to see adopted as goals of the collaboration. Again, the meta-interpreter will have to handle these. We leave the construction of such a meta-interpreter for future work.

4.2 Disadvantages

Applying the CaSAPI tool in a centralised manner, while being a computationally straight-forward approach, comes with several disadvantages. Combining the individual knowledge bases can lead to performance issues, since all the computational burden needs to be carried by one central entity which computes the optimal set of institutional goals. This agent (and the goal finding process) will quickly become the bottleneck of the system. It also increases the system's vulnerability to attacks, since without this entity's capability to execute the centralised algorithm, the agents cannot continue their efforts to form a society. Furthermore, this central agent needs to be trusted, which brings with it even more challenges.

However, the biggest disadvantage of the centralised approach concerns *privacy*. The individual agents will have to share their individual rules, knowledge and other internal details that they may want to keep secret from each other. For example, if Adrian, Betty and Carles want to form a market-place institution then while they need some common goals to make their venture happen, they would be forgiven for being hesitant to share all their business knowledge (such as detailed business rules) with one another.

5 Distributed approach

In order to allow the participating agents some privacy and to keep their individual rules private, we propose a distributed approach which does away with a

central entity which amalgamates information. Instead, each of the agents needs to be equipped with the CaSAPI engine in order to compute arguments and their supporting assumptions. Communication between the agents will be used to distribute arguments and attacks against these arguments are again computed locally. The only prerequisite we insist on is a shared understanding of the notion of contrary.

Each single goal will be treated separately, in a new conversation. Each conversation starts with one agent, the initiator whose individual goal is concerned, broadcasting a message with the goal in question and the supporting assumptions needed to defend this goal.

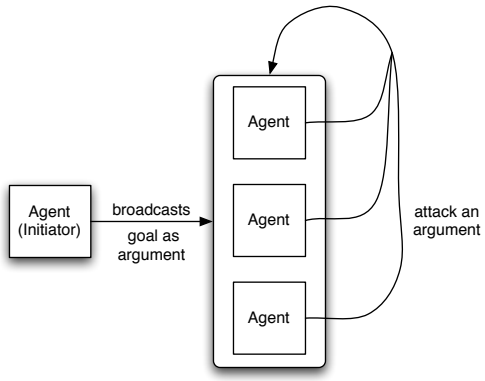

Fig. 2. Schematic description of interaction protocol

Every agent receiving this message then attempts to find an attack by looking for some support for the contrary of one of the assumptions in the initial message. This includes disputing one of the rules used to build the argument by showing the contrary of the applicability assumption. An agent who finds such an attack, broadcasts it (together with the assumptions needed to defend it). Everyone is then trying to counter-attack this attack (again by searching for supporting assumptions for an argument in support of a contrary of an assumption of the attack). This collaborative process implicitly constructs a tree of arguments and continues until no more attacks can be found and the initial argument either prevails or is defeated. If it prevails, it becomes an institutional goal.

This process can be clarified with an example from the running scenario. Imagine Betty initiates a new conversation by broadcasting her goal g_4 with the supporting set of assumptions $\{b_1\}$. Now Adrian and Carles both attempt to find an argument in favour of $\neg b_1$ since this is the only possibility to attack Betty's argument. Adrian is unsuccessful, but Carles finds an attack, namely an argument in favour of $\neg b_1$ supported by $\{c_2, c_3\}$. Since neither Adrian nor Betty can find arguments for either of $\neg c_2$ or $\neg c_3$, Carles' attack prevails and Betty's initial argument is defeated. Note that Adrian may withhold an attack

consciously in order to help Carles. We leave the issue of collusion for future work.

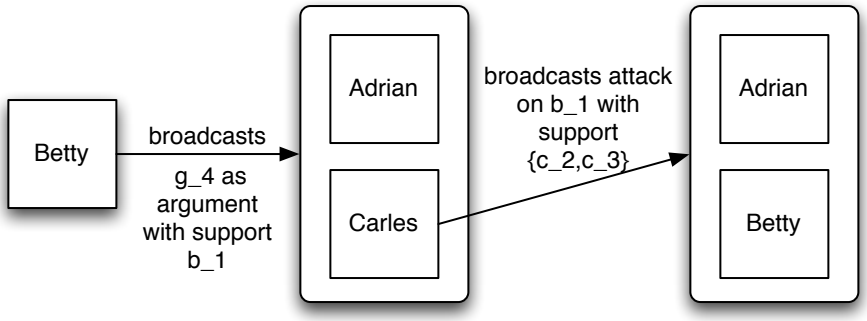

Fig. 3. An instance of the interaction protocol from the scenario

Each individual agent only initiates conversations about goals that it considers possible. Therefore, Betty will not propose goal g_5 to the other agents, as she herself is able to show its impossibility. Hence agents will never attack their own proposals. Further implementation issues are discussed in Section 5.2 of this paper but first we summarise the advantages of the distributed approach.

