Negotiating the Distribution of Obligations with Sanctions among Autonomous Agents

Guido Boella[1] and **Leendert van der Torre**[2]

Abstract. In this paper we study the distribution of obligations together with their associated sanctions among agents belonging to collective entities like groups and organizations. We model the distribution as a negotiation process among the agents, we model the behavior of the agents in a qualitative game theory, and we formalize them in a logical framework. We characterize collective obligations according to the way the responsibility in case of violation is attributed to individual agents or to the whole set of agents, where we distinguish among violations during the negotiation and during the execution of the task. We also show that in some cases it is a drawback to be the only agent able to see to the fulfilment of part of an obligation, but in other cases it may be an advantage, because of the power it gives to the agent during the negotiation.

1 Introduction

Organizational concepts are introduced in multiagent systems to control their emergent behavior. Agents are organized in groups [10] or assigned to roles, and obligations are assigned to agents based on their roles' responsibilities. Obligations are assigned also to collective entities like groups and organizations, for example, when these entities are recognized by the law as legal persons [13], or when obligations are delegated in an organization in a top down manner to the departments. The distribution of group and organizational obligations raises the question of which agents are responsible in case of violation and which ones are sanctioned.

In this paper we study how an obligation with explicit sanctions directed to a set of agents is distributed among them. We provide a formal model in which the distribution of obligations is the result of a negotiation process [14, 17] among the agents, where the negotiation model makes the dependencies between agents explicit. One issue we study in our framework is how the distribution of obligations – comparable to the distribution of group, joint and social goals – is related to the distribution of responsibility and sanctions, and another issue we study in our framework is the role played by the abilities of the agents in the distribution process.

Dignum and Royakkers [9] discuss the difference between general obligations, e.g., "cyclists have to give way to motor vehicles", and collective ones, e.g., "John and Paul have to move the table", and how the obligation is distributed to the agents, e.g., "John takes one end of the table and Paul the other". Cholvy and Garion [8] refine this approach and claim that if there is a set of agents subject to an obligation to perform some task, then "the derivation of individual obligations from collective obligations depends on several parameters, among which the ability of the agents". If an agent is the only one able to perform a part of that task, then it is obliged to do that part and it is also obliged to do that towards the other members of the set. Their motivating example is based on three children who are obliged by their mother to prepare the table for dinner. The oldest child is the only one who is tall enough to get the glasses on the cupboard. The whole group is responsible for the violation of the collective obligation, but in case the violation is due to the fact that the oldest boy did not bring the glasses, only he "can be taken responsible by the group because he was the only one able to take the glasses".

The formal framework we introduce can be used for a more fine grained analysis of the example, including the following issues.

- It considers various task distributions, because the agents have to find an agreement on how to distribute the tasks. This agreement is the result of a negotiation which has to take into account the dependence relations among the agents. E.g., the kids are first obliged to negotiate the distribution of the obligation, and then, only if they find a successful distribution of obligations, the oldest kid becomes obliged to see to the glasses.
- It distinguishes, on the one hand, between the obligations and the sanctions associated with the violations of these obligations, including violations during the distribution process. In the example, if the negotiation is unsuccessful, then the children have violated the negotiation obligation and are punished, but the oldest child is not punished for not seeing to the glasses.
- It explains the apparent conflict between the obligation of the oldest boy and the analysis based on dependence and power structures, as is common in social theory and which has been promoted in agent theory by Castelfranchi [7]. According to this alternative view, the ability possessed by only one agent makes the remaining agents depend on him, since they lack the power to do part of the task they are obliged to. The oldest boy is in the best position, rather than having an additional burden and being sanctioned both for not respecting the collective obligation and his own obligation; the reason is that he has more power in the negotiation for the distribution of the task, and, by exercising this power, he may end up doing less than the other boys.

We focus on "sets of agents" rather than "groups of agents" to avoid the assumption that a group is cooperating to a shared goal.

The paper is organized as follows. In Section 2 we introduce our conceptual model of normative multiagent systems, in Section 3 we introduce the negotiation model, and in Section 4 we discuss the role of the agent's abilities with respect to the dependencies between the agents, and we discuss the classification of collective obligations.

