Division of Powers in MAS Control

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

Leendert van der Torre
SEN-3
CWI Amsterdam
The Netherlands
E-mail: torre@cwi.nl

ABSTRACT

Decision making in multiagent systems has to deal with the norms regulating the system. In this paper we propose a logical framework based on three dimensions. First, we distinguish between agents whose behavior is governed by norms, defenders of these norms and autonomous normative systems; in this paper we call the latter two normative agents. Second, we distinguish some of the usual mental attitudes for all agents, including the normative agents. Third, we distinguish between behavior that counts as a violation, and sanctions that are applied. To formalize decision making we also extend this framework to a qualitative game theory. n-player games are based on recursive modelling: the bearer of obligations models the defender agents who have the duty to monitor violations and to apply sanctions, which in turn model the normative systems, which issue the norms and watch over the behavior of these defender agents. We show how normative systems can delegate monitoring and sanctioning of violations to autonomous defender agents, inspired by Montesquieu's trias politica.

1. INTRODUCTION

Recent approaches to distributed systems such as virtual communities of autonomous agents [14] raise the issue of the distribution of control in such MAS. In particular the management of such systems cannot be centralized in a single agent since this would risk to endanger the core business of the system [12]. Thus it seems useful to apply new metaphors taken from the regulations of human societies for dealing with this problem. One of the key concepts of the organization of modern societies is the separation of powers as proposed in the Montesquieu's *trias politica*: the representative, executive and judicial authorities should be kept distinct. Moreover decentralizing the control of the policies regulating a society supports the view that tasks can be better performed if they are dealt with by the local level in an autonomous way.

In this paper we model norms in qualitative decision theories, based on belief and desire rules. We study norms expressed in a logical framework with the following three dimensions:

1. A distinction between agents whose behavior is governed by norms and autonomous so-called normative agents that represent the normative system [2]. We distinguish a particular normative agent called the defender agent, which has the role to autonomously enforce that norms are respected.
2. A distinction between some of the usual mental attitudes for all agents, including the normative agents: their behavior is directed by beliefs, desires and goals.
3. A distinction, when an agent does not respect an obligation, between the fact that its behavior counts as a violation, and the fact that it can be sanctioned by the normative agent [2].

In this paper, we address the following two questions.

- How do the agents make decisions? In particular, do they fulfill or violate the norms?
- How can the role of monitoring and sanctioning violations be delegated by the normative agent to a *defender* agent?

This paper is organized as follows. In Section 2 we introduce the logical framework and we discuss its three dimensions. In Section 3 we present the qualitative game theory, the obligations and some examples.

2. THE THREE DIMENSIONS

In this section we discuss the three dimensions of our logical framework: normative agents, mental attitudes, and the violation / sanction distinction.

2.1 Normative agents

The first dimension of our framework is the set of agents that are distinguished. Normative systems are "sets of agents (human or artificial) whose interactions can fruitfully be regarded as norm-governed; the norms prescribe how the agents ideally should and should not behave [...]. Importantly, the norms allow for the possibility that actual behavior may at times deviate from the ideal, i.e. that violations of obligations, or of agents' rights, may occur" [17]. Boella and Lesmo [2] distinguish between agents whose behavior is governed by norms, and an autonomous normative system. They call the latter the normative agent.

In their approach, normative systems are modelled as a single agent. However, in modern states the power is separated between several autonomous roles: the role of the government, the judicial system and the legislative systems. Moreover, in the perspective of distributed multiagent systems, such a distinction between roles can make the social control more effective. In our model, we introduce a first preliminary distinction between agents who have only the judiciary power (defenders in Conte and Castelfranchi [10]'s terminology) and those who have also the legislative one, i.e.

the normative agents. The task of the normative agent is kept separate from the one delegated to the defender agents. So defenders can act autonomously on the basis of more local information.

As an example of such a scenario in a virtual community, consider agent a_1 who is subject to the obligation to share its file system space on the web. This obligation derives from the policy regulating the virtual community it belongs to. Since the central authority a_3 has not enough resources to control and punish every member of the community it delegates this control task to agent a_2. However, it obliges agent a_2 to perform correctly its task by punishing it in case of non compliance with its duties. Since the virtual community is composed of heterogeneous agents, a_3 cannot assume that a_2 is a respectful agent who fulfil every obligation is imposed on it. Hence, a_3 tries to control a_2's behavior by means of obligations concerning its task to monitor and punish a_1. Multiple levels of delegation of controls are possible: a_3 could delegate to another defender the task of controlling the defender a_2, so to have a hierarchy of agents.

2.2 The mental attitudes of agents

The second dimension is the set of mental attitudes assigned to the agents. In agent theory, concepts like beliefs, desires and intentions are proper abstraction tools for characterizing the behavior of agent systems [21]. Agents' behavior is governed by their specific balance between beliefs, desires and intentions. Moreover, norms and obligations seem to be a further ingredient in the control of agents' behavior, but there is much less consensus on how norms and obligations are integrated with the beliefs, desires and intentions.