5.1 Advantages

The advantages of this distributed approach are three-fold:

Less vulnerability of the system as a whole, since even if an agent fails to perform (e.g. is shut down), the other agents can still continue to look for an agreement.

Privacy is maintained to the extent that rules in the individual knowledge bases are not shared between agents. In the example above where Carles successfully attacked Betty's argument, he did not have to reveal his rule $r_4 \leftarrow r_5, c_2$, for example. This privacy is useful but requires that the agents are honest. If this is not the case, an agent could counter-attack any attack on his proposed goals with a fictional support set.

Efficiency is gained in two ways. ABA provides computational savings via several filtering mechanisms (all of which employed in CaSAPI). Additionally, each agent can locally store the successful and unsuccessful arguments from the dialectical structure that is computed each time the argumentation system is run. These stored arguments can then be re-used in future conversations to save recomputing their support sets. Some of these savings will however be offset by increased communication cost.

5.2 Implementation Issues

We want to briefly touch upon two issues that require further discussion. The first one is the *order of goal proposals*. If a no meta-interpreter (cf. Section 4.1) is used, then in the simplest case, a token-based approach can be employed where the n agents that want to reach an agreement form a circle and only the agent in possession of the token can initiate a new conversation. After the initiation it passes the token on to the next agent in clock-wise direction. An agent can also pass the token on without initiating a new conversation, if all his goals have been discussed or are in discussion. If the token moves n times without a new conversation being started, then the process finishes.

This simple approach can be optimised in various ways that we do not want to go into too deeply here. Suffice to say the order in which the goals are considered, while not changing the final result[6], has an impact on efficiency, since due to the storing of (sub-)arguments described in Section 5.1 conversations can be shortened significantly.

A second issue concerns the termination of a conversation. Above we said that the arguing stops when no more attacks can be found. We consider two ways in which this is implemented. If all agents operate on the same (or sufficiently similar) clock, a time-out mechanism can be used. If no attack has been broadcast within a specified number of seconds of the initiation of the conversation (or of the broadcasting of the most recent attack), then the argument (or the attack) prevails. A more elaborate approach has the agents explicitly stating that they cannot find an attack on a given set of supporting arguments. It involves a conversation protocol including messages to that effect. Further work is needed to formalise these protocols.

6 Related Work & Conclusions

In this paper we present original but preliminary work on the problem of finding a set of institutional goals for multi-agent systems from which institutions can be constructed. Research on agent organisations and institutions has mainly focused on specification languages (e.g. [2]), architectures and software tools and frameworks (e.g.[19, 20, 2]), agent reasoning (e.g. [29]), and understanding the evolution of norms [4]. Somewhat related approaches are found in [1] and [26], but to the best of our knowledge, no efforts on automating institutional design, from the conception of goals to the enactment of the rules, have been carried out. In this paper we take the first steps along this direction.

We are proposing to use assumption-based argumentation [5, 11, 17] and have lined-out two approaches to the problem of determining a set of institutional goals. One may argue that institutional goals should be more general and abstract than the goals of individual agents. For example, from the personal goal

[6] When a credulous semantics is used, such as admissibility, correctness does become an issue. For the GB-dispute derivations we use in this paper, the order in which the goals are considered has no impact on correctness.

"I want to finish negotiating by 8pm", one can deduce the institutional goal *"We should all finish by 8pm"*. We plan to investigate this in the future. For this paper we assume the individual goals are sufficiently general (i.e. as the institutional goal above).

A second line of investigation involves introducing trust and reputation into the argumentation and interaction mechanism. One could place more importance on the arguments of certain agents and then use a preference-based approach to resolve conflicts between goals as well as arguments. Another interesting notion is favouritism (i.e. not attacking an argument even though one could).

Finally, one can extend the same process to reasoning about joint norms. Some norms may be derived from institutional goals, others could be agreed upon using a similar approach to the one sketched in this paper.

A somewhat related field is that of team formation. An agent team consists of a number of cooperative agents which have agreed to work together toward a common goal [21]. The callenges associated with team formation involve determining how agents will be allocated to address the high-level problem, maintaining consistency among those agents during execution, and revising the team as the environment or agent population changes. In our case we focus on the negotiation that occurs prior to the formation process, namely on the agreement of high-level goals.

The work on joint intentions [7] can also be seen as relevant although it has primarily focused on understanding the motivations for a team of agents to jointly pursue/drop goals. Thus, the main focus has been on understanding cooperation as a collective (intentional) decision-making process that makes agents adopt joint actions. The working assumption is that collective intentional behaviour cannot be analysed in terms of individual intentions. In our case, we are not concerned on agents' collective mental state or decision-making, but on the argumentation machinery employed to eventually reach a collective agreement.