[1] Dipartimento di Informatica - Università di Torino - Italy, email: guido@di.unito.it

[2] CWI Amsterdam and Delft University of Technology - The Netherlands, email: torre@cwi.nl

2 Normative multiagent system

The conceptual model consists of a set of agents (A), which are described (AD) by a set of boolean variables (X), including *decision variables* representing the actions they can perform, and desires (D) guiding their decision making. Desires can be conflicting, and the way the agents resolve their conflicts is described by a priority relation ($\geq$) that expresses their agent characteristics [5]. The priority relation is defined on the powerset of the desires such that a wide range of characteristics can be described.

Definition 1 (AS) *An agent set is a tuple $\langle A, X, D, AD, \geq \rangle$:*

- *the agents A, variables X, desires D are three finite disjoint sets.*
- *an agent description $AD : A \to 2^{X \cup D}$ is a complete function that maps each agent to sets of variables (its decision variables) and desires. For each agent $a \in A$, we write X_a for $X \cap AD(a)$ and D_a for $D \cap AD(a)$. We write parameters $P = X \setminus \cup_{a \in A} X_a$.*
- *a priority relation $\geq$: $A \to 2^D \times 2^D$ is a function from agents to a transitive and reflexive relation on the powerset of the desires containing at least the subset relation. We write $\geq_a$ for $\geq(a)$.*

Desires are abstract concepts which are described by – though conceptually not identified with – rules (R) built from literals (L). Background knowledge and integrity constraints are formalized by a set of effect rules (E).

Definition 2 (MAS) *A multiagent system is a tuple $\langle A, X, D, AD, E, MD, \geq \rangle$, where $\langle A, X, D, AD, \geq \rangle$ is an agent set, and:*

- *the set of literals built from X, written as $Lit(X)$, is $X \cup \{\neg x \mid x \in X\}$, and the set of rules built from X, written as $Rul(X) = 2^{Lit(X)} \times Lit(X)$, is the set of pairs of a set of literals built from X and a literal built from X, written as $\{l_1, \ldots, l_n\} \to l$. We also write $l_1 \wedge \ldots \wedge l_n \to l$ and when $n = 0$ we write $\top \to l$. Moreover, for $x \in X$ we write $\sim x$ for $\neg x$ and $\sim\neg x$ for x.*
- *the set of effects $E \subseteq Rul(X)$ is a set of rules built from X.*
- *the motivational description $MD : D \to Rul(X)$ is a complete function from the sets of desires to the set of rules built from X. For $S \subseteq D$, we write $MD(S) = \{MD(s) \mid s \in S\}$.*

The following example represents our running example of the boys preparing the table as a multiagent system. For notational convenience we sometimes write agents as parameter of a proposition, e.g., $get\_glasses(boy_1)$, which does not mean that we have quantification in the logic, but which just means that the whole expression is a single proposition.

Example 1 *The following multiagent system MAS contains three boys that can prepare a table.*

$A = \{boy_1, boy_2, boy_3\}$
$X = \{prepared\_table, early\_to\_bed, get\_glasses, get\_drinks, get\_sheet, get\_forks, get\_dishes, get\_glasses(a), get\_drinks(a), get\_sheet(a), get\_forks(a), get\_dishes(a) \mid a \in A\}$
$X_{boy_1} = \{get\_glasses(boy_1), get\_drinks(boy_1), get\_sheet(boy_1), get\_forks(boy_1), get\_dishes(boy_1)\}$
$X_a = \{get\_drinks(a), get\_sheet(a), get\_forks(a), get\_dishes(a)\}$ *for* $a \in \{boy_2, boy_3\}$
$MD(D_a) \supseteq \{\top \to \neg get\_glasses(a), \top \to \neg get\_drinks(a), \top \to \neg get\_sheet(a), \top \to \neg get\_forks(a), \top \to \neg get\_dishes(a), \top \to \neg early\_to\_bed\}$ *for* $a \in A$
$E = \{get\_glasses(a) \to get\_glasses, get\_drinks(a) \to get\_drinks, get\_sheet(a) \to get\_sheet, get\_forks(a) \to get\_forks, get\_dishes(a) \to get\_dishes, get\_glasses \wedge get\_drinks \wedge get\_sheet \wedge get\_forks \wedge get\_dishes \to prepared\_table \mid a \in A\}$
$\geq_a \supseteq \{\top \to \neg early\_to\_bed\} > \{\top \to \neg get\_glasses(a)\} > \{\top \to \neg get\_drinks(a)\} > \ldots$ *for each* $a \in \{boy_1, boy_2, boy_3\}$.