Boella and Lesmo also attribute mental states to the normative agent (thus taking an intentional stance [11] towards normative systems). Roughly, the content of an obligation is a goal of the normative agent. In other words, in order to blend norms with the BDI (Beliefs, Desires and Intentions) agents model, norms and obligations are introduced basing them on beliefs and pre-existing motivational attitudes, without resorting to another primitive attitude. But once obligations are introduced in the decision process on a BDI agent, the agent must be allowed to decide to violate them. Otherwise, an agent cannot be said to be truly autonomous. Such agents base their decision process on a symbolic representation of their preferences, so it is more immediate to adopt a qualitative decision process, such as the one proposed by Broersen *et al.* [7]. We [3] explained this attribution of mental attitudes for normative *multiagent* systems as dynamic social orders, patterns of interactions among agents "such that it allows the satisfaction of the interests of some agent A" [8]. These interests can be a shared goal, a value that is good for everybody or for most of the members; for example, the interest may be to avoid accidents. A social order requires *social control*, "an incessant local (micro) activity of its units" [8], aimed at restoring the regularities prescribed by norms. Thus, the agents attribute to the normative system, besides goals, also the ability to autonomously enforce the conformity of the agents to the norms, because a dynamic social order requires a continuous activity for ensuring that the normative system's goals are achieved. The importance of punishment for the success of societies in evolutionary competition has been argued by Boyd *et al.* [6].

Thus a normative multiagent system has all the properties requested by Wooldridge and Jennings [21] for being an agent.

In this paper, we also attribute mental attributes to the defender agents. The relation between the two normative agents is that the normative system imposes obligations for the defenders, and it thus motivates the defenders to act in a certain way.

2.3 Violations and sanctions

The third dimension of our framework are the aspects of norms and obligations that are distinguished. The possibility that norms are violated is important in a multiagent system. For example, as Castelfranchi *et al.* [9] claim, because there are conflicting norms by different authorities and unforeseen cases where the respect of a norm leads to a worse result for the system. Boella and Lesmo therefore introduce a definition that does not presuppose that an agent always sticks to the norms it is subject to. Agents have to decide whether to respect a norm or not, thus facing the possibility of a sanction. To avoid sanctions they can counter agent a_2's action by misleading its beliefs or making its sanctions impossible to be applied (see [2]).

For what concerns the possibility of not respecting norms, we distinguish in [3] between behavior that *counts as* a violation - in the sense of the construction of social reality proposed by Searle [19] - and sanctions which are applied; "counts as a violation" can be roughly read as "punishable". Since both the recognition of something as a violation and the sanctioning actions are the result of the activity of the normative agent, as argued by the sociologist Goffman [16], obligations are inherently related with a game-theoretic setting.

In our framework, an agent a_1 is obliged by a norm n issued by agent a_2 to do x if:

- Agent a_2 wants that a_1 does x.
- Agent a_2 desires that there is no violation, but if $\neg x$ then it has the goal that it counts as a violation of the norm n.
- Agent a_2 desires not to sanction, but if there is a violation then it wants to sanction agent a_1. This goal of the normative system expresses that it only sanctions in case of violation.
- Agent a_1 does not desire to be sanctioned.

In this section we introduce defender agents in the definition of obligations with delegation of control. A defender agent a_2 is obliged by agent a_3 to decide that $\neg x$ counts as a violation by agent a_1. Moreover, if a_2 takes this decision it is also obliged to sanction this violation.

These obligations cannot prevent that an uncooperative defender agent decides not to perform its duties leaving unsanctioned the behavior of the bearer of an obligation. Moreover, delegated control paves the way to other possibilities of violating obligations without being sanctioned: the agent can try to mislead or to influence the behavior of the defender agent.

3. RECURSIVE MODELLING

In this section we present a logical framework for BDI agents based on recursive modelling [15]. This framework is extended to a qualitative game-theory for dealing with n-player games: each player considers the reaction of the subsequent agent in the hierarchy. We assume that the reaction of the subsequent agent affects only the outcome of the immediately preceding agent. Hence, each agent's behavior is watched by another agent whose behavior can be in control of another one and so on in a recursive way; until the highest level of authority whose behavior is not controlled is reached.

The basic picture is visualized with three agents in Figure 1 and reflects the deliberation of agent a_1 in various stages. a_1 is subject to an obligation by a_3 and a_2 is a defender agent (underscript numbers denote the agent, while superscript numbers the time instant).

Agent a_1 is obliged to make certain decisions, and it is deliberating about the effects of the fulfilment or the violation of these obligations. Agent a_2 is the defender which may recognize and sanction violations. 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, a_1 has to consider that a_2 is subject to some obligations (e.g., to make a_1 respect its obligations): so in modelling a_2, it considers that a_2 recursively models a_3, the normative agent who watches over agent a_2's behavior.

When agent a_1 makes its decision d_1, it believes that it is in state s_1^0. The expected consequences of this decision (due to belief rules B_1^1) are called state s_1^1. Then agent a_2 makes a decision d_2, typically whether it counts this decision as a violation and whether it sanctions agent a_1 or not. 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. Agent a_1 expects that agent a_2 believes that the expected result of decision d_2 is state s_2^2. Finally, agent a_1's expected consequences of d_2 from a_1's point of view are called state s_1^2. And a_2 makes a similar reasoning about a_3's decisions. Note however, that the recursion in modelling other agents stops here since agent a_3 has no authority watching over its behavior. Hence it has not to base its decisions on the expected reaction of another agent.

