Groups as Agents with Mental Attitudes

Guido Boella
Dipartimento di Informatica
Università di Torino - Italy
e-mail: guido@di.unito.it

Leendert van der Torre
CWI-Amsterdam and
Delft University of Technology
e-mail: torre@cwi.nl

Abstract

We discuss a model of cooperation among autonomous agents, based on the attribution of mental attitudes to groups: these attitudes represent the shared beliefs and objectives and the wish to reduce the costs for the members. When agents take a decision they have to recursively model what their partners are expected to do under the assumption that they are cooperative, and they have to adopt the goals and desires attributed to the group: otherwise, the other members consider them uncooperative and thus liable.

1. Introduction

In multiagent systems, autonomous agents interact with each other, they play roles in organizations [15], they are hold responsible for some tasks and they are subject to obligations and permissions [20]. According, e.g., to [13, 15, 28], a multiagent system should make minimal commitments on the structure of its heterogeneous members, e.g., that their autonomous behavior is driven by the representation of mental attitudes like beliefs, desires, goals or intentions. At the same time, in a multiagent system, agents form coalitions and groups to achieve goals which they are not able to pursue individually. Groups in a multiagent system interact as a whole with other agents and groups, they play roles in organizations, they are hold responsible for some tasks and they are subject to obligations and permissions [23]. But if groups act as agents in the multiagent system, groups should be described in the same terms as agents: they should be attributed mental attitudes, like beliefs, desires, goals or intentions, and an autonomous behavior.

In this paper, starting from Boella *et al.* [1]'s model of cooperation, we address the following research questions:

- How can groups be considered as agents and how can they be attributed beliefs, desires and goals?
- Which properties of cooperation can be shown in such a definition of a group?

Boella *et al.* [1] argue that the basic elements of a general model of cooperation among the members of a group of autonomous agents are: a) considering the overall advantage that the group gains from the decisions of the single agents by means of a shared utility function; b) adopting some goals of the partners, if their achievement increases the advantage for the group. These elements must be combined with c) the ability of social agents to model in a recursive way ([17]) the decisions of the other partners, so to predict their behavior. Boella *et al.* [1] show that, if these elements are present, the group's behavior satisfies the basic properties of cooperation required by Cohen and Levesque [12], Grosz and Kraus [19], Tambe [29] and Yen *et al.* [32], like helpful behavior, communication, conflict avoidance, *et cetera*. Boella *et al.* [1]'s approach, however, suffers from the lack of a precise model of beliefs, desires and goals and from the dichotomy between these qualitative notions and the quantitative approach of the decision theoretic planner they use.

In this paper, we propose a model of cooperation which provides a precise formalization of the notions of belief, desire and goal of the agents, using a logical framework; second, instead of using classical decision theory, we base the deliberation process of agents on a qualitative decision theory inspired to the BOID architecture of Broersen *et al.* [9]. The methodology we adopt is the same as the logical multiagent framework we used for normative multiagent systems [3, 6] and virtual communities [4]. In those papers, we use a similar metaphor: a normative system can be described as an agent. Here, we propose that groups are modelled as agents, too.

We assume in this paper that a group is already formed and, hence, we do not consider the problem of group creation (see, e.g., [24, 27]) or its dynamics. Moreover, we do not consider the problem of distributed planning in a group.

The structure of this paper is the following: in Section 2, we describe the attribution of mental attitudes to a group. Then, in Sections 3 and 4, we present the formal framework. In Section 5, we apply the framework to some scenarios typical of cooperation. A summary closes the paper.

2. The group as an agent

Bratman [7] considers the following key features of shared cooperative activity:

a. *Commitment to the joint activity:* "The participants each have an appropriate commitment (though perhaps for different reasons) to the joint activity", [7, p. 94].

b. *Commitment to the mutual support:* "Each agent is committed to supporting the efforts of the other to play her role in the joint activity", [7, p. 94].

c. *Mutual responsiveness:* "Each participating agent attempts to be responsive to the intentions and actions of the other knowing that the other is attempting to be similarly responsible", [7, p. 94]. Where "responsiveness" means "keeping an eye to the behavior of the other and to act on the expectations that an agent has on the partner's behavior".

The first two requirements are not sufficient to define cooperative behavior. In contrast, the third one, mutual responsiveness, is a general ability of autonomous agents that enables them to work in a social environment, so it is necessary beyond cooperation. This idea is due to the sociologist Goffman [18], who argues that human actions are always taken in a situation of "strategic interaction":

> "When an agent considers which course of action to follow, before he takes a decision, he depicts in his mind the consequences of his action for the other involved agents, their likely reaction, and the influence of this reaction on his own welfare" [18, p. 12].

In the field of agent theory this idea has been formalized by Gmytrasiewitcz and Durfee [17] as recursive modelling:

> "Recursive modelling method views a multiagent situation from the perspective of an agent that is individually trying to decide what physical and/or communicative actions it should take right now. [...] In order to solve its own decision-making situation, the agent needs an idea of what the other agents are likely to do. The fact that other agents could also be modelling others, including the original agent, leads to a recursive nesting of models."

With respect to game theory, recursive modelling considers the practical limitations of agents, such that they can build only a finite nesting of models about other agents' decisions.

In this paper, we explain cooperative behavior by considering the group as an entity of social reality (in the sense of the construction of social reality of Searle [26]): a group exists because it is collectively attributed by all its members mental attitudes like beliefs, desires and goals. Its beliefs represent the conventions of its members and the recipes they use to achieve their shared goals. Its goals and desires represent the shared goals of its members as well as their preferences about the means to fulfill their goals and about costs they incur into. Note that the group's motivations include not only the shared goals: rather, they include also the private desires to minimize the costs for each agent; otherwise, the partners would not agree to stay in the group.