We now extend the multiagent system to a normative multiagent system to take norms into account and formalize collective obligations. To describe the normative system, we introduce a set of norms (N) and a norm description that associates violations with variables (V).

Definition 3 (NMAS) *A normative multiagent system $NMAS$ is a tuple $\langle A, X, D, AD, E, MD, \geq, N, V \rangle$ where $MAS = \langle A, X, D, AD, E, \geq \rangle$ is our multiagent system, and moreover:*

- *the norms N is a set disjoint from A, X and D.*
- *the norm description $V : N \times 2^A \to P$ is a function that maps pairs of a norm and a set of agents to parameters, where $V(n, \{a_1, \ldots, a_i\})$ represents that the parameter counts as a violation of the norm n by the set of agents $\{a_1, \ldots, a_i\} \subseteq A$.*

We define sanction-based obligations in the normative multiagent system using an extension of Anderson's well-known reduction [1, 12] which we discuss in [3, 4]: violations and sanctions are consequences of not fulfilling a norm.

Definition 4 (Obligation) *Let $NMAS = \langle A, X, D, AD, E, MD, \geq, N, V \rangle$. We say that $NMAS \models O_G(x, s \mid Y)$ where $x \in Lit(X)$, $s \in Lit(P)$, $Y \subseteq Lit(X)$ and $G \subseteq A$ iff $\exists n \in N$ such that:*

- *$Y \wedge \sim x \to V(n, G) \in E$: there is a violation of norm n by the set of agents G if x is false in context Y.*
- *$V(n, G) \to s \in E$: the violation is sanctioned with s.*
- *$Y \to \sim s \in MD(D_a)$ for each $a \in G$: the sanction is disliked by each agent of the set G.*

If the set G of the addresses of the obligation is a singleton, then the obligation is individual, otherwise it is a collective obligation. The difference with a general obligation like "cyclists have to give way to motor vehicles" is that the $\sim x$ is a violation of the entire set of agents as a whole and that the sanction s is disliked by all the agents. An individual obligation requires instead that the violation is attributed to a single agent ($Y \wedge \sim x \to V(n, \{a\}) \in E$) and that the sanction is disliked by the agent who committed the violation ($Y \to \sim s_a \in MD(D_a)$).

The following example illustrates the normative multiagent system in the running example.

Example 2 *We extend the multiagent system MAS to a normative multiagent system $NMAS$ imposing the obligation to prepare the table to the children or else they are sanctioned by going to bed early:*
$N = \{n\}$
$V = \{V(n, \{boy_1, boy_2, boy_3\})\}$
$E = \{\neg prepared\_table \to V(n, \{boy_1, boy_2, boy_3\}), V(n, \{boy_1, boy_2, boy_3\}) \to early\_to\_bed\}$
hence,
$NMAS \models O_{\{boy_1, boy_2, boy_3\}}(prepared\_table, early\_to\_bed \mid \top)$

A decision of agents is a set of literals that does not lead to a contradiction. To define this notion, we have to introduce a logic for rules, for which we use a simple input/output logic [11].

