Attributing Mental Attitudes to Groups
Cooperation in a Qualitative Game Theory

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

Leendert van der Torre
CWI
Amsterdam
The Netherlands
e-mail: torre@cwi.nl

Abstract

We discuss a general model of cooperation among autonomous agents, based on a qualitative game theory. The basic elements of the model are the ability of agents to recursively model what their partners will do, and the idea that a group can be described as an agent whom goals and desires are attributed to: these represent the shared objective and the wish to save the members's resources. When the agents of the group take a decision they must adopt these goals and desires: if they don't do that, they are considered by the other members uncooperative and thus liable.

1 Introduction

Cooperation is necessary in many multiagent environments, from information integration, to interactive education and to collective robotics. Since also human users interact with such systems, the use of agent technology is appealing, because agent behavior is driven by notions like beliefs, desires and goals, inspired to human agenthood.

Boella *et al.* [1] show that two basic elements of a general model of cooperation among a group of BDI agents are: 1) considering the overall advantage that the group gains from the decisions of the single agents; 2) the recursive modelling ([15]) by each agent of the decisions of the other partners. [1] show that, if these two elements are present, the group's behavior satisfies the basic properties of cooperation required by Cohen and Levesque [11], Grosz and Kraus [17] and Tambe [25], like helpful behavior, communication, conflict avoidance, *et cetera*.

[1] propose their analysis of cooperation in an multiagent framework based on an extension of the decision theoretic planning paradigm proposed by Haddawy and Hanks [18]. In their model the planning activity is necessary to coordinate the actions of the agents, while decision theory is necessary to compare the different alternatives at disposal of the group. However, their approach 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 their planner which uses utility functions.

In this paper we reconsider Boella *et al.* [1]'s model of cooperation. First, we provide 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 like the one proposed by the BOID architecture of Broersen *et al.* [8]. We address the following research questions:

- How can a qualitative game theory based on recursive modelling be used in a model of cooperation among BDI agents?
- Can a group of agents be considered as an agent so that mental attitudes can be attributed to it?
- Which properties of cooperation can be shown in such a model?

Our logical multiagent framework is inspired to the one proposed for modelling the normative reasoning of agents subject to obligations and permissions in [2], [4] and [5]; the basic idea is that the normative system can be seen as an agent which is attributed beliefs, desires and goals.

In this paper, we use a similar metaphor: a group of agents can be described as an agent, and the shared goals and desires which the members aim to can be attributed to the group as its mental state.

The structure of this paper is the following: in Section 2, starting from [1], we describe the attribution of mental attitudes to a group. Then, in Section 3, we present the formal framework. Finally, in Section 4, we apply the framework to several scenarios typical of cooperation: communication, helpful behavior, conflict avoidance and correct conclusion of cooperation.

2 The group is an agent

The definition of [1] is inspired to Bratman [6], who considers the key features of shared cooperative activity:

- *Commitment to the joint activity:* "The participants each have an appropriate commitment (though perhaps for different reasons) to the joint activity", [6], p. 94.
- *Commitment to the mutual support:* "Each agent is committed to supporting the efforts of the other to play her role in the joint activity", [6], p. 94.
- *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", [6], 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 basic tenets of the definition of [1] are, thus, the following; a set of agents $\mathcal{A} = \{a_1, \ldots a_n\}$ cooperates to a shared goal x by means of a plan composed of subgoals $y_1, \ldots, y_n$ when:

1. Each agent $a_i \in \mathcal{A}$ has the goal to do its part y_i.
2. Each agent $a_i \in \mathcal{A}$ believes that the other agents of $\mathcal{A}$ have the goal to do their part.
3. Each agent $a_i \in \mathcal{A}$ believes that it shares with the other agents a (multi-attribute) utility function based on the weighed sum of the utility functions representing the shared goal and the resource consumption of the single agents. Each agent, when it plans its own part of the shared plan, has to consider also this global utility as part of its own individual utility function.
4. Each agent must adopt also the subgoals which contribute to the partners' doing their part of the plan if this adoption increases the shared utility.
5. Each agent must remain in the group as long as the adoption of some goal which contributes to the partners doing their part of the plan increases the shared utility.

This definition only partially conforms to the above requirement of mutual responsiveness; the reason is that, independently of cooperation, the authors assume that an agent is able not only to consider the effects of its decisions, but also to consider the reaction of the other agents interacting with it: an agent *recursively models* the other agents using the profile it has about their motivations and beliefs.

This idea comes from the philosophical view of the sociologist Goffman [16], 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" [16], p. 12.

In the field of agent theory this idea has been formalized by Gmytrasiewitcz and Durfee [15] with the name of recursive modelling:

> "Recursive modelling method views a multi agent 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. It can arrive at it by representing what it knows about the other agents' decision-making situations, thus modelling them in terms of their own payoff matrices. The fact that other agents could also be modelling others, including the original agent, leads to a recursive nesting of models."

With respect to pure game theory, recursive modelling considers the practical limitations of agents in realistic settings such as in acquiring knowledge and reasoning so that an agent can only build a finite nesting of models about other agents' decisions.

The combination of the definition of cooperation together with the reasoning ability of agents to do recursive modelling allows [1] to predict a number of phenomena which characterize cooperation, from helpful behavior, to conflict avoidance, to coordination by communication.

The definition of [1] assumes that the group is already in place and that the members agreed on a partial plan and distributed the subgoals composing the shared plan. The authors do not consider the negotiation phase leading to the formation of the group (see, e.g., Smith and Cohen [24]).

In this paper we reconsider the definition above, departing from it under two respects. First, 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. 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. Moreover, we avoid the problems that classical decision theory presents when dealing with plans rather than with decisions, as discussed, e.g., in Dastani *et al.* [12].

Second, our approach departs from the idea due to [6] that shared cooperative activity is defined in terms of individual mental states and their interrelationship, without resorting to collective form of attitudes that go beyond the mind of individuals and without introducing 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", [7], p. 111. This "broadly individualistic" approach contrasts with many other approaches like Gilbert [14] (the cooperating agents form "a plural subject which is no more reducible"), Tuomela and Miller [27] (who introduce *we-intentions* -"we shall do G" - which represent the internalization of the notion of group in its members) and Searle [23] ("collective intentional behavior is a primitive phenomenon").

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 [22]) which can be described as an agent, as Tuomela and Miller [27] do: i.e., mental attitudes like beliefs, desires and goals are attributed to the group. In particular, the goals and desires of the group represent the shared goal of the members as well as the desires about the means and resources they can use to fulfill the goal.

The group, however, as a social construction, is not an agent acting in the real world. It acts indirectly via the actions of its members. How the motivations of the group influence the behavior of the members? We propose that the members, to act as a group, should take into accounts the group's goals and desires. According to Castelfranchi [10], the ability of taking into account the goals of other agents - in this case, of the group - is one of the key capabilities 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, that is, to *adopt* those goals; where adoption is "having a state of affairs as a goal *because* another agent has the same state as a goal".

To be cooperative, when taking its own decision, each member of the group should adopt and give priority to the goals and desires of the group agent, and, only subordinately, it can continue to achieve its private goals. The idea of [1] of a shared utility function is substituted, in this qualitative theory, with the idea that the group can be described by an agent who has its own desires and goals. 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, when it evaluates a decision, first considers which goals and desires of the group are fulfilled by the decision and which are not; only after maximizing the fulfillment of these motivations it includes in its decision some actions fulfilling also its private goals. Note that the group's motivations include not only the shared goal and the agent's desire to preserve its own resources: rather, they include also the desire to preserve the resources of the other agents; otherwise, the partners would not agree to stay in a group where each agent takes care of its own resources only. So when a member takes a decision it has also to consider that the decision is fair for its partners.

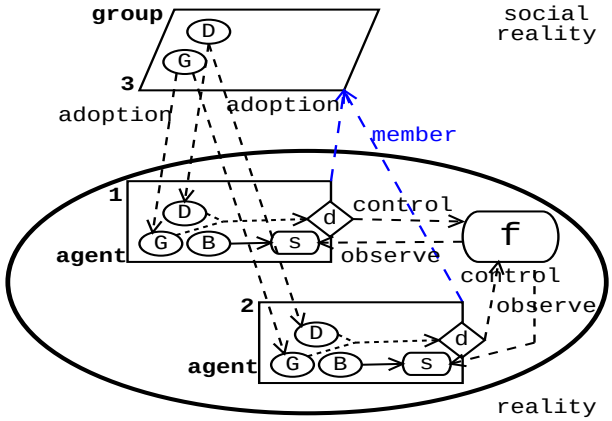

Figure 1. The adoption of group's mental attitudes.