The group, however, is a social construction, and is not an agent acting in the real world. It acts indirectly via the actions of its members. In our model, agents of a group coordinate with each other since Bratman's conditions are realized as follows:

a. When they take a decision, they consider first the goals of the group and they try to maximize their fulfillment. Hence, they are committed to the joint activity.

b. When they take a decision, they include in it some actions which contribute to the efforts of their partners. Hence, they are committed to mutual support.

c. When they take a decision, they recursively model the decisions of their partners and their effects under the assumption that the partners are cooperative, too. Hence, they are mutually responsive to each other.

In more detail, when an agent evaluates a decision, he first considers which goals and desires of the group are fulfilled by his decision and which are not (a); only after maximizing the fulfillment of these motivations he includes in his decision some actions fulfilling also his private goals. When agents base their decisions on the goals and desires of the group we will say that their agent type is cooperative. This classification of agents according to the way they give priority to desires, goals or obligations is inspired by the BOID agent architecture presented in [9]. In Boella and van der Torre [3], when an agent bases his decision on the obligations he is subject to, his agent type is called respectful. Taking into accounts the motivations of other agents, and, thus also the goals and desires of the group, is an ability called "adoption": "having a state of affairs as a goal *because* another agent has the same state as a goal", [11]. According to Castelfranchi [11], adoption is a key capability for an agent to be social: social agents must be able to consider the goals of other agents and to have attitudes towards those goals.

An agent, to understand the impact of his decisions on his partners and, thus, on the goals of the group, has to recursively model what his partners will decide and how their decisions will affect the group's motivations (c). For this reason, the logical framework described in the next sections allows an agent to take a decision under the light of his partners' expected reactions. First, by using recursive mod-

elling, the agent understands whether the group's performance can be improved by including in his decisions some actions which contribute to his partners' efforts (b). Second, the agent understands whether his decision conflicts with the predicted decisions of the other agents. Third, he understands when he needs to inform the partners when their goal has been achieved, or to proactively inform them about his decisions [32]. It must be noted that if an agent includes in his decisions some actions which contribute to his partners' efforts (b), these decisions could lead an agent to being exploited by his partners. E.g., assume the partner **b** of an agent **a**, besides doing his part x in the group, has some other private goal y: agent **a** could do something for helping his partner **b** to achieve y since this could lead to a better performance of x by his partner **b**. Instead, this is not possible since, when agent **a** recursively models his partner **b**, he assumes that his partner **b** has a cooperative agent type: so, he will not disregard his part x to achieve his private goal y.

We can motivate our view by means of the following example. A group of two agents **a** and **b** has the shared goal of finding some object lost at home. Their simple plan is that agent **a** looks in the kitchen and agent **b** in the living room. Besides the shared goal, the group's motivations include the desires of the two agents to save as much time as possible. Suddenly, agent **a** finds the object; he knows that his partner **b** is still looking under the sofa. Can agent **a** exit the group since he achieved the shared goal (which is also his own goal)? He cannot, he should not abandon the group. If there were no other shared motivations besides the shared goal, then we could not explain why agent **a** should still take care of his partner **b**. Hence, we must assume some other shared desire which agent **a** should attend to: that agent **b** does not waste his time and energy. Agent **a**'s further commitment to the group is explained by the fact that he can still take a decision which allows to fulfill this desire of his partner. If the object has been found, the action of searching it again does not reach any effect: so, no other goal or desire of the group can be satisfied by looking around. Even worse, looking again has some cost (e.g., wasting time, effort, messing up the living room) which is not justified by the shared goal anymore. What makes this desire to save time and energy different from the other private desires of agent **b** is that it is attributed to the group, and, thus, agent **a** must attend to it to be cooperative. Agent **a** must attend to this desire not only while he is doing his part, e.g., by avoiding interfering with agent **b**'s actions, but also when he cannot or should not do his part anymore. So, agent **a** decides to communicate to agent **b** that the object has been found, even if this action does not satisfy any of his private goals and, rather, it costs some effort to him. But, agent **a** does so since for the group the cost of communication is worth less than the cost of searching.

3. The conceptual model

First of all, we introduce the structural concepts and their relations. We have to describe the different aspects of the world and the relationships among them. We therefore introduce a set of propositional variables X and we extend it to consider also negative states of affairs: $L(X) = X \cup \{\neg x \mid x \in X\}$. Moreover, for $x \in X$ we write $\sim x$ for $\neg x$ and $\sim(\neg x)$ for x. The relations between the propositional variables are given by means of conditional rules written as $R(X) = 2^{L(X)} \times L(X)$: the set of pairs of a set of literals built from X and a literal built from X, written as $l_1 \wedge \ldots \wedge l_n \rightarrow l$, and, when $n = 0$, $\top \rightarrow l$. The rules represent the relations among propositional variables existing in beliefs, desires and goals of the agents.

Then there are the different sorts of agents A we consider. Besides real agents RA (either human or artificial) we consider as agents in the model also socially constructed agents like groups, normative systems and organizations SA. This does not mean that these agents exist. Rather, they exist only as they are attributed mental attitudes by other agents (either real or not). By mental attitudes we mean beliefs B, desires D and goals G.

Concerning the relations existing between these structural concepts, mental attitudes are represented by rules, even if they do not coincide with them: $MD : B \cup D \cup G \rightarrow R(X)$. When there is no risk of confusion we will abuse the notation by identifying rules and mental states. To resolve conflicts among motivations $M = D \cup G$ we introduce a priority relation by means of a function $\geq$: $A \rightarrow 2^M \times 2^M$ from agents to a transitive and reflexive relation on the powerset of the motivations containing at least the subset relation. We write $\geq_a$ for $\geq(a)$. Moreover, different mental attitudes are attributed to all the different sorts of agents by the agent description relation $AD : A \rightarrow 2^{B \cup D \cup G \cup A}$. We write $B_a = AD(a) \cap B$, $A_a = AD(a) \cap A$, for $a \in A$, *etc.*

Also agents are in the target of the AD relation for the following reason: groups, normative systems and organizations exist only as profiles attributed by other agents. So groups, normative systems and organizations exist only as they are described as agents by other agents, according to the agent description relation. The AD relation induces an exists-in-profile relation specifying that an agent $b \in SA$ exists only as some other agents attribute to it mental attitudes: $\{a \in RA \mid b \in A_a\} \neq \emptyset$.

