Norm Governed Multiagent Systems:
The Delegation of Control to Autonomous Agents

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

When agents make decisions, they have to deal with norms regulating the system. In this paper we therefore propose a rule-based qualitative decision and game theory combining ideas from multiagent systems and normative systems. Whereas normative systems are typically modelled as a single authority that imposes obligations and permissions on the agents, our theory is based on a multiagent structure of the normative system. We distinguish between agents whose behavior is governed by norms, so-called defender agents who have the duty to monitor violations of these norms and apply sanctions, and autonomous normative systems that issue norms and watch over the behavior of defender agents. We show that autonomous normative systems can delegate monitoring and sanctioning of violations to defender agents, when bearers of obligations model defender agents, which in turn model autonomous normative systems.

1 Introduction

Norms are studied in multiagent systems for various reasons, for example to increase the stability of such systems. However, there is no consensus how standard frameworks have to be extended to express norms. Moreover, the management of distributed systems such as virtual communities of autonomous agents [13, 4, 5] is not centralized in a single agent since this would endanger the core business of the system [11]. Also, sometimes tasks can be better performed if they are dealt with by the local level in an autonomous way.

In this paper, we address the following two questions.

- How do agents make decisions involving norms? In particular, when do they fulfill or violate norms?

- How can the role of monitoring and sanctioning violations be delegated by the normative system to autonomous agents called *defender* agents [10]?

To address these questions we model norms in qualitative decision and game theory, combining ideas from normative systems (see, e.g., [18] for an overview of theories and applications) and multiagent systems. Whereas normative systems are typically modelled as a single authority that imposes obligations and permissions on the agents, our theory is based on a multiagent structure of the normative system. We distinguish between agents whose behavior is governed by norms, so-called defender agents who have the duty to monitor violations of these norms and to apply sanctions, and autonomous normative systems that issue norms and watch over the behavior of defender agents.

Decision making in this multiagent systems is formalized as follows. Agents are modelled as cognitive belief-desire-intention or BDI agents. Autonomous normative systems create obligations for defender agents, such that a defender agent can be seen as a role defined on the basis of a set of obligations. When making decisions, agents model defender agents, which in turn model the autonomous normative system. In our theory we can study under which conditions this delegation of the control of norms to defender agents is effective, as is illustrated by an example.

In this paper we discuss the technical details of our theory, but some of its assumptions are treated elsewhere. In particular, the unification of normative and multiagent systems can be explained by an attribution of mental attitudes to normative systems, which itself can be explained by social delegation of goals to the normative system [3, 7, 4].

This paper is organized as follows. In Section 2 we discuss the role of defender agents using recursive modelling. In Section 3 we present the qualitative game theory, in Section 4 we define obligations in this theory and in Section 5 we formalize defender agents.

2 Defender agents

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" [15]. Boella and Lesmo [2] distinguish between agents whose behavior is governed by norms, and an autonomous normative system. In their approach, autonomous normative systems are called normative agents and modelled as a single cognitive BDI 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, such a distinction between roles can make social control more effective. In our model, we introduce a distinction between agents who have only the judiciary power, which following Conte and Castelfranchi [10] we call defender agents, and those who have also the legislative one. The task of the normative system 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 a scenario in a virtual community, consider agent 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 3 has not enough resources to control and punish every member of the community it delegates this control task to defender agent 2. Since the virtual community is composed of heterogeneous agents, system 3 cannot assume that defender 2 is a respectful agent who fulfils every obligation imposed on it. Hence, system 3 tries to control defender 2's behavior by means of obligations concerning its task to monitor and punish agent 1. The relation between defender agent 2 and system 3 is that the system motivates the defenders by imposing obligations for the defenders, and punishing their violations.

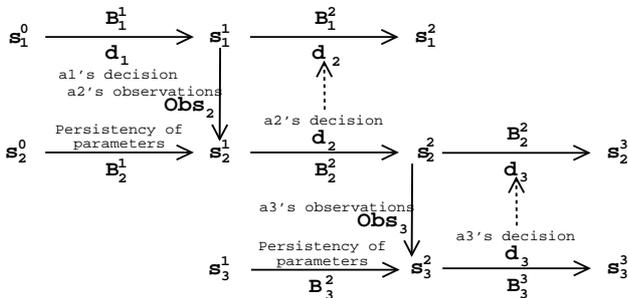The basic picture is visualized in Figure 1 and reflects the deliberation of agent 1 in various stages. Agent 1 is subject to obligations by system 3 that has delegated the control to defender agent 2. Agent 1 is deliberating about the effects of the fulfilment or the violation of the obligations. Agent 2 may recognize and sanction violations. Agent 1 recursively models agent 2's decision and bases its choice on the effects of agent 2's predicted actions. But in doing so, agent 1 has to consider that agent 2 is subject to some obligations to make agent 1 respect its obligations: so in modelling agent 2, it considers that agent 2 recursively models system 3, the normative system who watches over agent 2's behavior.

Figure 1 should be read as follows. Subscripts denote the agent, while superscripts denote the time instant. When agent 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 epistemic state s_1^1. Then agent 2 makes a decision d_2, typically whether it counts this decision as a violation and whether it sanctions agent 1 or not. Now, to find out which decision agent 2 will make, agent 1 has a *profile* of agent 2: it has a representation of the initial state which agent 2 believes to be in and of the following stages. When agent 1 makes its decision, it believes that agent 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 1 believes that its own decision d_1 will have the consequence that agent 2 believes that it is in state s_2^1, due to its observations Obs_2 and the expected consequences of these observations. Agent 1 expects that agent 2 believes that the expected result of decision d_2 is state s_2^2. Finally, agent 1's expected consequences of d_2 from agent 1's point of view are called state s_1^2. Moreover, agent 2 makes a similar reasoning about system 3's decisions. The recursion in modelling other agents stops here, since system 3 does not have to base its decisions on the expected reaction of another agent. Which decision an agent makes depends also on its motivational state which contains among others desire and goal rules (not depicted in Figure 1).