As in [1], to understand which is the impact of its decision on the decisions of the partners and, thus, on the goals of the group, an agent has to recursively model how its partners will decide and how their decisions affect the group's motivations. For this reason, the logical framework described in the next section allows an agent to take a decision under the light of its partners' expected reactions.

When a member of a group bases its decision on the goals and desires of the group agent we will say that its 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 [8]. Analogously, in [2], when an agent bases its decision on the obligations it is subject to, its agent type is called respectful.

In Figure 1 we summarize the model. The boxes 1 and 2 represent the agents acting in the world. They have a representation of the world (s), beliefs, desires and goals (the B, D, G circles), and, basing on them, they take their decision d which affects the facts f in the world, facts which they can observe. Moreover, they are members of a group (the box 3 belonging to social reality). The group is modelled as an agent with its own desires and goals, but it cannot act in the world, since it is only a social construction. When the agents 1 and 2 take a decision they must give priority to the goals and desires adopted from the group's with respect to their own goals and desires.

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

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 [9] argue, 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 its part, to complain and protest if it abandons the group and to require compensations for the consequent losses. Hence, cooperation is strictly connected with rights and obligations between agents. In Tuomela [26]'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". As we show in Section 4.5, 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 [2], [4] and [5] for modelling normative systems.

3 Recursive modelling

In this section we present a logical framework for BDI agents based on recursive modelling: each player considers the reaction of the subsequent agents.

The basic picture is visualized in Figure 2 and reflects the deliberation of agent a_1 in various stages. Agent a_1 is going to take a decision during the cooperation to some shared goal. Agent a_2 is another agent of the group, who is going to act after agent a_1. Agent a_1 recursively models agent a_2's decision (taken from its point of view) and bases its choice on the effects of agent a_2's predicted actions. But in doing so, agent a_1 has to consider also that agent a_2 can recursively model another member a_3 of the group to coordinate its behavior with it.

When agent a_1 makes its decision d_1, it believes that it is in state s_1^0 (subscript numbers denote agents, superscript ones the time instant). The expected consequences of this decision (due to belief rules B_1) are called state s_1^1. Then agent a_2 makes a decision d_2. Now, to find out which decision agent a_2 will make, agent a_1 has a *profile* of agent a_2: it has a representation of the initial state which agent a_2 believes to be in and of the following stages. When agent a_1 makes its decision, it believes that agent a_2 believes that it is in state s_2^0. This may be the same situation as state s_1^0, but it may also be different. Then, agent a_1 believes that its own decision d_1 will have the consequence that agent a_2 believes that it is in state s_2^1, due to its observations and the expected consequences of these observations according to belief rules B_2. Agent a_1 expects that agent a_2 believes that the expected result of decision d_2 is state s_2^2. Finally, the expected consequences of d_2 from a_1's point of view are called state s_1^2. And agent a_2 makes a similar reasoning about a_3's decisions. Note however, that the recursion in modelling other agents stops here since there is no agent acting after agent a_3. Hence it does not have to base its decisions on the expected reaction of another agent.

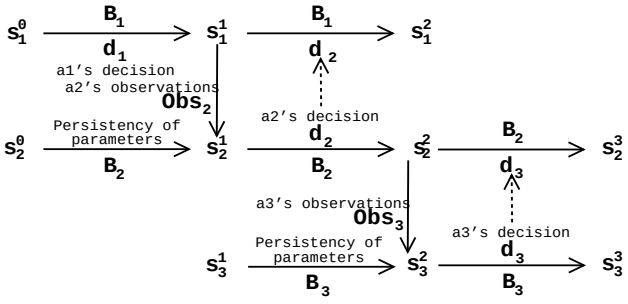

Figure 2. A three agent scenario.

3.1 Agent theory

The variables of the language are either *decision variables* of an agent, whose truth value is directly determined by it, or *parameters*, whose truth value can only be determined indirectly [19].

Definition 1 (Decisions) *Let $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ be a set of n distinct agents. $A_i = \{m, m', m'', \ldots\}$ (the decision variables) for $a_i \in \mathcal{A}$ and $P = \{p, p', p'', \ldots\}$ (the parameters) are $n+1$ disjoint sets of propositional variables. A literal is a variable or its negation. For a propositional variable p we write $\overline{p} = \neg p$ and $\overline{\neg p} = p$.*

A decision set is a tuple $\delta = \langle d_1, \ldots, d_n \rangle$ where d_i is a set of literals of A_i (the decision of agent a_i) for $1 \leq i \leq n$. Decisions are complete, in the sense that for each decision variable x in A_i, agent a_i takes a decision about it: either $x \in d_i$ or $\neg x \in d_i$.

The consequences of decisions are given by the agents' epistemic states, where we distinguish between the agents' beliefs about the world and the agents' beliefs about how a new state is constructed out of previous ones. The example in Figure 2 illustrates that we only consider games in which each agent a_i makes a decision at moment i. Second, the agents' beliefs about how a new state at moment t is constructed out of previous ones is expressed by a set of *belief rules*, denoted by B_i. Belief rules can conflict and agents can deal with such conflicts in different ways. The epistemic state therefore also contains an ordering on belief rules, denoted by $\geq_i^B$, to resolve such conflicts. Finally, to model the recursion the epistemic state of agent a_i, denoted by σ_i, includes the epistemic state of agent a_{i+1}, σ_{i+1}, unless it is the last agent a_n.

In order to distinguish the value of the propositional variables in the sequence of four stages, we use superscript numbers to label the parameters and states.

Definition 2 (Epistemic states) *Let P^0, P^1, …, P^{n+1} be the sets of propositional variables defined by $P^t = \{p^t \mid p \in P \text{ and } 0 \leq t \leq n+1\}$. We write L_{A_i}, $L_{A_i P^t}$, … for the propositional languages built up from A_i, $A_i \cup P^t$, … with the usual truth-functional connectives. We assume that the propositional language contains a symbol $\top$ for a tautology.*

Let a rule built from a set of literals be an ordered sequence of literals $l_1, \ldots, l_r, l$ written as $l_1 \wedge \ldots \wedge l_r \rightarrow l$ where $r \geq 0$. If $r = 0$, then we also write $\top \rightarrow l$.

The epistemic state *of agent a_i, $i < n$, is $\sigma_i = \langle B_i, \geq_i^B, s_i^{i-2}, s_i^{i-1}, s_i^i, s_i^{i+1}, \sigma_{i+1} \rangle$*
whereas the epistemic state of agent a_n is identical except that it does not contain the epistemic state of agent a_{n+1}. B_i is a set of rules of $L_{A_{i-1}A_iA_{i+1}P^{i-2}P^{i-1}P^iP^{i+1}}$; $\geq_i^B$ is a transitive and reflexive relation on the powerset of B_i containing at least the subset relation.

s_i^{i-2} is a set of literals of $L_{P^{i-2}}$ (the state before agent a_{i-1}'s decision). $s_i^{i-1} \subseteq L_{A_{i-1}P^{i-1}}$ (the initial state of agent a_i's decision), $s_i^i \subseteq L_{A_iP^i}$ (the state after the decision d_i of agent a_i), and $s_i^{i+1} \subseteq L_{A_{i+1}P^{i+1}}$ (the state after the decision d_{i+1} of agent a_{i+1}).
Moreover, let $s_i = s_i^{i-2} \cup s_i^{i-1} \cup s_i^i \cup s_i^{i+1}$. All states are assumed to be complete.

The agents' epistemic states depend on what it can observe. Here we accept a simple formalization of this complex phenomena, based on an explicit enumeration of all propositions which can be observed.