Moreover, we do not assume that an agent can observe every propositional variable: the set of observable variables $OP : A \rightarrow 2^X$ is a function from agents to the powerset of variables, where $OP(a)$ is the set of variables which agent $a \in A$ can observe.

Finally, the two sets of agents are disjoint and are all subsets of the set of agents A: $RA \cap SA = \emptyset$ and $RA \cup SA = A$.

We introduce now concepts concerning informational aspects. First of all, the set of variables whose truth value is determined by an agent (decision variables) [21] are distinguished from those which are not P (the parameters).

Concerning the relations among these concepts, we have that parameters P are a subset of the propositional variables X. The complement of X and P represents the decision variables controlled by the different agents. Hence we associate to each agent a subset of $X \setminus P$ by extending again the agent description relation $AD : A \rightarrow 2^{B \cup D \cup G \cup A \cup (X \setminus P)}$. We can now define a multiagent system as $MAS = \langle RA, SA, X, P, B, D, G, AD, MD, \geq \rangle$.

3.1. Plans

Some more words must be devoted to the representation of plans, since in our abstract model, we do not have an explicit notion of plan, with decompositions and causal links among actions, and we abstract from problems like the temporal ordering of actions. We consider a plan as a set of subgoals which imply the achievement of the goal. Each subgoal can be either a decision variable, i.e., an action directly executable by the agent, or a parameter, whose truth can be controlled indirectly via some decision variable. We focus only on how to express the notion of subgoal in our system.

If an agent $a \in A$ has a goal $r \rightarrow x \in G_a$, where r is its relevance condition, there are two possibilities: either x is directly executable by the agent or x is not directly executable. In the second case, if x is not a decision variable in X_a, it believes that it must make true some other propositional variables or to execute some actions: e.g., $y \wedge z \rightarrow x \in B_a$. To achieve x, the agent has to adopt y and z as subgoals. How can we represent this fact in our conditional rule based formalism? Certainly, saying that $\top \rightarrow y \in G_a$ and $\top \rightarrow z \in G_a$ are two unconditional goals of the agent is not enough, because we would lose the relation between x and $y \wedge z$; if x had been achieved, y and z would not be goals of the agent anymore. A first solution could be to use the fact that x has not been achieved as a condition of the goals: $\neg x \rightarrow y \in G_a$ and $\neg x \rightarrow z \in G_a$. Is this enough? It is also possible that while $\neg x$ is still true, x is not anymore a current goal of the agent since the relevance condition r is not true anymore: x is not anymore a goal to be fulfilled. The agent does not consider the possibility that the main goal becomes irrelevant before its satisfaction. Hence, the correct representation of subgoals of $r \rightarrow x \in G_a$ is $r \wedge \neg x \rightarrow y \in G_a$ and $r \wedge \neg x \rightarrow y \in G_a$. And so on, recursively, for the subgoals of y and z, if any.

In summary, a subgoal of another goal has among its conditions the relevance condition of the main goal as well as the fact that the main goal has not been achieved yet.

4. Games between agents in a group

The advantage of the attribution of mental attitudes to groups is that standard techniques developed in qualitative decision and game theory can be applied to cooperation. Here we consider a simple form of games between two agents **a** and **b** in A which form a group $\mathcal{A} \in SA$.

First of all, to incorporate the consequences of belief rules, we introduce a simple logic of rules called out: it takes the transitive closure of a set of rules, called reusable input/output logic in [22]; $out(E, S)$ be the closure of $S \subseteq L(X)$ under the rules E. We write $out*(E, S)$ for the closure under the so-called closed world assumption. For a propositional variable x, if x is not part of $out(E, S)$, then $\neg x$ is part of $out*(E, S)$.

- $out^0(E, S) = S$
- $out^{i+1}(E, S) = out^i(E, S) \cup \{l \mid L \rightarrow l \in E, L \subseteq out^i(E, S)\}$ for $i \geq 0$
- $out(E, S) = \cup_0^\infty(E, S)$
- $out*(E, S) = out(E, S) \cup \{\neg x \mid x \notin out(E, S)\}$

For an agent $a \in A$ and a decision $\delta \in \Delta$ we write δ_a for $\delta \cap L(X_a)$. When agent **a** takes its decision $\delta_{\mathbf{a}}$ it has to minimize the unfulfilled motivational attitudes it considers relevant: its own desires $D_{\mathbf{a}}$ and goals $G_{\mathbf{a}}$, and also the desires $D_{\mathcal{A}}$ and goals $G_{\mathcal{A}}$ of the group it belongs to. But when it considers these attitudes, it must not only consider its decision $\delta_{\mathbf{a}}$ and the consequences of this decision; as required by the third requirement of cooperative agents, it must consider also the decision $\delta_{\mathbf{b}}$ of its partner **b** and its consequences $out*(B_{\mathbf{a}}, \delta)$. So agent **a** recursively considers which decision agent **b** will take depending on its different decisions $\delta_{\mathbf{a}}$. Note that here agent **a** assumes that agent **b** is not aware of agent **a**'s decision and of its consequences, but only of those propositional variables $OP(\mathbf{b})$ in X that it can observe: $out*(B_{\mathbf{b}}, \delta_{\mathbf{b}} \cup (out(B_{\mathbf{a}}, \delta_{\mathbf{a}}) \cap OP(\mathbf{b}))$.

We can now define decisions of agents. The set of decisions Δ is the set of subsets $\delta = \delta_{\mathbf{a}} \cup \delta_{\mathbf{b}} \subseteq L(X)$ such that their closures under the beliefs $out*(B_{\mathbf{a}}, \delta)$ and $out*(B_{\mathbf{b}}, \delta_{\mathbf{b}} \cup (out(B_{\mathbf{a}}, \delta_{\mathbf{a}}) \cap OP(\mathbf{b}))$ do not contain a variable and its negation. Note that there is no restriction to the possibility that decisions include decision variables of X_a which do not contribute to the goals of the agent. In particular, the decisions can contain decision variables contributing to the goals to be achieved by the partner of the agent in a group. Our second requirement of cooperative agents in Section 2 is thus satisfied.