Definition 5 *Let $B \subseteq Rul(X)$ be a set of rules, and $C \subseteq Lit(X)$ a set of literals. The output of B applied to C, written as $out(B, C)$, is $\cup_{i=0\ldots\infty} out_i(B, C)$ with $out_0(B, C) = C$ and $out_{i+1}(B, C) = out_i(B, C) \cup \{l \mid l_1 \wedge \ldots \wedge l_n \rightarrow l \in B, \{l_1, \ldots, l_n\} \subseteq out_i(B, C)\}$. B is consistent in context C, written as $cons(B \mid C)$, iff there do not exist two contradictory literals p and $\neg p$ in $out(B, C)$.*

Agents evaluate states of affairs consequent to decisions according to which desires remain unfulfilled. There are various ways in which the preference relation can be lifted to a preference relation on sets of states; here we assume that every state of the former has to be preferred to each state of the latter. The set of optimal decisions is a subset of the decisions, which depend on the agents' preference relation, which, in turn, depends on the agents' desires. In this paper we do not detail the derivation of the optimal decisions from the preference relation; such definitions can be based on equilibria or on recursive modelling, and are discussed elsewhere [2, 3, 4].

Definition 6 *Let NMAS be a normative multiagent system.*

- *The set of decisions Δ in context $C \subseteq Lit(X)$ is $\Delta = \{\delta \subseteq Lit(X \setminus P) \mid cons(E \mid \delta \cup C)\}$*
- *Agent $a \in A$ prefers a state of affairs $S_1 \subseteq Lit(X)$ to another one $S_2 \subseteq Lit(X)$ iff $U(S_2, a) >_a U(S_1, a)$, where $U(S, a) = \{d \in D_a \mid MD(d) = L \rightarrow l, L \subseteq S$ and $l \notin S\}$.*
- *The preference relation $\succ$: $A \rightarrow 2^{2^{Lit(X)}} \times 2^{2^{Lit(X)}}$ of the agents is a relation on sets of sets of literals such that $T_1 \succ_a T_2$ iff $\forall S_1 \in T_1, S_2 \in T_2 : U(S_2, a) >_a U(S_1, a)$.*
- *The set of optimal decisions is a subset of Δ.*

The preference relation is used to evaluate the results of the negotiation, which is introduced in the following section. The next example illustrates decisions in the normative multiagent system.

Example 3 *Consider two alternative decisions, where we assume that all variables not explicitly mentioned are false (i.e., we assume the closed world assumption for decisions):*
$\delta_1 = \{get\_glasses(boy_1), get\_drinks(boy_2), get\_sheet(boy_2), get\_forks(boy_3), get\_dishes(boy_3)\}$
$\delta_2 = \{get\_forks(boy_3), get\_dishes(boy_3)\}$

Their consequent states of affairs are, respectively:
$S_1 = out(E, \delta_1) = \{prepared\_table, get\_glasses(boy_1), get\_drinks(boy_2), get\_sheet(boy_2), get\_forks(boy_3), get\_dishes(boy_3)\}$
$S_2 = out(E, \delta_2) = \{get\_forks(boy_3), get\_dishes(boy_3), V(n, \{boy_1, boy_2, boy_3\}), early\_to\_bed\}$

If we consider agent boy_1, given that it does not desire neither getting the glasses nor going to bed early ($MD(D_{boy_1}) \supseteq \{\top \rightarrow \neg get\_glasses(boy_1), \top \rightarrow \neg early\_to\_bed\}$), its unfulfilled desires are:
$U(S_1, boy_1) = \{\top \rightarrow get\_glasses(boy_1)\}$
$U(S_2, boy_1) = \{\top \rightarrow early\_to\_bed\}$

Since, from Example 1, $U(S_2, boy_1) >_{boy_1} U(S_1, boy_1)$, agent $boy_1 \in A$ prefers S_1 over S_2.

In the example, there is no incentive for the agents to choose one particular distribution of the decisions which lead to $prepare\_table$. In the following section we therefore describe the negotiation protocol to distribute the obligation among the agents.

3 Negotiation protocol

A negotiation protocol is described by a sequence of negotiation actions which either lead to success or failure. In this paper we only consider protocols in which the agents propose a so-called deal, and when an agent has made such a proposal, then the other agents can either accept or reject it. Moreover, they can also end the negotiation process without any result.