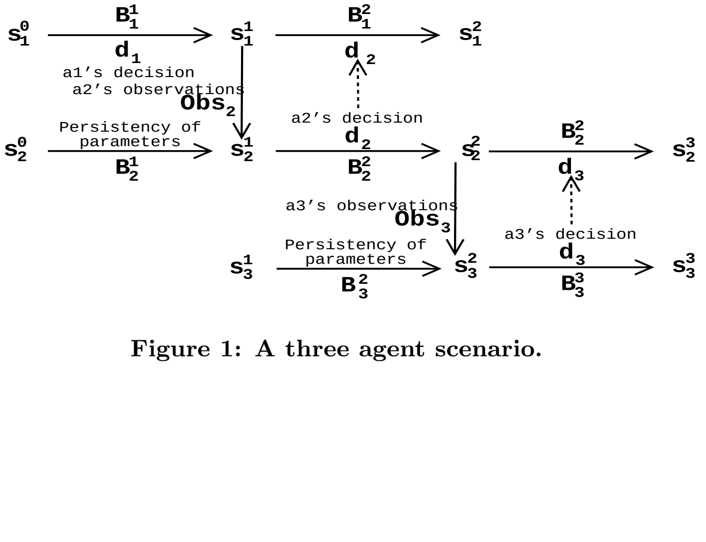

Figure 1: A three agent scenario.

3.1 Agent theory

In this section we define when states respect belief rules and violate or fulfill desire and goal rules. We start with the *decisions*. We assume that the base language contains boolean variables and logical connectives. The variables 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. The distinction between decision variables and parameters is a fundamental principle in all decision theories or decision logics. Our terminology is borrowed from Lang et al. [18], they are called respectively controllable and uncontrollable propositions by Boutilier [5].

Definition 1 (Decisions). *Let $A = \{a_1, a_2, \ldots, a_n\}$ be a set of n distinct agents. $A_i = \{m, m', m'', \ldots\}$ for $a_i \in A$ and $P = \{p, p', p'', \ldots\}$ be $n+1$ disjoint sets of propositional variables. By convention $A_0 = \emptyset$. 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 $1 \le i \le n$. 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 $\neg x \in d_i$. By convention $d_{n+1} = \emptyset$.

We distinguish between what we call the agent's epistemic states, i.e. its beliefs about the world, and its mental states, i.e. the sets of its belief, desire and goal rules. We first formalize the epistemic states: since an agent has a profile of the next agent's beliefs too, its epistemic state σ_i will include also the epistemic state of the next agent a_{i+1} (σ_{i+1}) unless it is the last agent a_n who does not recursively model the behavior of further agents. In order to distinguish the value of the propositional variables in the sequence of stages, we use superscript numbers to label the parameters and states according to the stage they describe.

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

The epistemic state of agent a_n is a tuple $\sigma_n = \langle s_n^{n-2}, s_n^{n-1}, s_n^n, s_n^{n+1}\rangle$, while for agents a_i $1 \le i < n$ it includes also the epistemic state of the next agent a_{i+1} $\sigma_i = \langle s_i^{i-2}, s_i^{i-1}, s_i^i, s_i^{i+1}, \sigma_{i+1}\rangle$. s_i^{i-2} is a set of literals of $L_{P^{i-2}}$ (the state before agent a_{i-1}'s action. $s_i^{i-1} \subseteq L_{A_{i-1}P^{i-1}}$ (the initial state of agent a_i's action), $s_i^i \subseteq L_{A_iP^i}$ (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.

States s_1^{-1} and s_n^{n+1} are dummies considered only for the sake of generality of the recursive definition.

The agent's mental state contains five sets of rules for each agent. Three sets of *belief rules* are used to calculate the expected consequences of decisions and *desire* and *goal rules* express the attitudes of the agents towards a given state, depending on the context.

DEFINITION 3 (MENTAL STATES). *Let a rule of one of the propositional languages* L_1, $L_{A_1P^i}$, … *be an ordered sequence of literals* $l_1, \ldots, l_r, l$ *of this language written as* $l_1 \wedge \ldots \wedge l_r \rightarrow l$ *where* $r \geq 0$.

The mental state M_n *of agent* a_n *is a tuple* $\langle B_n^{n-1}, B_n^n, B_n^{n+1}, D_n, G_n \rangle$, *while the mental state* M_i *of agent* a_i $1 \leq i < n$ *is a tuple* $\langle B_i^{i-1}, B_i^i, B_i^{i+1}, D_i, G_i, M_{i+1} \rangle$, *where* M_{i+1} *is the mental state that agent* a_i *attributes to agent* a_{i+1}. B_i^{i-1} *is a set of rules of* $L_{A_{i-1}P^{i-2}P^{i-1}}$, B_i^i *is a set of rules of* $L_{A_{i-1}A_iP^{i-2}P^{i-1}P^i}$, B_i^{i+1} *is a set of rules of* $L_{A_{i-1}A_iA_{i+1}P^{i-2}P^{i-1}P^iP^{i+1}}$, D_i, G_i *are sets of rules of* $L_{A_{i-1}A_iA_{i+1}P^{i-2}P^{i-1}P^iP^{i+1}}$. *Let* $B_i = B_i^{i-1} \cup B_i^i \cup B_i^{i+1}$; *let* $B_1^0 = B_n^{n+1} = \emptyset$ *for the sake of generality.*