Definition 3 (Observations) *The propositions observable by agent a_i, OP_i, are a subset of the stage s_{i-1}^{i-1} (according to agent a_{i-1}'s point of view) including agent a_{i-1}'s decision: $P^{i-1} \cup A_{i-1}$. The expected observations of agent a_i in state s_i^{i-1} are $Obs_i = \{l \in s_{i-1}^{i-1} \mid l \in OP_i \text{ or } \overline{l} \in OP_i\}$: if a proposition describing state s_{i-1}^{i-1} is observable, then agent a_i knows its value in s_{i-1}^{i-1}. By convention $OP_1 = \emptyset$ and $s_0^0 = \emptyset$.*

The observations of agent a_i depend on the state s_{i-1}^{i-1} containing the effects of the decision of agent a_{i-1} from a_{i-1}'s point of view. What is not observed persists from the initial state s_i^{i-2} from a_i's perspective.

The consequences of rules given a set of literals are defined using the out ([20]) and $maxfamily$ functions. Intuitively, out applies iteratively the rules and $maxfamily$ selects a consistent maximal set of rules with respect to the belief rule ordering $\geq_i^B$, using intermediate phases Q and Q'.

Definition 4 (Consequences) *A set of literals is called inconsistent if it contains p and $\neg p$ for some propositional variable p; otherwise it is called consistent. For s a set of literals (state), R a set of rules, and $\geq$ a transitive and reflexive relation on the powerset of R containing at least the superset relation, let:*

1. *$out(s, R) = \cup_0^\infty out^i(s, R)$ be the state obtained by $out^0(s, R) = s$ and $out^{i+1}(s, R) = out^i(s, R) \cup \{l \mid l_1 \wedge \ldots \wedge l_n \rightarrow l \in R \text{ and } \{l_1, \ldots, l_n\} \subseteq out^i(s, R)\}$*
2. *Q is the set of subsets of R which can be applied to s without leading to inconsistency:*
   $Q = \{R' \subseteq R \mid out(s, R') \text{ consistent }\}$
3. *Q' is the set of maximal elements of Q with respect to set inclusion:*
   $Q' = \{R' \in Q \mid \nexists R'' \in Q \text{ such that } R' \subset R''\}$
4. *$maxfamily$ is the set of maximal elements of Q' with respect to the $\geq$ ordering:*
   $maxfamily(s, R, \geq) =$
   $\{R' \in Q' \mid \nexists R'' \in Q' \text{ and } R'' \geq R', R' \not\geq R''\}$

This is not, however, sufficient to define the consequences of a decision in state s at instant t. First, besides the state s also another set f representing the decision or the observation must be considered. Second, the result of the rules in $maxfamily(s, R, \geq)$ contains also parameters of s from the preceding instant t: they must be filtered out, leaving only the ones describing the consequent state at instant $t + 1$. Third, and most importantly, the parameters which are not affected by the decision must persist from the state s at the instant before t to the state at instant $t + 1$.

Definition 5 (Respect) *$next(s, f, R, \geq, t)$ be the set of states obtained by:*

1. *O is the set of new elements in $out(s \cup f, R')$:*
   $O = \{(out(s \cup f, R') \cap Lit_{A_{t+1}P^{t+1}}) \mid R' \in maxfamily(s \cup f, R, \geq)\}$
2. *$next(s, f, R, \geq, t)$ is the set of states in O plus some elements persisting from s:*
   $next(s, f, R, \geq, t) = \{G \cup s''' \mid G \in O$ *and*
   $s''' = \{l^{t+1} \mid l^t \in (P^t \cap s)$ *and* $\overline{l^{t+1}} \notin G\}\}$

An epistemic state description
$\sigma_i = \langle B_i, \geq_i^B, s_i^{i-2}, s_i^{i-1}, s_i^i, s_i^{i+1}, \sigma_{i+1}\rangle$
respects the decision set $\delta = \langle d_1, \ldots, d_n\rangle$ and the expected observations Obs_i of agent a_i if
$s_i^{i-1} \in next(s_i^{i-2}, Obs_i, B_i, \geq_i^B, i-2)$,
$s_i^i \in next(s_i^{i-2} \cup s_i^{i-1}, d_i, B_i, \geq_i^B, i-1)$,
$s_i^{i+1} \in next(s_i^{i-2} \cup s_i^{i-1} \cup s_i^i, d_{i+1}, B_i, \geq_i^B, i)$,
and, if $i < n$, σ_{i+1} respects the decision set $\delta = \langle d_1, \ldots, d_n\rangle$ and the expected observations Obs_{i+1} of a_{i+1}.

Note that the second state s_1^0 and the last one s_n^{n+1} are obtained just by persistency from s_1^{-1} and s_n^n, respectively, since for the first agent there are no observations and the last one does not recursively model the decision of any other agent and $B^0 = B^{n+1} = \emptyset$.

The following example illustrates how the persistence of parameters that are not affected by any rules is modelled.

Example 1 *Let $s_1^0 = \{p^0, q^0\}$, $d_1 = \{a\}$, $B_1 = \{a \wedge p^0 \rightarrow \neg q^1\}$. We have $out(s_1^0, B_1) = \{p^0, q^0, \neg q^1, a\}$, maximally consistent rules and preferred rules $Q' = maxfamily(s_1^0 \cup d_1, B_1, \geq^B) = \{\{a \wedge p^0 \rightarrow \neg q^1\}\}$. Proposition p^0 persists from s_1^0 since $\neg p^1$ does not belong to the next state, while q^0 does not: $O = \{\{p^0, q^0, \neg q^1, a\}\}$ and $s_1^1 = next(s_1^0, d_1, B_1, \geq^B) = \{\{p^1, \neg q^1, a\}\}$.*

Next, an example of conflicting rules:

Example 2 *Let $s_1^0 = \{p^0, q^0\}$, $d_1 = \{a\}$, $B_1 = \{a \rightarrow q^1, a \wedge p^0 \rightarrow \neg q^1\}$ and $\geq^B = \{a \wedge p^0 \rightarrow \neg q^1\} > \{a \rightarrow q^1\}$. We have $out(s_1^0, \{a \rightarrow q^1\}) = \{p^0, q^0, q^1, a\}$ and $out(s_1^0, \{a \wedge q^0 \rightarrow \neg q^1\}) = \{p^0, q^0, \neg q^1, a\}$, maximally consistent rules $Q' = \{\{a \rightarrow q^1\}, \{a \wedge p^0 \rightarrow \neg q^1\}\}$; preferred rules $maxfamily(s^0 \cup d_1, B_1, \geq^B) = \{\{a \wedge p^0 \rightarrow \neg q^1\}\}$, since $\{a \wedge p^0 \rightarrow \neg q^1\} > \{a \rightarrow q^1\}$. Proposition p^0 persists from s^0 since $\neg p^1$ does not belong to the next state, while q^0 does not: $O = \{\{p^0, q^0, \neg q^1, a\}\}$ and $s_1^1 = next(s_1^0, d_1, B_1, \geq^B) = \{\{p^1, \neg q^1, a\}\}$.*

The agent's motivational state contains two sets of rules for each agent. *Desire* (D_i) and *goal (G_i) rules* express the attitudes of the agent a_i towards a given state, depending on the context. How the agents take decisions, and in particular how they deliberate whether to cooperate or not, depends not only on their desires and goals, but also on their *agent characteristics*. Given the same set of rules, distinct agents reason and act differently. For example, a cooperative agent always tries to fulfill the goals of the group, whereas a selfish agent first tries to achieve its own goals. We express these agent characteristics by a priority relation on the rules $\geq_i$ which encode, as detailed in Broersen *et al.* [8], how the agent resolves its conflicts.

Definition 6 (Motivational states) *The* motivational state *M_i of agent a_i, $1 \leq i < n$, is a tuple $\langle D_i, G_i, \geq_i, M_{i+1}\rangle$, where D_i, G_i are sets of rules of $L_{A_{i-1}A_iA_{i+1}P^{i-2}P^{i-1}P^iP^{i+1}}$, $\geq_i$ is a transitive and reflexive relation on the powerset of $D_i \cup G_i$ containing at least the subset relation, and M_{i+1} is the motivational state that agent a_i attributes to agent a_{i+1}. The motivational state M_n of agent a_n is a tuple $\langle D_n, G_n, \geq_n\rangle$.*

A group $\mathcal{A}$ is defined by the motivational state of an agent: its desires, goals and agent characteristic.