Definition 7 (Protocol) *A negotiation protocol is a tuple $NP = \langle Ag, deals, actions, valid, finished, broken \rangle$ where:*

- *the agents Ag, $deals$ and $actions$ are three disjoint sets, such that $actions = \{propose(a, d), accept(a, d), reject(a, d) \mid a \in Ag, d \in deals\} \cup \{breakit(a) \mid a \in Ag\}$.*
- *$valid$, $finished$, $broken$ are sets of finite sequences of $actions$.*

Given a normative multiagent system, the negotiation protocol for achieving an agreement on an obligation is an instantiation of a deal-based protocol. We assume that the agents involved are ordered ($\leq$), that a sequence of actions (a history) is valid when each agent does an action respecting this order. More precisely, one agent after the other has to make a proposal for distributing the obligation (τ_δ) and distributing the sanction (τ_σ), where the distribution of the obligation has not been proposed before by this agent. Then, after each proposal, the other agents have to accept or reject this proposal, again respecting the order, until they all accept it or one of them rejects it. When it is an agent's turn to make a proposal, it can also end the negotiation by breaking it. The history is $finished$ when all agent have accepted the last deal, and $broken$ when the last agent has ended the negotiations. For simplicity we assume that each agent has to contribute to the achievement of the task, that an agent can only propose or accept tasks it can execute ($\tau_\delta(a) \subseteq Lit(X_a)$), and that only deals can be proposed that are not partial and lead to the achievement of the task to be distributed.

Definition 8 (NMAS protocol) *Given a normative multiagent system $NMAS = \langle A, X, D, AD, E, MD, \geq, N, V \rangle$, a negotiation protocol for $NMAS \models O_G(x, s \mid Y)$ and total order $\leq$ is a $NP = \langle Ag, deals, actions, valid, finished, broken \rangle$, where:*

- *the set of agents $Ag = G \subseteq A$ consists of the agents which have to distribute the task $g = Y \rightarrow x$;*
- *$\leq \subseteq Ag \times Ag$ is a total order on Ag,*
- *the set of deals $\tau(g)$ is the set of pairs $\langle \tau_\delta, \tau_\sigma \rangle$, where τ_δ is a set of decisions $\tau_\delta : A \rightarrow 2^{Lit(X)}$ such that the task g is satisfied by the consequences of the consistent decision $\delta = \cup_{a \in Ag} \tau_\delta(a)$: $g \in out(E, \delta \cup Y)$ and $\forall a \in Ag : \delta_a \neq \emptyset$ and $cons(E, \delta \cup Y)$ together with a sanction $\tau_\sigma : A \rightarrow Lit(X)$ that maps each agent a to a sanction ($Y \rightarrow\sim \tau_\sigma(a) \in MD(D_a)$) which is imposed when the agent does not fulfill its part of the task $\tau_\delta(a)$;*
- *a history h is a sequence of actions, and $valid(h)$ holds if:*
  - *the $propose$ and $breakit$ actions in the sequence respect $\leq$,*
  - *each $propose$ is followed by a sequence of $accept$ or $reject$ actions respecting $\leq$ until either all agents have accepted the deal or one agent has rejected it,*
  - *there is no double occurrence of a proposal $propose(a, \tau(g))$ of same deal by any agent $a \in G$, and*
  - *the sequence h ends iff either all agents have accepted the last proposal ($finished(h)$) or the last agent has broken the negotiation ($broken(h)$) instead of making a new proposal.*

The following example illustrates the negotiation protocol.

Example 4 *Given an obligation $O_{\{boy_1,boy_2,boy_3\}}(prepared\_table, early\_to\_bed \mid \top)$ with its task $g = \top \rightarrow prepared\_table$, the $NMAS$ protocol is:*
$Ag = G = \{boy_1, boy_2, boy_3\}$
Deals = the sets of pairs $\langle \tau_\delta, \tau_\sigma \rangle$ where the set of decisions τ_δ belongs to the set $\Delta = \{\delta = \{get\_glasses(a_1), get\_drinks(a_2), get\_sheet(a_3), get\_forks(a_4), get\_dishes(a_5)\} \mid a_1, \ldots, a_5 \in Ag \wedge \delta \cap X_a \neq \emptyset$ for each $a \in Ag\}$. And τ_σ is a tuple of $|Ag|$ elements specifying a sanction $s_i \in X$ for every agent in Ag.
Here is a history h, where boy_1 proposes something which is not accepted, but boy_2 thereafter proposes a distribution which is accepted:
$action_1 : propose(boy_1, d_1 = \langle \tau_\delta, \langle s_1, s_2, s_3 \rangle\rangle)$ *where*
$\tau_\delta = \{get\_glasses(boy_1), get\_drinks(boy_2),$
$get\_sheet(boy_3), get\_forks(boy_3), get\_dishes(boy_3)\}$
$action_2 : accept(boy_2, d_1)$
$action_3 : reject(boy_3, d_1)$
$action_4 : propose(boy_2, d_2 = \langle \tau'_\delta, \langle s_1, s_2, s_3 \rangle\rangle)$ *where*
$\tau'_\delta = \{get\_glasses(boy_1), get\_drinks(boy_2),$
$get\_sheet(boy_2), get\_forks(boy_3), get\_dishes(boy_3)\}$
$action_5 : accept(boy_3, d_2)$
$action_6 : accept(boy_1, d_2)$