The normative agent's beliefs 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 4 (OBSERVATIONS). *The propositions observable by agent* a_i, OP_i, *are a subset of the description 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 observation of agent* a_i *in state* s_i^{i-1} *is* $Obs_i = \{p \mid p \in OP_i \text{ and } p \in s_{i-1}^{i-1}\} \cup \{\neg p \mid p \in OP_i \text{ and } \neg p \in s_{i-1}^{i-1}\}$: *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 decision sets and epistemic states are related to each other by the agent's mental state. There are two different kinds of relations. First, the belief rules express whether the states are the expected consequences of the decisions. Second, the desire and goal rules are used to evaluate the consequences of decisions.

How the agents reason about obligations, and in particular how they deliberate whether they fulfill or violate them, depends not only on their interpretation of the obligations in terms of their beliefs, desires and goals, but also on their *agent characteristics*. Given the same set of rules, distinct agents reason and act differently. For what concerns reasoning, different agents can deal with conflicts among belief rules in different ways. For what concerns acting, a respectful agent always tries to fulfill the goals of the normative system, whereas a selfish agent first tries to achieve its own goals. We express these agent characteristics by a priority relation on the rules which encode, as detailed in Broersen *et al.* [7], how the agent resolves its conflicts.

DEFINITION 5 (AGENT CHARACTERISTICS). *The characteristics of the agent* a_i *are a tuple* $C_i = \langle \geq_i^B, \geq_i \rangle$ *of transitive and reflexive relations containing at least the subset relation, defined, respectively,* $\geq_i^B$ *on the powerset of* B_i *and* $\geq_i$ *on the powerset of* $D_i \cup G_i$.

DEFINITION 6 (RESPECTING MENTAL STATES AND BELIEFS). *For* s *a state,* f *a set of literals in* L, R *a set of rules, and* $\geq$ *a transitive and reflexive relation on the powerset of* R *containing at least the superset relation, let* $\max(s, f, R, \geq)$ *be the set of states obtained by:*

1. *S is the set of states after applying a consistent subset of R to the union of the state s with f:*

   $$\begin{aligned} S = \{&\{l_1, \ldots, l_n\} \cup f \mid l_{i,1} \wedge \ldots \wedge l_{i,m_i} \rightarrow l_i \in R \text{ and} \\ &l_{i,j} \in s \cup f \text{ for } j = 1 \ldots m_i \text{ for } i = 1 \ldots n \text{ and} \\ &\{l_1, \ldots, l_n\} \cup f \text{ consistent } \} \end{aligned}$$

2. *S' is the set of maximal elements of S, i.e.*

   $$S' = \{s \in S \mid \nexists s' \in S \text{ such that } s \subset s'\}$$

3. *S'' is the set of maximal (with respect to the $\geq$ ordering) elements of S', i.e.*

   $$\{s \in S' \mid \nexists s' \in S' s' \geq s \text{ and } s \not\geq s'\}$$

4. *$\max(s, f, R, \geq)$ is the set of states that contain an element of S'' together with some elements from s:*

   $$\begin{aligned} \max(s, f, R, \geq) \quad &= \{s' \cup s'' \mid s' \in S'' \text{ and} \\ &s'' = \{l^i \in s \mid l^i \in P^i \text{ and } \overline{l^{i+1}} \notin s'\}\} \end{aligned}$$

A state description $\sigma_i = \langle s_i^{i-2}, s_i^{i-1}, s_i^i, s_i^{i+1}, \sigma_{i+1} \rangle$ *respects the decision set* $\delta = \langle d_1, \ldots, d_n \rangle$, *the expected observations* Obs_i *of agent* a_i *together with the mental state description* $M_i = \langle B_i^{i-1}, B_i^i, B_i^{i+1}, D_i, G_i, M_{i+1} \rangle$*if*
$s_i^{i-1} \in \max(s_i^{i-2}, Obs_i, B_i^{i-1}, \geq_i^B)$,
$s_i^i \in \max(s_i^{i-2} \cup s_i^{i-1}, d_i, B_i^i, \geq_i^B)$,
$s_i^{i+1} \in \max(s_i^{i-2} \cup s_i^{i-1} \cup s_i^i, d_{i+1}, B_i^{i+1}, \geq_i^B)$, *and,*
if $i < n$, σ_{i+1} *respects the decision set* $\delta = \langle d_1, \ldots, d_n \rangle$, *the expected observations* Obs_{i+1} *of agent* $i + 1$ *together with the mental state description* M_{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 $B_1^0 = B_n^{n+1} = \emptyset$.

The agent characteristics applied to the belief rules enable us to model the non-monotonic aspects of action effects. In fact, we model applicability conditions for actions and conditional effects. A decision variable has an effect unless some contextual condition holds.

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.