References

1. L. Amgoud and S. Kaci. An argumentation framework for merging conflicting knowledge bases. *International Journal of Approximate Reasoning*, 45:321–340, 2007.
2. J. Arcos, M. Esteva, P. Noriega, J. A. Rodriguez, and C. Sierra. Engineering open environments with electronic institutions. *Engineering Applications of Artificial Intelligence*, 18(2):191–204, 2005.
3. J. L. Arcos, M. Esteva, P. Noriega, J. A. Rodríguez-Aguilar, and C. Sierra. Engineering open environments with electronic institutions. *Engineering Applications of Artificial Intelligence*, 18(2):191–204, March 2005.
4. R. Axelrod. An evolutionary approach to norms. *The American Political Science Review*, 80(4), 1986.
5. A. Bondarenko, P. M. Dung, R. A. Kowalski, and F. Toni. An abstract, argumentation-theoretic framework for default reasoning. *Artificial Intelligence*, 93(1-2):63–101, 1997.
6. C. Castelfranchi. Social power: A point missed in multi-agent dai and hci. *Decentralized A.I.*, pages 49—62, 1990.

7. P. Cohen, H. Levesque, and I. Smith. On team formation, 1998.
8. R. K. Dash, N. R. Jennings, and D. C. Parkes. Computational-mechanism design: A call to arms. *IEEE Intelligent Systems*, 18(6):40–47, 2003.
9. V. Dignum. *A model for organizational interaction: based on agents, founded in logic*. PhD thesis, Utrecht University, 2003.
10. P. Dung, P. Mancarella, and F. Toni. Computing ideal sceptical argumentation. *Artificial Intelligence*, 2006.
11. P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Art. Int.*, 77, 1995.
12. P. M. Dung, R. A. Kowalski, and F. Toni. Dialectic proof procedures for assumption-based, admissible argumentation. *Artif. Intell.*, 170:114–159, 2006.
13. P. M. Dung, P. Mancarella, and F. Toni. A dialectic procedure for sceptical, assumption-based argumentation. In *COMMA*, 2006.
14. P. M. Dung, P. Mancarella, and F. Toni. Computing ideal sceptical argumentation. *Art. Int., Special Issue on Argumentation in Art. Int.*, 171(10–15):642–674, 2007.
15. J. Ferber, O. Gutknecht, and F. Michel. From agents to organizations: An organizational view of multi-agent systems. In *AOSE*, pages 214–230, 2003.
16. D. Gaertner, A. Garcia-Camino, P. Noriega, J. A. Rodriguez-Aguilar, and W. Vasconcelos. Distributed norm management in regulated multi-agent systems. In *Proc. of the Sixth Int'l AAMAS Conference*, 2007.
17. D. Gaertner and F. Toni. CaSAPI: a system for credulous and sceptical argumentation. In *Proceedings of the International Workshop on Argumentation and Non-Monotonic Reasoning (ArgNMR 2007)*, 2007.
18. D. Gaertner and F. Toni. Computing arguments and attacks in assumption-based argumentation. *IEEE Intelligent Systems*, November/December, 2007.
19. B. Gâteau, O. Boissier, D. Khadraoui, and E. Dubois. MoiseInst: an organizational model for specifying rights and duties of autonomous agents. In *EUMAS*, pages 484–485, 2005.
20. O. Gutknecht, J. Ferber, and F. Michel. Integrating tools and infrastructures for generic multi-agent systems. In *Agents*, pages 441–448, 2001.
21. B. Horling and V. Lesser. A survey of multi-agent organizational paradigms. *The Knowledge Engineering Review*, 19(4):281–316., 2005.
22. Y. Kalfoglou and M. Schorlemmer. Ontology mapping: The state of the art. *Knowledge Engineering Review*, 2003.
23. J. Patel, W. T. L. Teacy, N. R. Jennings, M. Luck, S. Chalmers, N. Oren, T. J. Norman, A. D. Preece, P. M. D. Gray, G. Shercliff, P. J. Stockreisser, J. Shao, W. A. Gray, N. J. Fiddian, and S. G. Thompson. CONOISE-G: agent-based virtual organisations. In *Proc. of the Int'l AAMAS Conference*, pages 1459–1460, 2006.
24. H. Prakken and G. Sartor. On the relation between legal language and legal argument: assumptions, applicability and dynamic priorities. In *Proc. of the 5th Int'l Conference on Artificial Intelligence and Law*, pages 1–10. ACM Press, 1995.
25. O. Shehory and S. Kraus. Methods for task allocation via agent coalition formation. *Artif. Intell.*, 101(1-2):165–200, 1998.
26. K. Sycara. Argumentation: Planning other agents' plans. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 1989.
27. M. Tambe and W. Zhang. Towards flexible teamwork in persistent teams: Extended report. *Autonomous Agents and Multi-Agent Systems*, 3(2):159–183, 2000.
28. D. Tapscott and A. D. Williams. *Wikinomics: How Mass Collaboration Changes Everything*. Portfolio Hardcover, 2006.
29. F. L. y López, M. Luck, and M. d'Inverno. Constraining autonomy through norms. In *Proceedings of the International AAMAS Conference*, pages 674–681, 2002.