We have $valid(h)$, because the order of action respects $\leq$, and we have $accepted(h)$, because the history ends with acceptance by all agents ($action_5$ and $action_6$) after a proposal ($action_4$).

Finally we define how agents make decisions in the normative multiagent system together with the interaction protocol. The games we define are as follows. First the agents negotiate the distribution of an obligation, then they make a decision in the normative multi-agent system to either fulfill the obligations or accept the associated sanctions. To define such games, we have to define the effect of the negotiation as an update of the system with the accepted deal together with its sanctions, and we have to define how the normative system sanctions the agent that has broken the negotiations. For the former, an agreement creates a set of obligations for the set of agents. An agreement works like a contract [2]: the agreement creates the obligations for the agents to perform the part of the task they have agreed upon. For the latter, we assume that there is a known penalty (π) for the agent who breaks the negotiations (which may depend on the agent, e.g., older boys may be punished harder for breaking negotiations than the younger ones).

Definition 9 (Effect of negotiation) *Let:*

- *$NMAS$ be the system $\langle A, X, D, AD, E, MD, \geq, N, V\rangle$ and*
- *$NP = \langle Ag, deals, actions, valid, finished, broken\rangle$ be a negotiation protocol for the obligation $O_G(x, s \mid Y)$ based on a total order $\leq \subseteq Ag \times Ag$, and*
- *a break penalty $\pi : Ag \rightarrow Lit(X)$ be a function from agents to literals, such that $\top \rightarrow \sim\pi(a) \in MD(D_a)$ for $a \in Ag$.*

The effect of the negotiation with history h is the normative multiagent system $NMAS'$ defined as follows:

- *If the negotiation is successful ($finished(h)$), with deal and sanctions $\tau(g)$, then $NMAS'$ is $NMAS$ together with $Y \wedge \sim \tau_\delta(a) \rightarrow V(n, \{a\})$ and $V(n, \{a\}) \rightarrow \tau_\sigma(a)$ added to E for each agent $a \in G$.*
- *If the negotiation is broken by agent a, then $\top \rightarrow \pi(a)$ is added to E.*

The set of possible outcomes of the history h, written as $outcomes(h) = \{out(E, \delta) \mid \delta \in \Delta$ is optimal$\}$, is the set of the consequences of the optimal decisions of $NMAS'$.

The games the agents can play in this extended qualitative game theory are as follows. Due to the fact that the number of possible distributions of obligations is finite, the histories and the number of histories are finite too, thus the definition is well founded.

Definition 10 *A history h_1 dominates a history h_2 at step i if they have the same set of actions at step $1 \ldots i-1$, are optimal for step $i+1 \ldots$, and $outcomes(h_1) \succ_a outcomes(h_2)$, agent a performing action i prefers the set of possible outcomes of h_1 to the set of possible outcomes of h_2.*

A history is optimal at step i if it is not dominated by another history, and it is optimal at all steps $j > i$.

A history is optimal if it is optimal at step 1.

The behavior of agents in the negotiation protocol is illustrated in the following example.