DEFINITION 7 (UNFULFILLED MENTAL STATES). *Let* $U(R, s)$ *be the unfulfilled rules of state* s, $U(R, s) = \{l_1 \wedge \ldots \wedge l_n \rightarrow l \in R \mid \{l_1, \ldots, l_n\} \subseteq s \text{ and } l \notin s\}$. *The* unfulfilled mental state description *of agent* a_i *is the tuple* $U_i = \langle U_i^D = U(D_i, s_i), U_i^G = U(G_i, s_i) \rangle$.

For what concerns the priorities on desire and goal rules, agents can be classified according to the way they solve the conflicts among the rules belonging to different components: desires, goals and desires and goals of the normative system that can be adopted. We defined agent types as they have been introduced in the BOID architecture [7]. Here for space reasons, we introduce only a selfish stable agent type, which bases its decisions only on its unsatisfied goals and desires.

DEFINITION 8 (AGENT TYPES).

$s_i \leq s'_i$ *iff* $U'^G_i = U(G_i, s'_i) \geq_i U_i^G = U(G_i, s_i)$ *and if* $U'^G_i \geq_i U_i^G$ *and* $U_i^G \geq_i U'^G_i$ *then* ${U'^D}_i \geq_i {U^D}_i$

We finally can define the optimal decisions.

Definition 9 (Optimal decisions). *For $1 \leq i \leq n$, given initial state s_i^{i-2}, a mental state M_i, observations OP_i and agent characteristics C_i, the decision set $\delta = \langle d_1, \ldots, d_n \rangle$ is optimal for agent i if for all state descriptions $\sigma_i = \langle s_i^{i-2}, s_i^{i-1}, s_i^i, s_i^{i+1}, \sigma_{i+1} \rangle$ which respect δ, for all other decision sets $\delta' = \langle d'_1, \ldots, d'_n \rangle$, where $d_j = d'_j$ if $1 \leq j < i$, which are optimal for agents j, $i < j \leq n$, for all state descriptions σ'_i respecting δ', $s_i \leq s'_i$ (where $s_i = s_i^{i-2} \cup s_i^{i-1} \cup s_i^i \cup s_i^{i+1}$).*

3.2 Obligations

Our approach is inspired to the so-called Anderson's reduction of modal logic [1], which may be written as $O(p) = \Box(\neg p \rightarrow V)$.

Definition 10 (Obligations). *Let NS be a normative system or set of norms $\{n, n', n'', \ldots\}$ and let the decision variables A_{i+1} of agent a_{i+1} have a non empty intersection with a set of so-called violation variables $V = \{V_i(n) \mid n \in NS$ and $a_i \in A_i\}$, which represent the fact that something counts as a violation of norm n by agent a_i.*

Agent a_i believes that it is obliged to decide to do x, a literal built from a parameter in $P^i \cup P^{i+1}$ or from a decision variable in A_i, $O_{i,i+1}(x)$, iff agent a_i believes that there is a norm $n \in NS$ such that:

1. *$\top \rightarrow x \in D_{i+1} \cap G_{i+1}$: agent a_i believes that agent a_{i+1} desires and has as a goal that x and wants agent a_i to adopt x as a goal.*

2. *$\neg x \rightarrow V_i(n) \in D_{i+1} \cap G_{i+1}$: agent a_i believes that if (agent a_{i+1} believes that) $\neg x$ then agent a_{i+1} has the goal and the desire to recognize it as $V_i(n)$: a violation of norm n by agent a_i.*

3. *$\top \rightarrow \neg V_i(n) \in D_{i+1}$: agent a_i believes that agent a_{i+1} desires that there is no violation.*

When the literal x is built from a decision variable, then we call the obligation an ought-to-do obligation, and when it is built from a parameter then we call it an ought-to-be obligation.

We now extend this definition to conditional sanction-based obligations

Definition 11 (Conditional obligations with sanction). *Agent a_i believes that it is obliged to decide to do x (a literal built from a propositional variable in $P^i \cup P^{i+1} \cup A_i$) with sanction s (a decision variable in A_{i+1}) under condition q (a proposition of $L_{A_i P^i P^{i+1}}$), $O_{i,i+1}(x, s|q)$, iff for some $n \in NS$:*

1. *$q \rightarrow x \in D_{i+1} \cap G_{i+1}$: agent a_i believes that (if agent a_{i+1} believes to be) in context q agent a_{i+1} desires and has as a goal that x and wants agent a_i to adopt x as a goal.*

2. *$q \wedge \neg x \rightarrow V_i(n) \in D_{i+1} \cap G_{i+1}$: agent a_i believes that if (agent a_{i+1} believes that) $q \wedge \neg x$ then agent a_{i+1} has the goal and the desire $V_i(n)$: to recognize it as a violation of agent a_i.*

3. *$\top \rightarrow \neg V_i(n) \in D_{i+1}$*

4. *$V_i(n) \rightarrow s \in D_{i+1} \cap G_{i+1}$: agent a_i believes that if agent a_{i+1} decides $V_i(n)$ then it desires and has as a goal that it sanctions agent a_i.*

5. *$\top \rightarrow \neg s \in D_{i+1}$: agent a_i believes that agent a_{i+1} desires not to sanction.*

6. *$\top \rightarrow \neg s \in D_i$: agent a_i has the desire not to be sanctioned.*

Finally, we introduce the distinction between a defender agent a_{i+1} who has the duty to enforce a norm n and a normative agent a_{i+2} who imposes by means of norms n' and n'' to a_{i+1} the duty to watch over a norm n.