Given a decision $\delta_{\mathbf{a}}$, a decision $\delta_{\mathbf{b}}$ is optimal for agent **b** if it minimizes the unfulfilled motivational attitudes in $D_{\mathbf{b}}$ and $G_{\mathbf{b}}$ according to the $\geq_{\mathbf{b}}$ relation. The decision of agent $\mathbf{a}$ is more complex: for each decision $\delta_{\mathbf{a}}$ it must consider which is the optimal decision $\delta_{\mathbf{b}}$ for agent **b**.

Definition 1 (Recursive modelling) *Let:*

- *the unfulfilled motivations of decision δ according to agent $\mathbf{a} \in A$ be the set of motivations whose body is part of the closure of the decision under belief rules but whose head is not.*
  $U(\delta, \mathbf{a}) = \{m \in M \mid MD(m) = l_1 \wedge \ldots \wedge l_n \rightarrow l, \{l_1, \ldots, l_n\} \subseteq out*(B_\mathbf{a}, \delta)$ *and* $l \notin out*(B_\mathbf{a}, \delta)\}$.
- *the unfulfilled motivations of decision $\delta = \delta_\mathbf{a} \cup \delta_\mathbf{b}$ according to agent $\mathbf{b}$ be the set of motivations whose body is in the observable part of the closure of the decision under belief rules, but whose head is not:*
  $U(\delta, \mathbf{b}) = \{m \in M \mid MD(m) = l_1 \wedge \ldots \wedge l_n \rightarrow l, \{l_1, \ldots, l_n\} \subseteq out*(B_\mathbf{b}, \delta_\mathbf{b} \cup (out(B_\mathbf{a}, \delta_\mathbf{a}) \cap OP(\mathbf{b})))$ *and* $l \notin out*(B_\mathbf{b}, \delta_\mathbf{b} \cup (out(B_\mathbf{a}, \delta_\mathbf{a}) \cap OP(\mathbf{b})))\}$.
- *a decision δ is* optimal *for agent $\mathbf{b}$ if and only if there is no decision $\delta'_\mathbf{b}$ such that $U(\delta, \mathbf{b}) >_\mathbf{b} U(\delta_\mathbf{a} \cup \delta'_\mathbf{b}, \mathbf{b})$. A decision δ is optimal for agent $\mathbf{a}$ and agent $\mathbf{b}$ if and only if it is optimal for agent $\mathbf{b}$ and there is no decision $\delta'_\mathbf{a}$ such that for all decisions $\delta' = \delta'_\mathbf{a} \cup \delta'_\mathbf{b}$ and $\delta_\mathbf{a} \cup \delta''_\mathbf{b}$ optimal for agent $\mathbf{b}$ we have that $U(\delta', \mathbf{a}) >_\mathbf{a} U(\delta_\mathbf{a} \cup \delta''_\mathbf{b}, \mathbf{a})$.*

4.1. Decision making in groups

The agents value decisions according to the desires and goals which have been fulfilled and which have not. The agents can be classified according to the way they solve the conflicts among the rules belonging to different components: private desires and goals and desires and goals of the group $\mathcal{A}$ that can be adopted. We define agent types as they have been introduced in the BOID architecture [9].

In Section 2, we define as a requirement of cooperative agents the fact that they give priority to the desires and goals of the group; they pursue their private goals only if they do not prevent the achievement of the group's objectives.

Definition 2 (Agent types)

Selfish agent *A selfish agent always tries to minimize its own unfulfilled desires and goals. An agent $a \in A$ has a selfish agent type iff:*

- *if* $U(\delta, a) \geq_a U(\delta', a)$ *then* $U(\delta, a) \cap (D_a \cup G_a) \geq_a U(\delta', a) \cap (D_a \cup G_a)$

Cooperative agent *A cooperative agent always tries to minimize the unfulfilled desires and goals of the group $\mathcal{A}$, before minimizing its private goals and desires. An agent $a \in A$ has a cooperative agent type iff:*

- *if* $U(\delta, a) \geq_a U(\delta', a)$ *then* $U(\delta, a) \cap (D_\mathcal{A} \cup G_\mathcal{A}) \geq_\mathcal{A} U(\delta', a) \cap (D_\mathcal{A} \cup G_\mathcal{A})$

Similar definitions can be provided for agents who give precedence to goals with respect to desires, agents who adopt as their goals the obligations they are subject to, *etc*.

5. Properties of cooperation

In this section we discuss the properties of our model of groups using some typical scenarios.

5.1. Communication

"Any theory of joint action should indicate when communication is necessary", [12, p. 4]. The prototypical communication phenomena necessary to avoid miscoordination in a group are illustrated by Cohen and Levesque [12]: e.g., as discussed in Section 2, when an agent believes that the shared goal has been achieved, it is not yet allowed to leave the group; rather, it should ensure that all the other agents know this fact as well. We can model the necessity of this communication thanks to the interplay of the attribution of mental attitudes to the group with recursive modelling.