Definition 7 (Group $\mathcal{A}$) $\langle D_{\mathcal{A}}, G_{\mathcal{A}}, \geq_{\mathcal{A}}\rangle$

3.2 Plans

In Section 3.1 we define an agent in a minimal way, as characterized by sets of conditional beliefs, desires and goals concerning propositional variables and actions. In this model, we do not have an explicit notion of plan, with decompositions and causal links among actions, and we abstract away from problems like the temporal ordering of actions. We consider a plan as a set of subgoals whose achievement implies 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_i has a goal $r \rightarrow x \in G_i$, 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 the agent is able to achieve x, it believes that it must make true some other propositional

variables or to execute some actions: e.g., $y \wedge z \rightarrow x \in B_i$. 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 $\top \rightarrow y \in G_i$ and $\top \rightarrow z \in G_i$ 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: $\neg x \rightarrow y \in G_i$ and $\neg x \rightarrow z \in G_i$. 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 proposed representation does not consider the possibility that the main goal becomes irrelevant before its satisfaction. Hence, the correct representation of subgoals of $r \rightarrow x \in G_i$ is $r \wedge \neg x \rightarrow y \in G_i$ and $r \wedge \neg x \rightarrow y \in G_i$. 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.

In this paper, we do not consider further the problem of planning, i.e., the selection of subgoals to achieve a main goal. For further planning issues, refer to [1].

3.3 Decision making in groups

The agents value, and thus induce an ordering $\leq$ on, the epistemic states by considering which desires and goals 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, 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 [8].

Definition 8 (Agent types) *Let $U(R, s)$ be the unfulfilled rules of state s,*

$$\{l_1 \wedge \ldots \wedge l_n \rightarrow l \in R \mid \{l_1, \ldots, l_n\} \subseteq s \text{ and } l \notin s\}$$

The unfulfilled motivational state description *of agent a_i belonging to group $\mathcal{A}$ is $U_i = \langle U_i^{D_i} = U(D_i, s_i), U_i^{G_i} = U(G_i, s_i), U_i^{G_\mathcal{A}} = U(G_\mathcal{A}, s_i), U_i^{D_\mathcal{A}} = U(D_\mathcal{A}, s_i)\rangle$.*

The unfulfilled motivational state description determines an ordering on the state descriptions $s_i \leq s_i'$.

Different ordering are induced by different agent types.

Selfish agent *A selfish agent always tries to minimize its own unfulfilled goals and, when there is a tie among goals, it tries to minimize its unfulfilled desires. State s_i is preferred to state s_i', $s_i \leq s_i'$, iff*

1. $U'^{G_i}_i = U(G_i, s_i') \geq_i U_i^{G_i} = U(G_i, s_i)$
2. *if* $U'^{G_i}_i =_i U_i^{G_i}$ *then* $U'^{D_i}_i \geq_i U_i^{D_i}$

Cooperative agent *A cooperative agent always tries to minimize the unfulfilled goals of the group $\mathcal{A}$ (using the agent characteristic $\geq_\mathcal{A}$ of the group) and, subordinately, the group's desires, before minimizing its private goals and desires. $s_i \leq s_i'$ iff*

1. $U'^{G_\mathcal{A}}_i = U(G_\mathcal{A}, s_i') \geq_\mathcal{A} U_i^{G_\mathcal{A}} = U(G_\mathcal{A}, s_i)$
2. *if* $U'^{G_\mathcal{A}}_i =_\mathcal{A} U_i^{G_\mathcal{A}}$ *and then* $U'^{D_\mathcal{A}}_i \geq_\mathcal{A} U_i^{D_\mathcal{A}}$
3. *if* $U'^{G_\mathcal{A}}_i =_i U_i^{G_\mathcal{A}}$ *and* $U'^{D_\mathcal{A}}_i =_i U_i^{D_\mathcal{A}}$ *then* $U'^{G_i}_i \geq_i U_i^{G_i}$
4. *if* $U'^{G_\mathcal{A}}_i =_i U_i^{G_\mathcal{A}}$ *and* $U'^{D_\mathcal{A}}_i =_i U_i^{D_\mathcal{A}}$ *and* $U'^{G_i}_i =_i U_i^{G_i}$ *then* $U'^{D_i}_i \geq_i U_i^{D_i}$

Mixed agent *A mixed agent type considers also the goals and desires of the group but does not give them priority. $s_i \leq s_i'$ iff*

1. $U'^{G_iG_\mathcal{A}}_i = U(G_i, s_i') \cup U(G_\mathcal{A}, s_i') \geq_i U_i^{G_iG_\mathcal{A}} = U(G_i, s_i) \cup U(G_\mathcal{A}, s_i)$
2. *if* $U'^{G_iG_\mathcal{A}}_i = U_i^{G_iG_\mathcal{A}}$ *then* $U'^{D_iD_\mathcal{A}}_i \geq_i U_i^{D_iD_\mathcal{A}}$

Example 3 *Given the motivational state of the group $\mathcal{A}$ $\langle D_\mathcal{A} = \{\top \rightarrow y\}, G_\mathcal{A} = \{\top \rightarrow x\}, \geq_\mathcal{A}\rangle$ and the motivational state of agent a_1 $\langle D_1 = \{\top \rightarrow z\}, G_1 = \{\top \rightarrow x, \top \rightarrow w\}, \geq_1\rangle$, the unfulfilled motivational state description of agent a_1 in state $s = \{x, y\}$ is*
$\langle U_1^{D_1} = \{\top \rightarrow z\},\ U_1^{G_1} = \{\top \rightarrow w\},\ U_1^{D_\mathcal{A}} = \emptyset,\ U_1^{G_\mathcal{A}} = \emptyset\rangle$
While in state $s' = \{x, z\}$ is
$\langle U_1^{D_1} = \emptyset,\ U_1^{G_1} = \{\top \rightarrow w\},\ U_1^{D_\mathcal{A}} = \{\top \rightarrow y\},\ U_1^{G_\mathcal{A}} = \emptyset\rangle$
A cooperative agent prefers s and a selfish one s'.

We finally define the optimal decisions. It is again a recursive definition.

Definition 9 (Optimal decisions) *A partial epistemic state is an epistemic state excluding for each agent the last three states s_i^{i-1}, s_i^i and s_i^{i+1}. A decision problem consists of a partial epistemic state, observable propositions OP_i for all agents a_i, and a motivational state M_1. A decision set is optimal for a decision problem if it is optimal for each agent a_i. A decision set is* optimal *for agent a_i if there is no decision set that dominates it for agent a_i. A decision set $\delta_i = \langle d_1, \ldots, d_n\rangle$ dominates decision set $\delta_i' = \langle d_1', \ldots, d_n'\rangle$ for agent a_i iff $d_j = d_j'$ for $1 \leq j < i$, they are both optimal for agent a_j for $i < j \leq n$, and we have $s_i < s_i'$*

- *for all s_i in an epistemic state description that contains the partial epistemic state and that respects the decision set δ_i and Obs_i, and*
- *for all s_i' in an epistemic state description that contains the partial epistemic state and that respects the decision set δ_i' and Obs_i (defined on this epistemic state).*

4 Properties of cooperation

4.1 Communication

"Any theory of joint action should indicate when communication is necessary", [11], p. 4. The prototypical communication phenomena necessary to avoid miscoordination in a group are illustrated by [11]: 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 and recursive modelling.

In the next scenario, the two agents a_1 and a_2 form a group $\mathcal{A}$. The shared goal of the group is to achieve x ($\top \rightarrow x \in G_{\mathcal{A}}$), and to achieve x the members should achieve $y \wedge z$ ($y \wedge z \rightarrow x \in B_1 \cap B_2$); e.g., $x \in P$ means finding an object searched for, $y \in A_1$ is an action of a_1 for looking in some room and $z \in A_2$ an action of a_2 for looking in another one. Moreover the group desires not to make too much effort. E.g., the group desires preventing the fuel or time consumption (fy) due to executing action y ($\top \rightarrow fy \in D_{\mathcal{A}}$); where $\neg fy$ is the side effect of doing action y ($y \rightarrow \neg fy \in B_1 \cap B_2$); analogously for fz and fc. However, not all actions have the same costs: e.g., fy and fz are worth more than fc (see $\geq_{\mathcal{A}}$), where fc is the cost of the communication action c of agent a_1; this action makes agent a_2 believe that the object has been found, i.e., the shared goal (x) has been achieved ($c \rightarrow x \in B_2$).[1]

Assume that agent a_1 is going to perform its action y, but that for some reason x is already true ($x \in s_1^0$): e.g., the object has been found by someone else who gave it to a_1. The agent believes that agent a_2 is not aware of that ($\neg x \in s_2^0$) since x is not observable by it in state s_1^1 ($OP_2 = A_1 \cup P^1 \setminus \{x\}$). Agent a_1 has to figure out which is the best decision d_1, among doing nothing, doing its part y of the plan or communicating to agent a_2 that x is true or, to do both. However, agent a_1's private desires D_1 and goals G_1 are different from those of the group ($D_{\mathcal{A}}$ and $G_{\mathcal{A}}$): it does not care about the resource fz of agent a_2 ($\top \rightarrow fz \notin D_1$) and it has as a subgoal its part of the plan y: $\neg x \rightarrow y \in G_1$ (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).