Example 5 *Consider two histories h as above and h' which is like h until $action_3$ while it continues in this way:*
$action_4 : propose(boy_3, d_3 = \langle \tau''_\delta, \langle s_1, s_2, s_3\rangle\rangle)$ *where* $\tau''_\delta = \{get\_glasses(boy_1), get\_drinks(boy_2), get\_sheet(boy_2), get\_forks(boy_2), get\_dishes(boy_3)\}$
$action_5 : accept(boy_1, d_3)$
$action_6 : reject(boy_3, d_3)$
$action_7 : breakit(boy_1)$

If we consider history h, the outcomes for agent boy_1 depend on whether it respects the task assigned to it or not and on whether the other agents do their part:
$\{\{get\_glasses(boy_1), get\_drinks(boy_2), ...\},$
$\{early\_to\_bed, \neg get\_glasses(boy_1), \tau_\sigma(boy_1), ...\},$
$\{early\_to\_bed, get\_glasses(boy_1), \neg get\_drinks(boy_2), ...\}, ...\}$

In case of history h', since agent boy_1 breaks the negotiation, the outcome contains a double sanction: $\{\{early\_to\_bed, \pi(boy_1)\}\}$. If doing its part of the task is preferred to the sanction $\pi(boy_1)$, then h dominates h'.

4 Analysis

In this section we use our formal framework to analyze the role of abilities of agents in the distribution of obligations and sanctions, and to classify collective obligations.

4.1 Abilities of agents

In some cases it is a drawback to be the only agent able to see to the fulfilment of part of an obligation, but in other cases it may be an advantage, because of the power it gives to the agent over the other agents during the negotiation.

The argument for the latter has a backward induction character which is characteristic of game theory paradoxes like the centipede [6]. The agents know that they will break the negotiation if they are not able to accept or to reply to the last proposal with another proposal. But an agent who is not able to perform some action cannot propose a deal where it assigns this action to himself. Hence, it knows that it has a smaller number of possible proposals to reply with to its partners when it rejects a proposal. At a certain point of the negotiation, it will run out of counterproposals, so it will be compelled to break the negotiation. Being responsible for a breakdown of the negotiation can lead to be punished or punished more than the partners. As a consequence, the rational strategy for such an agent is to accept a proposal of its partners or to propose a deal which is less favorable to him but which is executable by him, before running out of its alternatives.

In the following table we represent an example where agents a and b negotiate how to distribute actions x, y and z. Agent b, however, is not able to perform action z. Assume the costs of actions are respectively 5, 3 and 1. The first two columns represent the distribution of actions in a deal, the third one the costs for the agents, the fourth one their preferences and the last one the order of proposals.

a	b	cost	preference		bid
x y	z	8 1	6		a4
x z	y	6 3	5	1	b1
y z	x	4 5	3	2	b2
z	x y	1 8	1	3	a1
y	x z	3 6	2		a2
x	y z	5 4	4		a3

Agent b should accept the first proposal ($a1$) of agent a (that a does only z and b does both x and y): it knows that, if it continues, after the proposal $b2$ it cannot accept any proposal of agent a ($a2 - a4$). These proposals all assign to b the action z it cannot execute, and it is compelled to refuse them, since it has no counterproposal to do.

4.2 Characterizing collective obligations

The distinguishing feature of obligation distribution protocols is that deals consist of the distribution of the obligations (τ_δ), as well as the distribution of the associated sanctions (τ_σ). We assume that both distributions are made at the same time. Another option is a two stage negotiation protocol, in which first the obligations are distributed, and in a second round the agents negotiate the sanctions (though this approach seems to have some drawbacks). Yet another issue is the distinction between two interpretations of sanctions, either as cues that the other agents see to their task, or as decommitment possibilities for the agents themselves [15, 16].

Our model distinguishes among three types of sanctions:

- the sanction associated with the obligation, which is imposed when the obligation is violated, regardless which agent is responsible for it;
- the sanctions associated with the negotiated deal, which are imposed if one of the agents does not fulfill its part of the deal; this is at least a sanction for the agent that does not fulfill its part, i.e., its absence is desired by this agent, but it may also be a sanction for more or even all agents involved;
- the sanctions associated with the break penalty π, which are at least a sanction for the agent that breaks the negotiation, but which also may be a sanction for more or all of the agents involved, e.g., we may have not only $\top \rightarrow \sim\pi(a) \in MD(D_a)$, but the normative system can also see to it that $\top \rightarrow \sim\pi(a) \in MD(D_b)$ for all agents $b \in Ag$.