In the next scenario, the agents **a** and **b** in A form a group $\mathcal{A} \in SA$. The shared goal of the group is to achieve x ($\top \rightarrow x \in G_\mathcal{A}$), and to achieve x the members should use the plan $y \wedge z \rightarrow x \in B_\mathcal{A} \cap B_\mathbf{a} \cap B_\mathbf{b}$; e.g., $x \in P$ means finding an object searched for, $y \in X_\mathbf{a}$ is an action of agent **a** for looking in some room and $z \in X_\mathbf{b}$ an action of **b** for looking in another one. Moreover the group agreed not to make too much effort; e.g., the group desires preventing fuel or time consumption due to executing action y ($\top \rightarrow \neg y \in D_\mathcal{A}$); analogously for actions $z \in X_\mathbf{b}$ and $c \in X_\mathbf{a}$, where c is the communication action of agent **a**; this action makes agent **b** believe that the object has been found, i.e., the shared goal (x) has been achieved ($c \rightarrow x \in B_\mathbf{b}$).[1] However, not all actions have the same costs: e.g., y and z cost more than c (see $\geq_\mathcal{A}$),

Assume that agent **a** is going to perform its action y, but that for some reason x is already true ($\top \rightarrow x \in B_\mathbf{a}$). The agent believes that agent **b** is not aware of that ($\top \rightarrow x \notin B_\mathbf{b}$) and x is not observable by it ($OP(\mathbf{b}) = X \setminus \{x\}$). Agent **a** has to figure out which is the best decision $\delta_\mathbf{a}$, among doing nothing, doing its part y of the plan or communicating to agent **b** that x is true, or doing both. However, agent **a**'s private desires $D_\mathbf{a}$ and goals $G_\mathbf{a}$ are different from those of the group: agent **a** does not care about the cost for agent **b** of doing z ($\top \rightarrow \neg z \notin D_\mathbf{a}$) and it has as a subgoal its part of the plan y: $\neg x \rightarrow y \in G_\mathbf{a}$ (where the condition $\neg x$ expresses the fact that y is a goal only as far as the main goal x has not been achieved yet).

1 A communication action in our framework is represented in a simplified way as an action whose effects influence the beliefs of another agent. In the formalization below, c has the effect x in the beliefs of agent **b**: $c \rightarrow x \in B_\mathbf{b}$, but $c \rightarrow x \notin B_\mathbf{a}$, since $c \rightarrow x \in B_\mathbf{a}$ would mean that, according to agent **a**, c achieves x in the world.

Situation 1

Group $\mathcal{A}$:
$B_{\mathcal{A}} = \{y \wedge z \rightarrow x\}$,
$G_{\mathcal{A}} = \{\top \rightarrow x\}$,
$D_{\mathcal{A}} = \{\top \rightarrow \neg y, \top \rightarrow \neg z, \top \rightarrow \neg c\}$,
$\geq_{\mathcal{A}} \supseteq \{\top \rightarrow x\} > \{\top \rightarrow \neg y, \top \rightarrow \neg z\} > \{\top \rightarrow \neg z\} > \{\top \rightarrow \neg c\}$,

Agent **a**:
$\{y, c\} \subseteq X_{\mathbf{a}}, x \in P$,
$B_{\mathbf{a}} = \{\top \rightarrow x, y \wedge z \rightarrow x\}$,
$G_{\mathbf{a}} = \{\top \rightarrow x, \neg x \rightarrow y\}$,
$D_{\mathbf{a}} = \{\top \rightarrow \neg y, \top \rightarrow \neg c\}$,
$\geq_{\mathbf{a}} \supseteq \{\top \rightarrow x\} > \{\top \rightarrow \neg y, \top \rightarrow \neg z\} > \{\top \rightarrow \neg z\} > \{\top \rightarrow \neg c\}$,

Agent **b**:
$z \in X_{\mathbf{b}}, OP(\mathbf{b}) = X \setminus \{x\}$,
$B_{\mathbf{b}} = \{y \wedge z \rightarrow x, c \rightarrow x\}$,
$G_{\mathbf{b}} = \{\top \rightarrow x, \neg x \rightarrow z\}$,
$D_{\mathbf{b}} = \{\top \rightarrow \neg z\}$,
$\geq_{\mathbf{b}} \supseteq \{\top \rightarrow x\} > \{\top \rightarrow \neg y, \top \rightarrow \neg z\} > \{\top \rightarrow \neg z\} > \{\top \rightarrow \neg c\}$,

Optimal decision: $\delta_{\mathbf{a}} = \{c, \neg y\}, \delta_{\mathbf{b}} = \{\neg z\}$
Consequences of beliefs:
$out*(B_{\mathbf{a}}, \{c, \neg y, \neg z\}) = \{x, c, \neg y, \neg z\}$
$out*(B_{\mathbf{b}}, \{\neg z\} \cup (out(B_{\mathbf{a}}, \{c, \neg y\}) \cap OP(\mathbf{b}))) = \{x, c, \neg y, \neg z\}$
Unfulfilled motivational attitudes:
$U(\delta, \mathbf{a}) \cap (D_{\mathbf{a}} \cup G_{\mathbf{a}}) = \{\top \rightarrow \neg c\}$
$U(\delta, \mathbf{b}) \cap (D_{\mathbf{b}} \cup G_{\mathbf{b}}) = \emptyset$.

Since agent **a** decides to do $\{c\}$, then agent **a**'s unconditional (and hence applicable) desire $\top \rightarrow \neg y \in D_{\mathbf{a}}$ is fulfilled in $out*(B_{\mathbf{a}}, \{c, \neg y, \neg z\})$ (the antecedent $\top$ of the unconditional rule $\top \rightarrow \neg y$ is true and also the consequent $\neg y$ is), while $\top \rightarrow \neg c$ remains unsatisfied ($\neg c \notin out*(B_{\mathbf{a}}, \{c, \neg y, \neg z\}\})$). Moreover, the shared goal $\top \rightarrow x \in G_{\mathcal{A}}$ is satisfied and $\neg x \rightarrow y \in G_{\mathbf{a}}$ is not applicable ($\neg x \notin out*(B_{\mathbf{a}}, \{c, \neg y, \neg z\}\})$).