Situation 1

Group $\mathcal{A}$:
$G_{\mathcal{A}} = \{\top \rightarrow x\}$,
$D_{\mathcal{A}} = \{\top \rightarrow fy, \top \rightarrow fz, \top \rightarrow fc\}$,
$\geq_{\mathcal{A}} \supseteq \{\top \rightarrow x\} > \{\top \rightarrow fz\} > \{\top \rightarrow fc\}$,

Agent 1:
$y, c \in A_1, x, fy, fz, fc \in P$,
$s_1^0 = \{fx, fy, fc, x\}$,
$B_1 = \{y \wedge z \rightarrow x, y \rightarrow \neg fy, z \rightarrow \neg fz, c \rightarrow \neg fc\}$,
$G_1 = \{\top \rightarrow x, \neg x \rightarrow y\}$,
$D_1 = \{\top \rightarrow fy, \top \rightarrow fc\}$,
$\geq_1 \supseteq \{\top \rightarrow x\} > \{\neg x \rightarrow y\} > \{\top \rightarrow fy\} > \{\top \rightarrow fc\}$,

Agent 2:
$z \in A_2, OP_2 = A_1 \cup P^1 \setminus \{x\}$,
$s_2^0 = \{fx, fy, fc, \neg x\}$,
$B_2 = \{y \wedge z \rightarrow x, c \rightarrow x, y \rightarrow \neg fy, z \rightarrow \neg fz, c \rightarrow \neg fc\}$,
$G_2 = \{\top \rightarrow x, \neg x \rightarrow z\}$,
$D_2 = \{\top \rightarrow fz\}$,
$\geq_2 \supseteq \{\top \rightarrow x\} > \{\neg x \rightarrow z\} > \{\top \rightarrow fz\}$

Optimal decision set: $\langle d_1 = \{c\}, d_2 = \emptyset \rangle$
Expected state description:
$s_1^1 = s_2^1 = \{fy, fz, \neg fc, x, c\}$,
$s_2^2 = s_1^2 = \{fy, fz, \neg fc, x, c\}$,
Unfulfilled motivational states:
$U_1^{D_1} = \{\top \rightarrow fc\}, U_1^{G_1} = \emptyset$,
$U_2^{D_2} = \emptyset, U_2^{G_2} = \emptyset$,
$U_1^{D_{\mathcal{A}}} = \{\top \rightarrow fc\}, U_1^{G_{\mathcal{A}}} = \emptyset$,
$U_2^{D_{\mathcal{A}}} = \{\top \rightarrow fc\}, U_2^{G_{\mathcal{A}}} = \emptyset$

Since agent a_1 decides to do c, then the next state is $s_1^1 = next(s_1^0, d_1, B_1, \geq_1^B, 0) = \{fx, fy, \neg fc, x, c\}$: $\neg fc$ is true as an effect of c ($c \rightarrow \neg fc \in B_1$); agent a_1 unconditional (and hence applicable) desire $\top \rightarrow fy$ is achieved in state s_1^1 (the antecedent $\top$ of the unconditional rule $\top \rightarrow fy$ is true and also the consequent fy is), while $\top \rightarrow fc$ remains unsatisfied ($fc \notin s_1^1$). Moreover, the shared goal $\top \rightarrow x$ is satisfied and $\neg x \rightarrow y \in G_1$ is not applicable ($\neg x \notin s_1^1$).

For what concerns agent a_2, it believes that the next state is $s_2^1 = \{fx, fy, \neg fc, x, c\}$, since $\neg x$ cannot persist from the initial state s_2^0 due to the effect of c ($c \rightarrow x \in B_2$ and c can be observed, $c \in OP_2$). In state s_2^1 its part of the plan $\neg x \rightarrow z$ is not relevant and, thus, is not a goal to be satisfied anymore.

Had agent a_1's decision been $d_1' = \emptyset$ it would fulfill a_1's and group's desire to save the resource fc ($\top \rightarrow fc \in D_1 \cap D_{\mathcal{A}}$). However, it would leave agent a_2 unaware of the satisfaction of the shared goal: $s'^1_2 = \{fx, fy, fc, \neg x\}$.

How does agent a_1 take a decision between d_1 and d_1'? It compares which of its goals and desires remain unsatisfied under the light of agent a_2's decision: $d_2' = \{z\}$. Agent a_1 knows that d_2' is the optimal decision after d_1' for agent a_2 since d_2' would achieve its goal $\neg x \rightarrow z$ (which is applicable since $\neg x$ persists in s_2^1 from s_2^0). So the unfulfilled desires of the group would have been $U'^{D_{\mathcal{A}}}_1 = \{\top \rightarrow fz\}$.

Since $\geq_{\mathcal{A}} \supseteq \{\top \rightarrow fz\} > \{\top \rightarrow fc\}$ (i.e., communication is less costly than doing z) d_1 is preferred over d_1' by a cooperative agent a_1: $U'^{D_{\mathcal{A}}}_1 \geq_{\mathcal{A}} U_1^{D_{\mathcal{A}}}$.

Had agent a_1 been a selfish agent, its decision would have been d_1': $s' \leq s$ since $U_1^{D_1} = \{\top \rightarrow fc\} \geq_1 U'^{D_1}_1 = \emptyset$.

[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 a_2: $c \rightarrow x \in B_2$, but $c \rightarrow x \notin B_1$, since $c \rightarrow x \in B_1$ would mean that, according to agent a_1, c achieves x in the world.

4.2 Helpful behavior

When, due to recursive modelling, agent a_1 believes that agent a_2 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 $y \wedge z$ for achieving x is composed by an action $y \in A_1$ of agent a_1 and a parameter $z \in P$ which can be made true by agent a_2 my means of action $j \in A_2$, but only under condition p ($j \wedge p \rightarrow z \in B_2$); in the initial agreement agent a_2 has the goal of doing j for achieving z: $\neg x \wedge \neg z \rightarrow j \in G_2$.

What happens if j cannot achieve z since the precondition p is false and a_2 cannot do anything for making p true?

Situation 2

Group $\mathcal{A}$:
$G_{\mathcal{A}} = \{\top \rightarrow x\}$,
$D_{\mathcal{A}} = \{\top \rightarrow fy, \top \rightarrow fz, \top \rightarrow fh\}$,
$\geq_{\mathcal{A}} \supseteq \{\top \rightarrow x\} > \{\top \rightarrow fy, \top \rightarrow fz\} > \{\top \rightarrow fy, \top \rightarrow fz, \top \rightarrow fh\}$

Agent 1:
$y, h \in A_1, x, z, fy, fz, fh, p \in P$,
$s_1^0 = \{fx, fy, fh, \neg p\}$,
$B_1 = \{y \wedge z \rightarrow x, y \rightarrow \neg fy, j \wedge p \rightarrow z, z \rightarrow \neg fz, h \rightarrow p, h \rightarrow \neg fh\}$,
$G_1 = \{\top \rightarrow x, \neg x \rightarrow y\}$,
$D_1 = \{\top \rightarrow fy, \top \rightarrow fh\}$,
$\geq_1 \supseteq \{\top \rightarrow x\} > \{\neg x \rightarrow y\} > \{\top \rightarrow fy, \top \rightarrow fh\}$,