Definition 12 (Delegated obligations). *Agent a_i believes that it is obliged to decide to do x (a literal built from a propositional variable in $P^i \cup P^{i+1} \cup A_1$) with sanction s, a decision variable of A_{i+1} performed by defender a_{i+1}, on behalf of the normative agent a_{i+2} (where $V_i(n) \in V \cap A_{i+1}$), $O_{i,i+1,i+2}(x, s)$, iff for some $n, n', n'' \in NS$:*

1. *$\top \rightarrow x \in D_{i+2} \cap G_{i+2}$: agent a_i believes that agent a_{i+2} desires and has as a goal that x and wants agent a_i to adopt x as a goal.*

2. *$O_{i+1,i+2}(V_i(n), s'|\neg x)$: agent a_i believes that if (agent a_{i+2} believes that) $\neg x$ then agent a_{i+1} is conditionally obliged by agent a_{i+2} to determine that this counts as a violation $V_i(n)$ by agent a_i.*

3. *$O_{i+1,i+2}(s, s'|V_i(n))$: agent a_i believes that if (agent a_{i+2} believes that) agent a_{i+1} decides that $\neg x$ counts as a violation $V_i(n)$ then it is conditionally obliged by agent a_{i+2} to sanction agent a_i.*

4. *$\top \rightarrow \neg V_i(n) \in D_{i+2}$*

5. *$\top \rightarrow \neg s \in D_{i+2}$*

6. *$\top \rightarrow \neg s \in D_i$*

Obligations at Items 2 and 3 imply by their definitions:

1. $\neg x \rightarrow V_i(n) \in D_{i+2} \cap G_{i+2}$: If $\neg x$ then agent a_{i+2} has the goal and the desire that agent a_{i+1} does $V_i(n)$: it recognizes $\neg x$ as a violation by agent a_i.

2. $V_i(n) \rightarrow s \in D_{i+2} \cap G_{i+2}$: if $V_i(n)$ then agent a_{i+2} desires and has as a goal that agent a_{i+1} sanctions agent a_i.

Given that these two goals are normative goals for agent a_{i+1}, if it adopts them, then a_i is in the same situation as in the definitions above.

Further definitions with multiple levels of defenders are possible since obligations at Items 2 and 3 can be delegated to a second defender agent and so on.

3.3 Examples

The first example represents in a two agent scenario an ought to be obligation - Definition 10 - to achieve p^1 of a stable agent a_1 which adopts p^1 only for the fear of the sanction s even if it desires not to do anything for achieving p^1.

Example 1. $O_{1,2}(p^1, s)$
$s_1^0 = \emptyset, B_1 = \{x \rightarrow p^1\}, \geq_1^B = \emptyset, x \in A_1, p^1 \in P^1,$
$G_1 = \emptyset, D_1 = \{\top \rightarrow \neg x, \top \rightarrow \neg s\},$
$\geq_1 = \{\top \rightarrow \neg s\} \geq \{\top \rightarrow \neg x\}$
$s_2^0 = \emptyset, Obs_2 = A_1 \cup P^1, B_2 = \{x \rightarrow p^1\}, \geq_2^B = \emptyset,$

$V_1(n) \in V \cap A_2, s \in A_2, n \in NS,$
G_2=$\{\top \rightarrow p^1, \neg p^1 \rightarrow V_1(n), V_1(n) \rightarrow s\},$
$D_2 = \{\top{\rightarrow}p^1, \neg p^1{\rightarrow}V_1(n), V_1(n){\rightarrow}s, \top{\rightarrow}\neg V_1(n), \top{\rightarrow}\neg s\},$
$\geq_2 \supseteq \{\neg p^1 \rightarrow V_1(n)\} > \{\top \rightarrow \neg V_1(n), \top \rightarrow \neg s\}$
Optimal decision set: $\langle d_1 = \{x\}, d_2 = \emptyset\rangle$
Expected state description: $s_1^1 = \{x, p^1\}, s_2^1 = \{x, p^1\}, s_2^2 = \{p^2\}, s_1^2 = \{p^2\}$
Unfulfilled mental states: $U_1^D = \{\top \rightarrow \neg x\}, U_1^G = \emptyset, U_2^D = U_2^G = \emptyset$

Since agent a_1 decides to do x, then $s_1^1 = \max(s_1^0, d_1, B_1^1, \geq_1^B) = \{x, p^1\}$ by Definition 6 of respecting mental states. Agent a_1's desire not to be sanctioned is fulfilled: the antecedent $\top$ of the unconditional rule $\top \rightarrow \neg s$ is true, and the consequent is consistent with state $s_1^2 = \{p^2\}$ since agent a_2 decides not to sanction ($\neg s$) (recall that $s \in A_2$, so it is implicitly a variable of the last stage - Definition 2 - while p^2 by persistency of the parameter p^1 from s_2^1 - Definition 6). In contrast, the unconditional (and hence applicable) goal $\top \rightarrow \neg x$ is in conflict with state $s_1^1 = \{x, p^1\}$ ($x \in A_1$, so it is a decision variable describing second stage) and it remains unsatisfied (see Definition 7).