Concerning agent **b**, it believes that the consequences of the decision $\delta = \delta_{\mathbf{a}} \cup \delta_{\mathbf{b}}$ are $\{x, c, \neg y, \neg z\}$ due to the effect of c (even if x cannot be observed, $c \rightarrow x \in B_{\mathbf{b}}$ and c can be observed, $c \in OP(\mathbf{b})$). Given these consequences, its part of the plan $\neg x \rightarrow z \in G_{\mathbf{b}}$ is not relevant and, thus, has not to be satisfied ($\neg x \notin out*(B_{\mathbf{b}}, \{\neg z\} \cup (out(B_{\mathbf{a}}, \{c, \neg y\}) \cap OP(\mathbf{b})))$).
Had agent **a**'s decision been $\delta'_{\mathbf{a}} = \{\neg c, \neg y\}$ it would fulfill **a**'s and group's desire to avoid the cost $\neg c$ ($\top \rightarrow c \in D_{\mathbf{a}} \cap D_{\mathcal{A}}$). However, it would leave agent **b** unaware of the satisfaction of the shared goal: $out*(B_{\mathbf{b}}, \{z\} \cup (out(B_{\mathbf{a}}, \{\neg c, \neg y\}\}) \cap OP(\mathbf{b}))) = \{z, \neg x, \neg c, \neg y\}$.

How does agent **a** take a decision between $\delta_{\mathbf{a}}$ and $\delta'_{\mathbf{a}}$? It compares which of its goals and desires remain unsatisfied in the light of agent **b**'s decision: $\delta'_{\mathbf{b}} = \{z\}$. Agent **a** knows that $\delta'_{\mathbf{b}}$ is the optimal decision after $\delta'_{\mathbf{a}}$ for agent **b** since $\delta'_{\mathbf{b}}$ would achieve its goal $\neg x \rightarrow z$. So the unfulfilled desires of the group would have been $U(\delta'_{\mathbf{a}} \cup \delta'_{\mathbf{b}}, \mathbf{a}) \cap (D_{\mathcal{A}} \cup G_{\mathcal{A}}) = \{\top \rightarrow \neg z\}$. Since $\geq_{\mathcal{A}} \supseteq \{\top \rightarrow \neg z\} > \{\top \rightarrow \neg c\}$ (i.e., communication is less costly than doing z) $\delta_{\mathbf{a}}$ is preferred over $\delta'_{\mathbf{a}}$ by a cooperative agent **a**: $U(\delta_{\mathbf{a}} \cup \delta'_{\mathbf{b}}, \mathbf{a}) \cap (D_{\mathcal{A}} \cup G_{\mathcal{A}}) \geq_{\mathbf{a}} U(\delta'_{\mathbf{a}} \cup \delta'_{\mathbf{b}}, \mathbf{a}) \cap (D_{\mathcal{A}} \cup G_{\mathcal{A}})$.

Had agent **a** been a selfish agent, its decision would have been $\delta'_{\mathbf{a}}$, since $U(\delta'_{\mathbf{a}} \cup \delta'_{\mathbf{b}}, \mathbf{a}) \cap (D_{\mathbf{a}} \cup G_{\mathbf{a}}) = \{\top \rightarrow \neg c\} \geq_{\mathbf{a}} U(\delta_{\mathbf{a}} \cup \delta_{\mathbf{b}}, \mathbf{a}) \cap (D_{\mathbf{a}} \cup G_{\mathbf{a}}) = \emptyset$.

5.2. Helpful behavior

When, due to recursive modelling, agent **a** believes that agent **b** is experiencing some difficulties in doing its part, it decides to do something to resolve them, but only in case its intervention ensures less costs for the group.

In the next scenario the plan $y \wedge z \rightarrow x \in B_{\mathcal{A}}$ for achieving x is composed by an action $y \in X_{\mathbf{a}}$ of agent **a** and a parameter $z \in P$ which can be made true by agent **b** my means of action $j \in X_{\mathbf{b}}$, but only under condition p ($j \wedge p \rightarrow z \in B_{\mathcal{A}}$); agent **a** can achieve p by doing $h \in X_{\mathbf{a}}$; agent **b** has the goal of doing j for achieving z: $\neg x \wedge \neg z \rightarrow j \in G_{\mathbf{b}}$. What happens if j cannot achieve z since precondition p is not true and agent **b** cannot do anything for making p true?

Situation 2

Group $\mathcal{A}$:
$B_{\mathcal{A}} = \{y \wedge z \rightarrow x, j \wedge p \rightarrow z, h \rightarrow p\}$,
$G_{\mathcal{A}} = \{\top \rightarrow x\}$,
$D_{\mathcal{A}} = \{\top \rightarrow \neg y, \top \rightarrow \neg z, \top \rightarrow \neg h\}$,
$\geq_{\mathcal{A}} \supseteq \{\top \rightarrow x\} > \{\top \rightarrow \neg y, \top \rightarrow \neg z, \top \rightarrow \neg h\}$

Agent **a**:
$y, h \in X_{\mathbf{a}}, \{x, z, p\} \subseteq P$,
$B_{\mathbf{a}} = \{y \wedge z \rightarrow x, j \wedge p \rightarrow z, h \rightarrow p\}$,
$G_{\mathbf{a}} = \{\top \rightarrow x, \neg x \rightarrow y\}$,
$D_{\mathbf{a}} = \{\top \rightarrow \neg y, \top \rightarrow \neg h\}$,
$\geq_{\mathbf{a}} \supseteq \{\top \rightarrow x\} > \{\top \rightarrow \neg y, \top \rightarrow \neg z, \top \rightarrow \neg h\}$

Agent **b**:
$j \in X_{\mathbf{b}}, OP(\mathbf{b}) = X$,
$B_{\mathbf{b}} = \{y \wedge z \rightarrow x, j \wedge p \rightarrow z, h \rightarrow p\}$,
$G_{\mathbf{b}} = \{\top \rightarrow x, \neg x \rightarrow z, \neg x \wedge \neg z \rightarrow j\}$,
$D_{\mathbf{b}} = \{\top \rightarrow \neg z\}$,
$\geq_{\mathbf{b}} \supseteq \{\top \rightarrow x\} > \{\top \rightarrow \neg y, \top \rightarrow \neg z, \top \rightarrow \neg h\}$

Optimal decision: $\delta_{\mathbf{a}} = \{y, h\}, \delta_{\mathbf{b}} = \{j\}$
Consequences of beliefs:
$out*(B_{\mathbf{a}}, \{y, h, j\}) = \{p, y, h, j, \neg x, \neg z\}$
$out*(B_{\mathbf{b}}, \{j\} \cup (out(B_{\mathbf{a}}, \{y, h\}) \cap OP(\mathbf{b}))) = \{p, x, z, j, h, y\}$
Unfulfilled motivational attitudes:
$U(\delta, \mathbf{a}) \cap (D_{\mathbf{a}} \cup G_{\mathbf{a}}) = \{\top \rightarrow \neg y, \top \rightarrow \neg h\}$
$U(\delta, \mathbf{b}) \cap (D_{\mathbf{a}} \cup G_{\mathbf{a}}) = \{\top \rightarrow \neg z\}$.