Agent 2:
$j \in A_2, OP_2 = A_1 \cup P^1$,
$s_2^0 = \{fx, fy, fh, \neg p\}$,
$B_2 = \{y \wedge z \rightarrow x, y \rightarrow \neg fy, j \wedge p \rightarrow z, z \rightarrow \neg fz, h \rightarrow p, h \rightarrow \neg fh\}$,
$G_2 = \{\top \rightarrow x, \neg x \rightarrow z, \neg x \wedge \neg z \rightarrow j\}$,
$D_2 = \{\top \rightarrow fz\}$,
$\geq_2 \supseteq \{\top \rightarrow x\} > \{\neg x \rightarrow z\} > \{\top \rightarrow fz\}$

Optimal decision set: $\langle d_1 = \{y, h\}, d_2 = \{j\}\rangle$
Expected state description:
$s_1^1 = s_2^1 = \{\neg fy, fz, \neg fh, p, y, h\}$,
$s_2^2 = s_1^2 = \{\neg fy, \neg fz, \neg fh, p, x, z, j\}$
Unfulfilled motivational states:
$U_1^{D_1} = \{\top \rightarrow fy, \top \rightarrow fh\}, U_1^{G_1} = \emptyset$,
$U_2^{D_2} = \{\top \rightarrow fz\}, U_2^{G_2} = \emptyset$,
$U_1^{D_{\mathcal{A}}} = \{\top \rightarrow fy, \top \rightarrow fz, \top \rightarrow fh\}, U_1^{G_{\mathcal{A}}} = \emptyset$,
$U_2^{D_{\mathcal{A}}} = \{\top \rightarrow fy, \top \rightarrow fz, \top \rightarrow fh\}, U_2^{G_{\mathcal{A}}} = \emptyset$

Agent a_1 accepts to do also action h to achieve p ($h \rightarrow p \in B_1$), so that agent a_2'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 fh than to give up the shared objective: $\geq_{\mathcal{A}} \subseteq \{\top \rightarrow x\} > \{\top \rightarrow fy, \top \rightarrow fz, \top \rightarrow fh\}$.

Sometimes, helpful behavior is not sufficient: what happens in the previous situation if agent a_2 is not aware of the contribute of agent a_1 to achieve p? If p is not observable ($OP_2 = A_1 \cup P^1 \setminus \{p\}$), then agent a_1 has to consider whether to communicate to agent a_2 that p is true by doing action c ($c \rightarrow p \in B_2$): if agent a_1 decides for $\neg c$, then it can predict that agent a_2 wrongly believes that it cannot do its part and it will give up the cooperation (correctly, from its point of view).

Situation 3

Group $\mathcal{A}$:
$G_{\mathcal{A}} = \{\top \rightarrow x\}$,
$D_{\mathcal{A}} = \{\top \rightarrow fy, \top \rightarrow fz, \top \rightarrow fh, \top \rightarrow fc\}$,
$\geq_{\mathcal{A}} \supseteq \{\top \rightarrow x\} > \{\top \rightarrow fy, \top \rightarrow fz, \top \rightarrow fh\} > \{\top \rightarrow fy, \top \rightarrow fz, \top \rightarrow fh, \top \rightarrow fc\}$

Agent 1:
$y, h, c \in A_1, x, z, fy, fz, fh, fc, p \in P$,
$s_1^0 = \{fx, fy, fh, fc, \neg p\}$,
$B_1 = \{y \wedge z \rightarrow x, y \rightarrow \neg fy, j \wedge p \rightarrow z, z \rightarrow \neg fz, h \rightarrow p, h \rightarrow \neg fh, c \rightarrow \neg fc\}$,
$G_1 = \{\top \rightarrow x, \neg x \rightarrow y\}$,
$D_1 = \{\top \rightarrow fy, \top \rightarrow fh, \top \rightarrow fc\}$,
$\geq_1 \supseteq \{\top \rightarrow x\} > \{\neg x \rightarrow y\} > \{\top \rightarrow fy, \top \rightarrow fh, \top \rightarrow fc\}$,

Agent 2:
$j \in A_2, OP_2 = A_1 \cup P^1 \setminus \{p\}$,
$s_2^0 = \{fx, fy, fh, fc, \neg p\}$,
$B_2 = \{y \wedge z \rightarrow x, y \rightarrow \neg fy, j \wedge p \rightarrow z, z \rightarrow \neg fz, h \rightarrow p, h \rightarrow \neg fh, c \rightarrow \neg fc, c \rightarrow p\}$,
$G_2 = \{\top \rightarrow x, \neg x \rightarrow z, \neg x \wedge \neg x \rightarrow j\}$,
$D_2 = \{\top \rightarrow fz\}$,
$\geq_2 \supseteq \{\top \rightarrow x\} > \{\neg x \rightarrow z\} > \{\top \rightarrow fz\}$

Optimal decision set: $\langle d_1 = \{y, h, c\}, d_2 = \{j\}\rangle$
Expected state description:
$s_1^1 = s_2^1 = \{\neg fy, fz, \neg fh, \neg fc, p, y, h, c\}$,
$s_2^2 = s_1^2 = \{\neg fy, \neg fz, \neg fh, \neg fc, p, z, x, j\}$
Unfulfilled motivational states:
$U_1^{D_1} = \{\top \rightarrow fy, \top \rightarrow fh, \top \rightarrow fc\}, U_1^{G_1} = \emptyset$,
$U_2^{D_2} = \{\top \rightarrow fz\}, U_2^{G_2} = \emptyset$,
$U_1^{D_{\mathcal{A}}} = \{\top \rightarrow fy, \top \rightarrow fz, \top \rightarrow fh, \top \rightarrow fc\}, U_1^{G_{\mathcal{A}}} = \emptyset$,
$U_2^{D_{\mathcal{A}}} = \{\top \rightarrow fy, \top \rightarrow fz, \top \rightarrow fh, \top \rightarrow fc\}, U_2^{G_{\mathcal{A}}} = \emptyset$

Helpful behavior should be constrained to the adoption of goals which contribute to the shared goal of the group. However, no explicit constraint is present in our model. Rather, this constraint emerges due to the fact that, when agent a_1 recursively model agent a_2, agent a_1 attributes to agent a_2 a cooperative agent type; in this way, agent a_1 is certain that its decision to help a_2 will contribute to the shared goal, since agent a_2 will give precedence to the satisfaction of shared goals.

4.3 Conflict avoidance

When agents have the possibility to 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.

In the next scenario agent a_1 can achieve its part of the shared plan $y \in P$ (a parameter) by doing $j \in A_1$ or $k \in A_1$; action k is less costly than j: $\geq_1 \supseteq \{\top \rightarrow fj\} > \{\top \rightarrow fk\}$ and $\{\top \rightarrow fj, \top \rightarrow fk\} \subseteq D_1$ (but the two desires do not belong to $D_{\mathcal{A}}$). However, if k is true, the agent a_2 cannot achieve its goal $z \in P$ (a parameter) by doing action $h \in A_2$: $h \rightarrow z \in B_2$ but $h \wedge k \rightarrow \neg z \in B_2$ and the second rule is an exception to the first one since it has priority over the other: $\geq_2^B \subseteq \{h \wedge j \rightarrow \neg z\} > \{h \rightarrow z\}$.