For what concerns agent a_2's attitudes, its unconditional desire and goal that agent a_1 adopts the content of the obligation $\top \rightarrow p^1$ is satisfied in s_2^1. Analogously are the desires not to prosecute and sanction indiscriminately: $\top \rightarrow \neg V_1(n)$ and $\top \rightarrow \neg s$ (recall that states are complete - Definition 2 - so $\neg V_1(n)$ and $\neg s$ are true in $s_2^2 = \{p^2\}$). The remaining conditional attitudes $\neg x \rightarrow V_1(n)$, etc. are not applicable and hence they are not unsatisfied.

Whatever other decision agent a_2 would have taken, it could not satisfy more goals or desires, so $d_2 = \emptyset$ is a minimal and optimal decision - Definition 9. E.g. $d_2'' = \{s\}$ leaves $\top \rightarrow \neg s$ unsatisfied: $\{\top \rightarrow \neg s\} \geq_2 \emptyset$ (in fact, $\geq_2$ contains the subset relation) and then $U''^D_2 = \{\top \rightarrow \neg s\} \geq U_2^D = \emptyset$ for a stable agent.

Had agent a_1's decision been $d_1' = \emptyset$, agent a_2 would have chosen $d_2' = \{V_1(n), s\}$. The unfulfilled desires and goals in state $s_1' = s_2' = \{V_1(n), s\}$ would have been: $U'^D_1 = \{\top \rightarrow \neg s\}, U'^G_1 = \emptyset$, $=U'^D_2$=$\{\top \rightarrow p^1, \top \rightarrow \neg V_1(n), \top \rightarrow \neg s\}, U'^G_2$=$\{\top \rightarrow p^1\}$

How does agent a_1 take a decision between d_1 and d_1'? Since its agent type is *stable* (Definition 8) it compares which of its goals and desires remain unsatisfied: $U_1^G = U'^G_1 = \emptyset$ but $U'^D_1 = \{\top \rightarrow \neg s\} \geq U_1^D = \{\top \rightarrow \neg x\}$. And hence, the optimal state (Definition 9) is s_1: $s_1 = \{x, p^1, p^2\} \leq s_1' = \{V_1(n), s\}$.

In the second example we show a three agent situation where a_2 is the defender of the obligation to do x on behalf of the normative agent a_3 (Definition 12). However, a_1 prefers to violate the obligation with respect to not being sanctioned.

EXAMPLE 2. *$O_{1,2,3}(x, s)$ and thus $O_{2,3}(V_1(n), s'|\neg x)$ and $O_{2,3}(s, s''|V_1(n))$*
$s_1^0 = \emptyset, B_1 = \emptyset, \geq_1^B = \emptyset, x \in A_1,$
$G_1 = \emptyset, D_1 = \{\top \rightarrow \neg x, \top \rightarrow \neg s\},$
$\geq_1 = \{\top \rightarrow \neg x\} \geq \{\top \rightarrow \neg s\}$
$s_2^0 = \emptyset, Obs_2 = A_1 \cup P^1, B_2 = \emptyset,\ \geq_2^B = \emptyset,$
$V_1(n) \in V \cap A_2, s \in A_2,$
$G_2 = \emptyset, D_2 = \{\top \rightarrow \neg s', \top \rightarrow \neg s''\},$
$s_3^0 = \emptyset, Obs_3 = A_2 \cup P^2, B_3 = \emptyset,\ \geq_3^B = \emptyset,$
$V_2(n'), V_2(n'') \in V \cap A_3, s', s'' \in A_3, n, n', n'' \in NS,$
G_3=$\{\top{\rightarrow}x, \neg x{\rightarrow}V_1(n), V_1(n){\rightarrow}s, \neg x \wedge \neg V_1(n){\rightarrow}V_2(n'),$
$V_2(n'){\rightarrow}s', V_1(n) \wedge \neg s{\rightarrow}V_2(n''), V_2(n''){\rightarrow}s''\},$
$D_3 = \{\top \rightarrow x, \neg x \rightarrow V_1(n), V_1(n) \rightarrow s, \neg x \wedge \neg V_1(n) \rightarrow V_2(n'), V_2(n') \rightarrow s', V_1(n) \wedge \neg s \rightarrow V_2(n''), V_2(n'') \rightarrow s'', \top \rightarrow \neg V_1(n), \top \rightarrow \neg s, \top \rightarrow \neg V_2(n'), \top \rightarrow \neg s', \top \rightarrow \neg V_2(n''), \top \rightarrow \neg s''\},,$
$\geq_3 \supseteq \{\neg x \wedge \neg V_1(n) \rightarrow V_2(n'), V_1(n) \wedge \neg s \rightarrow V_2(n'')\} \geq \{\top \rightarrow \neg V_2(n'), \top \rightarrow \neg s', \top \rightarrow \neg V_2(n''), \top \rightarrow \neg s''\}$
Optimal decision set: $\langle d_1 = \emptyset, d_2 = \{V_1(n), s\}, d_3 = \emptyset\rangle$
Expected state description: $s_1^1 = s_2^1 = \emptyset$, $s_2^2 = s_1^2 = \{V_1(n), s\}$, $s_3^3 = s_2^3 = \emptyset$
Unfulfilled mental states: $U_1^D = \{\top \rightarrow \neg s\}, U_1^G = \emptyset, U_2^D = U_2^G = \emptyset, U_3^D = \{\top \rightarrow x, \top \rightarrow \neg V_1(n), \top \rightarrow \neg s\}, U_3^G = \{\top \rightarrow x\}$