Agent **a** accepts to do also action h to achieve p ($h \rightarrow p \in B_{\mathcal{A}}$), so that agent **b**'s action j can achieve z. Thanks to recursive modelling, it can predict that if it does not do h, the group cannot achieve the shared goal. It does so since for the group it is better to face the additional cost of doing h than to give up the shared objective: $\geq_{\mathcal{A}} \subseteq \{\top \rightarrow x\} > \{\top \rightarrow \neg y, \top \rightarrow \neg z, \top \rightarrow \neg h\}$.

Sometimes, helpful behavior is not sufficient; in the previous situation if agent **b** is not aware of the contribute of agent **a** to achieve p and p and h are not observable ($OP(\mathbf{b}) = X \backslash \{h, p\}$), then agent **a** has to consider whether to communicate to agent **b** that p is true ($c \rightarrow p \in B_{\mathbf{b}}$).

5.3. Conflict avoidance

When agents can choose how to do their part, they can minimize their private costs - i.e., desires not contained in $D_{\mathcal{A}}$ - but, in doing so, they have to ensure that they do not prevent other agents from doing their part.

Consider an example where agent **a** can achieve its part of the shared plan $y \in P$ (a parameter) by doing $j \in X_{\mathbf{a}}$ or $k \in X_{\mathbf{a}}$; action k is less costly than j: $\geq_{\mathbf{a}} \supseteq \{\top \rightarrow \neg j\} > \{\top \rightarrow \neg k\}$ and $\{\top \rightarrow \neg j, \top \rightarrow \neg k\} \subseteq D_{\mathbf{a}}$ (but the two desires do not belong to $D_{\mathcal{A}}$). However, if k is true, the agent **b** cannot achieve its goal $z \in P$ (a parameter) by doing action $h \in X_{\mathbf{b}}$: $h \rightarrow z \in B_{\mathbf{b}}$ but $h \wedge k \rightarrow \neg z \in B_{\mathbf{b}}$. Hence, agent **a** decides to do the more costly action j.

5.4. Ending cooperation

When agent **a**, whatever action it chooses, cannot do anything for the group, it can consider itself as out of the group and it is entitled to return to its private goals. As a particular case we have the situation requested by Cohen and Levesque [12] that the group terminates when there is the mutual belief that every agent believes that the shared goal has been achieved. Consider a scenario similar to Situation 1: this time, both agents are aware that x has been satisfied. So no communication is necessary and cooperation ends without leaving any goal of the group unsatisfied.

Analogously, the agent can leave the group when it believes that the other agent knows that the shared goal has become irrelevant or that it is impossible to be achieved.

Agent **a** gives up the cooperation not only when the final conditions are met for all the other members, but also when there is nothing to do for preventing the other members incur in some cost for the group. For example, consider on Situation 1, assuming this time that agent **a** knows that its attempt to communicate to **b** that the shared goal x has been achieved will fail, since a precondition g does not hold and agent **a** cannot do anything for making it true: $\top \rightarrow g \notin B_{\mathbf{b}}$ and $c \wedge g \rightarrow \neg x \in B_{\mathbf{b}}$.

6. Summary and concluding remarks

In this paper, groups are considered as agents: each member of the group has to adopt the goals and desires attributed to the group when it takes a decision and to add some actions to help his partners. Moreover, they recursively model the decisions of their partners to predict the result of their actions. The group is not a real agent, but an entity belonging to the social reality and constructed by the agents when they join together. However, a group can interact with other agents, it can play roles in organizations, it can be hold responsible for some tasks and it can be subject to obligations and permissions. This model allows to explain cooperation phenomena like communication, helpful behavior, conflict avoidance, termination of cooperation.

In this paper, we consider a framework based on a qualitative decision theory: decisions are taken on the basis of the desires and goals of the agents, rather than on a quantitative representation. The idea that the group can be described by an agent who has its own desires and goals plays the same role as that of the shared utility function in Boella *et al.* [1]. Moreover, the idea that this shared utility function is part of the members' individual utility functions is substituted by the fact that a cooperative agent adopts the goals of the group. In this way, we need not distinguish anymore between the desires and goals attributed to the agents (like the goal of doing their part of the shared plan) on the one side, and the utility functions expressing their preferences on the other side like in [1]. We avoid also the problems that classical decision theory presents when dealing with plans rather than with decisions, as discussed, e.g., in Dastani *et al.* [14].

It must be noted that our approach departs from the idea due to Bratman [7] that shared cooperative activity is defined by individual mental states and their interrelationships, without collective forms of attitudes that go beyond the mind of individuals and without further mental states characterizing cooperative behavior: "a shared intention is not an attitude in the mind of some super-agent consisting literally of some fusion of the two agents", [8, p. 111]. This "broadly individualistic" approach contrasts with many other approaches like Gilbert [16] (the cooperating agents form "a plural subject which is no more reducible"), Tuomela [30] (who introduce *we-intentions* - "we shall do G" - which represent the internalization of the notion of group in its members) and Searle [25] ("collective intentional behavior is a primitive phenomenon").