Situation 4

Group $\mathcal{A}$:

$G_{\mathcal{A}} = \{\top \rightarrow x\}$,
$D_{\mathcal{A}} = \{\top \rightarrow fy, \top \rightarrow fz\}$,
$\geq_{\mathcal{A}} \supseteq \{\top \rightarrow x\} > \{\top \rightarrow fy, \top \rightarrow fz\}$

Agent 1:

$j, k \in A_1, x, y, z, fy, fz, fj, fk \in P$,
$s_1^0 = \{fx, fy, fj, fk\}$,
$B_1 = \{y \wedge z \rightarrow x,\ y \rightarrow \neg fy,\ j \rightarrow y,\ k \rightarrow y,$
$z \rightarrow \neg fz, j \rightarrow \neg fj, k \rightarrow \neg fk, h \rightarrow z, h \wedge k \rightarrow \neg z\}$,
$\geq_1^B \supseteq \{h \wedge k \rightarrow \neg z\} > \{h \rightarrow z\}$,
$G_1 = \{\top \rightarrow x, \neg x \rightarrow y\}$,
$D_1 = \{\top \rightarrow fy, \top \rightarrow fj, \top \rightarrow fk\}$,
$\geq_1 \supseteq \{\top \rightarrow x\} > \{\neg x \rightarrow y\} > \{\top \rightarrow fy, \top \rightarrow fj\} > \{\top \rightarrow fy, \top \rightarrow fk\}$,

Agent 2:

$h \in A_2, OP_2 = A_1 \cup P^1$,
$s_2^0 = \{fx, fy, fj, fk\}$,
$B_2 = \{y \wedge z \rightarrow x,\ y \rightarrow \neg fy,\ j \rightarrow y,\ k \rightarrow y,$
$z \rightarrow \neg fz, j \rightarrow \neg fj, k \rightarrow \neg fk, h \rightarrow z, h \wedge k \rightarrow \neg z\}$,
$\geq_2^B \supseteq \{h \wedge k \rightarrow \neg z\} > \{h \rightarrow z\}$,
$G_2 = \{\top \rightarrow x, \neg x \rightarrow z\}$,
$D_2 = \{\top \rightarrow fz\}$,
$\geq_2 \supseteq \{\top \rightarrow x\} > \{\neg x \rightarrow z\} > \{\top \rightarrow fz\}$

Optimal decision set: $\langle d_1 = \{j\}, d_2 = \{h\}\rangle$
Expected state description:
$s_1^1 = s_2^1 = \{\neg fy, fz, \neg fj, y, j\}$,
$s_2^2 = s_1^2 = \{\neg fy, \neg fz, \neg fj, y, z, x, h\}$
Unfulfilled motivational states:
$U_1^{D_1} = \{\top \rightarrow fy, \top \rightarrow fj\}, U_1^{G_1} = \emptyset$,
$U_2^{D_2} = \{\top \rightarrow fz\}, U_2^{G_2} = \emptyset$,
$U_1^{D_{\mathcal{A}}} = \{\top \rightarrow fy, \top \rightarrow fz\}, U_1^{G_{\mathcal{A}}} = \emptyset$,
$U_2^{D_{\mathcal{A}}} = \{\top \rightarrow fy, \top \rightarrow fz\}, U_2^{G_{\mathcal{A}}} = \emptyset$

4.4 Ending cooperation

When agent a_1, 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 [11] that the group terminates when there is the mutual belief that every agent believes that the shared goal has been achieved. We analyze a scenario similar to Situation 1: x has already been achieved, and, this time, both agents are aware of this fact. So no communication is necessary and cooperation ends without leaving any goal of the group unsatisfied.

Situation 5

Group $\mathcal{A}$:

$G_{\mathcal{A}} = \{\top \rightarrow x\}$,
$D_{\mathcal{A}} = \{\top \rightarrow fy, \top \rightarrow fz, \top \rightarrow fc\}$,
$\geq_{\mathcal{A}} \supseteq \{\top \rightarrow x\} > \{\top \rightarrow fy, \top \rightarrow fz, \top \rightarrow fc\}$,

Agent 1:

$y, c \in A_1, x, fy, fz, fc \in P$,
$s_1^0 = \{fx, fy, fc, x\}$,
$B_1 = \{y \wedge z \rightarrow x, y \rightarrow \neg fy, z \rightarrow \neg fz, c \rightarrow \neg fc\}$,
$G_1 = \{\top \rightarrow x, \neg x \rightarrow y\}$,
$D_1 = \{\top \rightarrow fy, \top \rightarrow fc\}$,
$\geq_1 \supseteq \{\top \rightarrow x\} > \{\neg x \rightarrow y\} > \{\top \rightarrow fy\} > \{\top \rightarrow fc\}$,

Agent 2:

$z \in A_2, OP_2 = A_1 \cup P^1 \setminus \{x\}$,
$s_2^0 = \{fx, fy, fc, x\}$,
$B_2 = \{y \wedge z \rightarrow x, c \rightarrow x, y \rightarrow \neg fy, z \rightarrow \neg fz, c \rightarrow \neg fc\}$,
$G_2 = \{\top \rightarrow x, \neg x \rightarrow z\}$,
$D_2 = \{\top \rightarrow fz\}$,
$\geq_2 \supseteq \{\top \rightarrow x\} > \{\neg x \rightarrow z\} > \{\top \rightarrow fz\}$
Optimal decision set: $\langle d_1 = \emptyset, d_2 = \emptyset\rangle$
Expected state description:
$s_1^1 = s_2^1 = \{c, fy, fz, fc, x\}, s_2^2 = s_1^2 = \{fy, fz, fc, x\}$,
Unfulfilled motivational states:
$U_1^{D_1} = \emptyset, U_1^{G_1} = \emptyset$,
$U_2^{D_2} = \emptyset, U_2^{G_2} = \emptyset$,
$U_1^{D_{\mathcal{A}}} = \emptyset, U_1^{G_{\mathcal{A}}} = \emptyset$,
$U_2^{D_{\mathcal{A}}} = \emptyset, U_2^{G_{\mathcal{A}}} = \emptyset$

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_1 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 to waste the resources of the group. For example, return on Situation 1, assuming this time that agent a_1 knows that its attempt to communicate to a_2 that the shared goal x has been achieved will fail, since a precondition g does not hold and agent a_1 cannot do anything for making it true: $\neg g \in s_1^0$, $c \wedge g \rightarrow \neg x \in B_2$ and $\{c \wedge g \rightarrow \neg x\} > \{c \rightarrow x\}$.

4.5 Defeating cooperation

In the previous scenarios we assumed always cooperative agent types. This unrealistic assumption must be released: in a community of heterogeneous agents it is possible that some agents take advantage of the cooperation only as long as it is fruitful for them. In the next scenario we consider a variation of Situation 1 where now agent a_1 has a selfish agent type: it takes decisions without giving precedence to the motivations of the group. The shared goal (which is also its private goal) has been achieved: communicating this fact to a_2 has a cost for agent a_1 while the cost faced by the community due to the waste of resources by agent a_2 ($\top \rightarrow fz \in D_{\mathcal{A}}$) is not a desire of agent a_1. Hence, agent a_1 decides to give up cooperation even if the group still needs its contribution:

Situation 6
Group $\mathcal{A}$:
$G_{\mathcal{A}} = \{\top \rightarrow x\}$,
$D_{\mathcal{A}} = \{\top \rightarrow fy, \top \rightarrow fz, \top \rightarrow fc\}$,
$\geq_{\mathcal{A}} \supseteq \{\top \rightarrow x\} > \{\top \rightarrow fz\} > \{\top \rightarrow fc\}$,
Agent 1:
$y, c \in A_1, x, fy, fz, fc \in P$,
$s_1^0 = \{fx, fy, fc, x\}$,
$B_1 = \{y \wedge z \rightarrow x, y \rightarrow \neg fy, z \rightarrow \neg fz, c \rightarrow \neg fc\}$,
$G_1 = \{\top \rightarrow x, \neg x \rightarrow y\}$,
$D_1 = \{\top \rightarrow fy, \top \rightarrow fc\}$,
$\geq_1 \supseteq \{\top \rightarrow x\} > \{\neg x \rightarrow y\} > \{\top \rightarrow fy\} > \{\top \rightarrow fc\}$,
Agent 2:
$z \in A_2, OP_2 = A_1 \cup P^1 \setminus \{x\}$,
$s_2^0 = \{fx, fy, fc, \neg x\}$,
$B_2 = \{y \wedge z \rightarrow x, c \rightarrow x, y \rightarrow \neg fy, z \rightarrow \neg fz, c \rightarrow \neg fc\}$,
$G_2 = \{\top \rightarrow x, \neg x \rightarrow z\}$,
$D_2 = \{\top \rightarrow fz\}$,
$\geq_2 \supseteq \{\top \rightarrow x\} > \{\neg x \rightarrow z\} > \{\top \rightarrow fz\}$
Optimal decision set: $\langle d_1 = \emptyset, d_2 = \{z\}\rangle$
Expected state description:
$s_1^1 = \{fy, fz, fc, x\}$,
$s_2^1 = \{fy, fz, fc, \neg x\}$,
$s_2^2 = s_1^2 = \{fy, \neg fz, fc, \neg x, z\}$,
Unfulfilled motivational states:
$U_1^{D_1} = \emptyset, U_1^{G_1} = \emptyset$,
$U_2^{D_2} = \{\top \rightarrow fz\}, U_2^{G_2} = \{\top \rightarrow x\}$,
$U_1^{D_{\mathcal{A}}} = \{\top \rightarrow fz\}, U_1^{G_{\mathcal{A}}} = \emptyset$,
$U_2^{D_{\mathcal{A}}} = \{\top \rightarrow fz\}, U_2^{G_{\mathcal{A}}} = \{\top \rightarrow x\}$

If non cooperative agents do not guarantee not to abandon the group, how can cooperation be ensured? The notion of the group's motivations represents an optimum which each agent should stick to. Joining a group creates an obligation towards the partners to stick to this optimum. Departing from this optimum represents a violation of the social commitment of the agent towards the other partners. A violation of this obligation can be sanctioned by the other agents when they become aware of the uncooperative behavior.