Given that agent a_1 prefers not to comply with its obligation, agent a_2 has to choose to determine that $\neg x$ is a violation and thus to sanction it. Agent a_2 has no direct motivation to do so apart from the fact that if it decides otherwise, then it can be sanctioned by agent a_3. In fact, agent a_3 has the goal $V_2(n')$ in a context where $\neg x$ is true but agent a_2 does not decide for $V_1(n)$. To deal with this reasoning, agent a_1 has to recursively model agent a_2's decision process: in doing so agent a_1 assumes that agent a_2 recursively models agent a_3 since it depends on agent a_3's decisions for what concerns the obligation to determine violations by agent a_1 and to punish it.

4. SUMMARY

In this paper we propose a logical framework with three dimensions. The first dimension is the set of agents involved, where we distinguished the agent whose behavior is norm governed, the defender agent who monitors and sanctions violations, and the normative agent who issues norms and monitors the defender agent. The second dimension is the mental attitudes attributed to each agent, where we distinguished beliefs, desires and goals each represented by conditional rules. The third dimension are the elements of the norms and obligations, where we distinguished between behavior that counts as a violation, and sanctions. We extended the framework with recursive modelling to model decision making of agents, in particular the interaction between normative and defender agents.

We argue that this approach allows facing the problem of controlling distributed systems, such as virtual communities, by delegating to defender agents the task of monitoring and sanctioning violations. However, these agents are not assumed to be fully cooperative so that they are kept under the control of a judicial authority. Since, as [13] claim, at higher levels the control routines become less risky and require less effort, there is no need of a infinite regression of authorities controlling each other. As [12] discuss, centralized control is not feasible in virtual communities where each participant is both a resource consumer and a resource provider. In fact, there is no authority which is in control of all the resources. Rather the central authority can only issue *meta-policies* [20] concerning the policies regulating the access to the single resources: for example, the central authority can oblige local authorities to grant access to their resources to authorized users, who are thus *entitled* to use the resources.

Since we propose to model delegation of control by means of obligations concerning what is obligatory and what must be sanctioned, our framework can be extended with meta-policies. We can extend this framework for representing obligations by the central authority that local authorities permit of forbid access as well as permissions to forbid or permit access.

In [4], our framework is extended with permissions; while permissions are usually modelled as the dual of obligations, [4] argue that permissions should be modelled as exceptions to obligations under some circumstances: in those contexts, the normative agent adopts the goal not to consider a forbidden behavior as a violation and thus it does not sanction the agent.

Other issues for further research are a complete separation of all three elements of the *trias politica* and the problem of rational norm creation.

5. REFERENCES

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

[2] G. Boella and L. Lesmo. A game theoretic approach to norms. *Cognitive Science Quarterly*, 2(3-4):492–512, 2002.

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

[4] G. Boella and L. van der Torre. Obligations and permissions as mental entities. In *Procs. of IJCAI Workshop on Cognitive Modeling of Agents and Multi-Agent Interactions*, Acapulco, 2003.

[5] C. Boutilier. Toward a logic for qualitative decision theory. In *Procs. of KR-94*, pages 75–86, Bonn, 1994.

[6] R. Boyd, H. Gintis, S. Bowles, and P. J. Richerson. The evolution of altruistic punishment. In *Proceedings of the National Academy of Sciences (USA)*, volume 100, pages 3531–3535, 2003.

[7] 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.

[8] C. Castelfranchi. Engineering social order. In *Proceedings of ESAW00*, Berlin, 2000.

[9] C. Castelfranchi, F. Dignum, C. Jonker, and J. Treur. Deliberate normative agents: Principles and architecture. In *Proceedings of The Sixth International Workshop on Agent Theories, Architectures, and Languages (ATAL-99)*, Orlando, FL, 1999.

[10] R. Conte and C. Castelfranchi. *Cognitive and Social Action*. UCL Press, 1995.

[11] D. Dennett. *The intentional stance*. Bradford Books/MIT Press, Cambridge (MA), 1987.

[12] B. Sadighi Firozabadi and M. Sergot. Contractual access control. In *Procs. of 10th International Workshop of Security Protocols*, Cambridge (UK), 2002.

[13] B. Sadighi Firozabadi and L. van der Torre. Formal models of control systems. In *Procs. of ECAI 1998*, pages 317–318, 1998.

[14] I. Foster, C. Kesselman, and S. Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. *International J. Supercomputer Applications*, 15(3), 2001.

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

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

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

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

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

[20] M. S. Sloman. Policy driven management for distributed systems. *Journal of Network and Systems Management*, 2(4):333–360, 1994.

[21] M. J. Wooldridge and N. R. Jennings. Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 10(2):115–152, 1995.