Finally, the definition of cooperation we presented is a prescriptive model: it explains how the members of a group should behave if they want to be cooperative. We make no assumption about why an agent is cooperative and, thus, adopts the goals and desires of the group. But, as Castelfranchi [10] argues, when an agent enters a group, a social commitment is created: this determines the right of the other

members of the group to control that the agent does his part, to complain and protest if he abandons the group and to require compensations for the consequent losses. Hence, cooperation is strictly connected with rights and obligations between agents. In Tuomela [31]'s terminology, the groups' attitudes are binding, in the sense of "an objective obligation to accept the attitude (goal, intention, belief, action) as applicable to all group members". This normative character can be described in our model thanks to the fact in this paper we exploit a multiagent framework similar to the one proposed by [3, 4] for modelling normative systems.

Related work is [3, 4, 5, 6] which analyze in a similar qualitative game theory the problem of normative reasoning in multiagent systems. Analogously to this paper, the basic idea is the attribution of mental attitudes - beliefs, desires and goals - to social entities like normative systems. In particular, the role of beliefs attributed to normative systems for modelling constitutive norms is addressed in [6].

Further research issues are also the power and dependence relations in groups [2], the creation of the group and its dynamics, and the distribution of obligations in a group.

References

[1] G. Boella, R. Damiano, and L. Lesmo. Cooperation and group utility. In *Intelligent Agents VI (ATAL'99)*, pages 319–333, Berlin, 1999. Springer Verlag.

[2] G. Boella, L. Sauro, and L. van der Torre. Social viewpoints on multiagent systems. In *Procs. of AAMAS'04*, New York, 2004.

[3] G. Boella and L. van der Torre. Attributing mental attitudes to normative systems. In *Procs. of AAMAS'03*, pages 942–943, Melbourne, 2003. ACM Press.

[4] G. Boella and L. van der Torre. Local policies for the control of virtual communities. In *Procs. of IEEE/WIC Web Intelligence Conference*, pages 161–167. IEEE Press, 2003.

[5] G. Boella and L. van der Torre. Contracts as legal institutions in organizations of autonomous agents. In *Procs. of AAMAS'04*, New York, 2004.

[6] G. Boella and L. van der Torre. Regulative and constitutive norms in normative multiagent systems. In *Procs. of KR'04*, Whistler (CA), 2004.

[7] M. Bratman. Shared cooperative activity. *The philosophical Review*, 101:327–341, 1992.

[8] M. Bratman. *Faces of intention: selected essays on intention and agency*. Cambridge University Press, Cambridge (UK), 1999.

[9] J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre. Goal generation in the BOID architecture. *Cognitive Science Quarterly*, 2(3-4):428–447, 2002.

[10] C. Castelfranchi. Commitment: from intentions to groups and organizations. In *Proc. of ICMAS'95*, pages 41–48, Cambridge (MA), 1995. AAAI/MIT Press.

[11] C. Castelfranchi. Modeling social action for AI agents. *Artificial Intelligence*, 103:157–182, 1998.

[12] P. R. Cohen and H. J. Levesque. Confirmation and joint action. In *Procs. of IJCAI'91*, pages 951–957, Sydney, 1991.

[13] M. Dastani, V. Dignum, and F. Dignum. Role-assignment in open agent societies. In *Procs. of AAMAS'03*, pages 489–496, Melbourne, 2003. ACM Press.

[14] M. Dastani, J. Hulstijn, and L. van der Torre. How to decide what to do? *European Journal of Operational Research*, 2003.

[15] V. Dignum, J.-J. Meyer, and H. Weigand. Towards an organizational-oriented model for agent societies using contracts. In *Procs. of AAMAS'02*, pages 694–695, Bologna, 2002. ACM Press.

[16] M. Gilbert. Walking together: a paradigmatic social phenomenon. *Midwest Studies*, 15:1–14, 1990.

[17] P. J. Gmytrasiewicz and E. H. Durfee. Formalization of recursive modeling. In *Procs. of ICMAS'95*, pages 125–132, 1995.

[18] E. Goffman. *Strategic Interaction*. Basil Blackwell, Oxford, 1970.

[19] B. Grosz and S. Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

[20] A. Jones and J. Carmo. Deontic logic and contrary-to-duties. In D. Gabbay, editor, *Handbook of Philosophical Logic*, pages 203–279. Kluwer, 2001.

[21] J. Lang, L. van der Torre, and E. Weydert. Utilitarian desires. *Autonomous Agents and Multiagent Systems*, pages 329–363, 2002.

[22] D. Makinson and L. van der Torre. Input-output logics. *Journal of Philosophical Logic*, 29:383–408, 2000.

[23] O. Pacheco and J. Carmo. A role based model of normative specification of organized collective agency and agents interaction. *Autonomous Agents and Multiagent Systems*, 6:145–184, 2003.

[24] T. Sandholm, K. Larson, M. Andersson, O. Shehory, and F. Tohme. Coalition structure generation with worst case guarantee. *Artificial Intelligence*, 111 (1-2):209–238, 1999.

[25] J. Searle. Collective intentionality. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in communication*. MIT Press, 1990.

[26] J. Searle. *The Construction of Social Reality*. The Free Press, New York, 1995.

[27] I. Smith and P. Cohen. Toward a semantics for an agent communications language based on speech-acts. In *Proc. 14th Conf. AAAI*, pages 24–31, Portland, 1996.

[28] V. Subrahmanian, P. Bonatti, J. Dix, T. Eiter, S. Kraus, F. Özcan, and R. Ross. *Heterogenous Active Agents*. MIT-Press, 2000.

[29] M. Tambe. Towards flexible teamwork. *Journal of Artificial Intelligence Research*, 7(7):83–124, 1997.

[30] R. Tuomela. *The Importance of Us: A Philosophical Study of Basic Social Notions*. Stanford University Press, 1995.

[31] R. Tuomela. *Cooperation: A Philosophical Study*. Kluwer, Dordrecht, 2000.

[32] J. Yen, X. Fan, S. Sun, R. Wang, C. Chen, K. Kamali, M. S. Miller, and R. A. Volz. Formal semantics and communication strategies for proactive information delivery among team-based agents. In *Procs. of AAMAS'03*, pages 1166–1167, Melbourne, 2003. ACM Press.