We can model social commitment, since our framework is inspired to [2] who model obligations and normative reasoning in multiagent systems. In brief, in [2], an obligation is associated with a sanction: when an agent is aware of a violation, it has the goal of considering the other agent as a violator and to sanction it; the sanction is an action which is not desired by the bearer of the obligation.

We include in the following scenario the obligation of agent a_1 to be cooperative, otherwise it is sanctioned by a_2 by doing $s \in A_2$ ($\top \rightarrow \neg s \in D_1$): $O_{1,2}(coop(a_1), s)$. $coop(a_1)$ is a parameter which is true after the decision d_1 of agent a_1 if d_1 is the same as a decision d_1' taken as if a_1 were a cooperative agent. If agent a_1 is not cooperative, agent a_2 will sanction it ($\neg coop(a_1) \rightarrow s \in G_2$). Here, the non cooperative agent a_1 decides to abandon the group since the main goal has become irrelevant for it ($r \rightarrow x \in G_1$ and $\neg r \in s_1^0$); its decision d_1 is not the decision d_1' that a cooperative agent would have taken in the same situation: $d_1 \neq d_1' = \{y\}$. However, the sanction does not prevent agent a_1 from exiting the group: it prefers to be sanctioned with respect to doing its part y of the plan: $\geq_1 \{\top \rightarrow fy\} > \{\top \rightarrow \neg s\}$.

Situation 7
Group $\mathcal{A}$:
$G_{\mathcal{A}} = \{\top \rightarrow x\}$,
$D_{\mathcal{A}} = \{\top \rightarrow fy, \top \rightarrow fz, \top \rightarrow fc\}$,
$\geq_{\mathcal{A}} \supseteq \{\top \rightarrow x\} > \{\top \rightarrow fz\} > \{\top \rightarrow fc\}$,
Agent 1:
$y \in A_1, x, fy, fz \in P$,
$s_1^0 = \{fx, fy, \neg x, \neg r\}$,
$B_1 = \{y \wedge z \rightarrow x, y \rightarrow \neg fy, z \rightarrow \neg fz\}$,
$G_1 = \{r \rightarrow x, r \wedge \neg x \rightarrow y\}$,
$D_1 = \{\top \rightarrow fy, \top \rightarrow \neg s\}$,
$\geq_1 \supseteq \{r \rightarrow x\} > \{\neg x \rightarrow y\} > \{\top \rightarrow fy\} > \{\top \rightarrow \neg s\}$
Agent 2:
$z, s \in A_2, OP_2 = A_1 \cup P^1 \setminus \{x\}$,
$s_2^0 = \{fx, fy, \neg x, \neg r\}$,
$B_2 = \{y \wedge z \rightarrow x, y \rightarrow \neg fy, z \rightarrow \neg fz\}$,
$G_2 = \{\top \rightarrow x, \neg x \rightarrow z, \neg coop(a_1) \rightarrow s\}$,
$D_2 = \{\top \rightarrow fz\}$,
$\geq_2 \supseteq \{\top \rightarrow x\} > \{\neg x \rightarrow z\} > \{\top \rightarrow fz\}$
Optimal decision set: $\langle d_1 = \emptyset, d_2 = \{s\}\rangle$
Expected state description:
$s_1^1 = s_2^1 = \{fy, fz, \neg r, \neg x, \neg coop(a_1)\}$,
$s_2^2 = s_1^2 = \{fy, fz, \neg r, \neg x, s\}$,
Unfulfilled motivational states:
$U_1^{D_1} = \{\top \rightarrow \neg s\}, U_1^{G_1} = \emptyset$,
$U_2^{D_2} = \emptyset, U_2^{G_2} = \{\top \rightarrow x\}$,
$U_1^{D_{\mathcal{A}}} = \emptyset, U_1^{G_{\mathcal{A}}} = \{\top \rightarrow x\}$,
$U_2^{D_{\mathcal{A}}} = \emptyset, U_2^{G_{\mathcal{A}}} = \{\top \rightarrow x\}$

5 Summary and concluding remarks

In this paper we show how a qualitative game theory can be used to model cooperation among BDI agents. Rather than basing decisions on a quantitative decision theory, the agents are assumed to decide basing on their goals and desires. Moreover, they recursively model the decisions of their partners to predict the result of their actions.

We do not reduce cooperation to the individual attitudes of the members, but we assume that the group can be considered as an agent: each member of the group has to adopt the goals and desires attributed to the group agent when it takes a decision. The group, however, is not a real agent, but an entity belonging to the social reality and constructed by the agents when they join together. This model allows to explain cooperation phenomena like communication, helpful behavior, conflict avoidance, correct termination of cooperation, and commitment to the group.

In this paper, we attribute goals and desires to the group, but not beliefs. According to Tuomela [26], it is possible to attribute also beliefs to a group to represent what the members collectively accept.

The logical formalism makes precise the notions of beliefs, desires and goals used informally in [1]. On the other hand, [1] consider also uncertainty in the world, nondeterministic actions and sensing actions, so that further phenomena can be modelled, like, e.g., unreliable communication and monitoring of the partners' behavior. Moreover, here we do not consider the problem of planning, but we only compare the possible decisions, while [1] use for this purpose an extension of the DRIPS planner ([18]).

Related work is [2], [4] and [5] 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 the normative system. In [3] a similar model is formalized using the standard BDI_{CTL} logic [21] for agent verification.

Another related work is Dastani and van der Torre [13] who consider the notion of joint goal in a qualitative decision theory. They show that groups of agents which end up in equilibria act as if they maximize joint goals.

References

[1] G. Boella, R. Damiano, and L. Lesmo. Cooperation and group utility. In *Intelligent Agents VI*, pages 319–333, Berlin, 2000. Springer Verlag.

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

[3] G. Boella and L. van der Torre. Game specification in the trias politica. In *Procs. of BNAIC'03*, 2003.

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

[5] G. Boella and L. van der Torre. Norm governed multiagent systems: The delegation of control to autonomous agents. In *Procs. of IEEE/WIC IAT Conference*. IEEE Press, 2003.

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

[7] M. E. Bratman. *Faces of intention : selected essays on intention and agency*. Cambridge Univ. Press, Cambrdige (UK), 1999.

[8] 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.

[9] C. Castelfranchi. Commitment: from intentions to groups and organizations. In *Proc. of ICMAS-96*, Cambridge (MA), 1996. AAAI/MIT Press.

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

[11] P. R. Cohen and H. J. Levesque. Confirmation and joint action. In *Proc. 12th IJCAI*, pages 951–957, Sydney, 1991.

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

[13] M. Dastani and L. van der Torre. What is a joint goal? Games with beliefs and defeasible desires. In *Procs. of NMR02*, Toulouse, 2002.

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

[15] P. J. Gmytrasiewicz and E. H. Durfee. Formalization of recursive modeling. In *Procs. of first ICMAS-95*, 1995.

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

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

[18] P. Haddawy and S. Hanks. Utility models for goal-directed, decision-theoretic planners. *Computational Intelligence*, 14:392–429, 1998.

[19] J. Lang, L. van der Torre, and E. Weydert. Utilitarian desires. *Autonomous agents and Multi-agent systems*, pages 329–363, 2002.

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

[21] A. S. Rao and M. P. Georgeff. Decision procedures for BDI logics. *Journal of Logic and Computation*, 8(3):293–343, 1998.

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

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

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

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

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

[27] R. Tuomela and K. Miller. We-intentions, free-riding and being in reserve. *Erkenntnis*, 36:25–52, 1992.