We characterize collective obligations according to the way the responsibility in case of violation is attributed to individual agents or the whole set of agents, that is the balance between the first two types of sanction, and the responsibility is attributed in case of broken negotiation, that is the third type of sanction. For example, consider the case of the United Nations which obliges two conflicting nations for make peace. The UNO may consider liable for the violation of the obligation the nation which causes a failure of negotiations rather than both the parties.

5 Summary and closing remarks

In this paper we consider the distribution of obligations directed to collectives. We claim that the distribution of obligations is the result of a negotiation process among the agents: they have to find an agreement about how to execute a set of actions which fulfill the collective obligation. We distinguish different types of collective obligations depending on the responsibilities assigned during the negotiation or execution phase.

The normative multiagent system we propose can be extended with obligations defined in terms of desires and goals of the normative systems, permissions as exceptions, and norms for policies in virtual communities. In this paper, violations are consequences of the lack of the fulfilment of an obligation and the sanctions are consequences of violations. In reality, both considering something as a violation and sanctioning an agent are autonomous actions performed by the normative system which can be thought as having the goal of sanctioning violations. So, the normative system can be considered as an agent, which the agents subject to obligations play games with [4]. Further work is the development of the process of creating obligations as result of institutional actions like contracts, as in [2].

REFERENCES

[1] A. Anderson, ‘A reduction of deontic logic to alethic modal logic’, *Mind*, **67**, 100–103, (1958).

[2] G. Boella and L. van der Torre, ‘Contracts as legal institutions in organizations of autonomous agents’, in *Procs. of AAMAS’04*, New York, (2004).

[3] G. Boella and L. van der Torre, ‘Δ: The social delegation cycle’, in *LNAI n.3065: Procs. of ΔEON’04*, pp. 29–42, Madeira (PG), (2004).

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

[5] 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).

[6] J. Broome and W. Rabinowicz, ‘Backwards induction in the centipede game’, *Analysis*, **59(4)**, 237–242, (1999).

[7] C. Castelfranchi, ‘The micro-macro constitution of power’, *Protosociology*, **18**, 208–269, (2003).

[8] L. Cholvy and C. Garion, ‘Collective obligations, commitments and individual obligations: a preliminary study’, in *Procs. of ΔEON’02*, London, (2002).

[9] F. Dignum and L. Royakkers, ‘Collective obligation and commitment’, in *Procs. of 5th Int. Conference on Law in the Information Society*, Florence, (1998).

[10] J. Ferber, O. Gutknecht, and F. Michel, ‘From agents to organizations: an organizational view of multiagent systems’, in *LNCS n. 2935: Procs. of AOSE’03*, pp. 214–230. Springer Verlag, (2003).

[11] D. Makinson and L. van der Torre, ‘Input-output logics’, *Journal of Philosophical Logic*, **29**, 383–408, (2000).

[12] J. J. Ch. Meyer, ‘A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic’, *Notre Dame Journal of Formal Logic*, **29(1)**, 109–136, (1988).

[13] 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).

[14] J. S. Rosenschein and G. Zlotkin, *Rules of Encounter. Designing Conventions for Automated Negotiation among Computers*, MIT Press, Cambridge, MA, 1994.

[15] T. Sandholm, S. Sikka, and S. Norden, ‘Algorithms for optimizing leveled commitment contracts’, in *Procs. of IJCAI’99*, Stockholm, (1999).

[16] V. Teague and L. Sonenberg, ‘Investigating commitment flexibility in multiagent contracts’, in *Game Theory and Decision Theory in Agent-Based Systems*, eds., S. Parsons, P. Gymtrasiewicz, and M. Wooldridge, 267–292, Kluwer, (2002).

[17] M. Wooldridge and S. Parsons, ‘Languages for negotiation’, in *Procs. of ECAI’00*, pp. 393–397, Berlin, (1998).