Multiple levels of delegation of controls are also possible. For example, system 3 can delegate the task of controlling the defender agent 2 to another defender. This leads a hierarchy of agents, in which each agent 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 reaching the highest level of authority whose behavior is not controlled. In this construction, the first and the last agent consider three moments in time (like agent 1 and 3 above), and each other agent considers four moments in time (like agent 2 above).

Figure 1. A three agent scenario.

3 Decision and game theory

In this section we present a qualitative game theory for BDI agents based on the iterative model visualized in Figure 1, which is an instance of the game theoretic technique *recursive modelling* introduced by Gmytrasiewicz and Durfee [14]. We start with *decisions*. We write d_i for the complete decision of agent i, which we define as a truth assignment to each *decision variable* of the agent.

Definition 1 (Decisions) *Assume n distinct agents, and let $\{A_1, \ldots, A_n\}$ be n disjoint sets of propositional variables. A literal is a variable or its negation. We write Lit_{A_i} for the set of literals built from A_i. 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 $d_i \subseteq Lit_{A_i}$ such that for each decision variable x in A_i, 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. First, we assume that beliefs of agent i about the world at moment t, written as s_i^t, contains – besides decision variables whose truth value is determined directly by an agent – also *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 [8, 16]. In action languages, they correspond to events and fluents respectively. The example in Figure 1 illustrates that we only consider games in which each agent i makes a decision at moment i, such that an agent has to consider – besides the recursively modelled states – only the moments just before its decision when it observes the previous decision, s_i^{i-2} and s_i^{i-1}, and just after its decision to evaluate the consequences of it, s_i^i and s_i^{i+1}.

Second, the agents' beliefs about how a new state at moment t is constructed out of previous ones is expressed by three sets of *belief rules*, denoted by B_i^t. 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 i includes the epistemic state of agent $i+1$, denoted by σ_{i+1}, unless it is the last agent n.

An epistemic state thus contains most elements depicted in Figure 1. However, when making decisions, not all these elements are available. In particular, for each decision the agent i only knows the belief rules and the initial state s_i^{i-2}; the other states s_i^{i-1}, s_i^i and s_i^{i+1} have to be calculated by applying belief rules. That is, a decision problem contains only a partial epistemic state. How belief rules are applied is formalized in Definition 4 and decision problems are given in Definition 7.

Definition 2 (Epistemic states) *For n agents and $0 \leq t \leq n+1$, let $P^t = \{p^t \mid p \in P = \{p, p', p'', \ldots\}\}$. We write Lit_{P^t} for the literals built from P^t, $Lit_{A_i P^t}$ for literals built from $A_i \cup P^t$, et cetera. 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 $\top \rightarrow l$, where $\top$ stands for tautology. The epistemic state of agent $i < n$ is*

$$\sigma_i = \langle B_i^{i-1}, B_i^i, B_i^{i+1}, \geq_i^B, s_i^{i-2}, s_i^{i-1}, s_i^i, s_i^{i+1}, \sigma_{i+1} \rangle$$

whereas the epistemic state of agent n is identical except that it does not contain the epistemic state of agent $n+1$.

B_i^{i-1} *is a set of rules of* $Lit_{A_{i-1} P^{i-2} P^{i-1}}$;
B_i^i *is a set of rules of* $Lit_{A_{i-1} A_i P^{i-2} P^{i-1} P^i}$;
B_i^{i+1} *is a set of rules of* $Lit_{A_{i-1} A_i A_{i+1} P^{i-2} P^{i-1} P^i P^{i+1}}$;
$B_i = B_i^{i-1} \cup B_i^i \cup B_i^{i+1}$;
$\geq_i^B$ *is a transitive and reflexive relation on the powerset of B_i containing at least the subset relation;*
$s_i^{i-2} \subseteq Lit_{P^{i-2}}$ *is the state before agent $i-1$'s action;*
$s_i^{i-1} \subseteq Lit_{A_{i-1} P^{i-1}}$ *is initial state of agent i's action;*
$s_i^i \subseteq Lit_{A_i P^i}$ *is state after decision d_i of agent i;*
$s_i^{i+1} \subseteq Lit_{A_{i+1} P^{i+1}}$ *is state after decision d_{i+1} of $i+1$;*
$s_i = s_i^{i-2} \cup s_i^{i-1} \cup s_i^i \cup s_i^{i+1}$. *All states are complete.*

The agents' beliefs depend not only on their belief rules, but also on what they can observe. Here we use a simple formalization of this complex phenomena based on an explicit enumeration of all propositions which can be observed. If a proposition describing state s_{i-1}^{i-1} is observable, then agent i knows its value in s_{i-1}^{i-1}. The observations of agent i depend on the state s_{i-1}^{i-1} containing the effects of the decision of agent $i-1$ from agent $i-1$'s point of view.

Definition 3 (Observations) *The propositions observable by agent i, OP_i, are a subset of the description of the stage s_{i-1}^{i-1} (according to agent $i-1$'s point of view) including agent $i-1$'s decision: $P^{i-1} \cup A_{i-1}$. The expected observation of agent i in state s_i^{i-1} is $Obs_i =*

$$\{l \in s_{i-1}^{i-1} \mid l \in OP_i \text{ or } \overline{l} \in OP_i\}$$

What is not observed persists from the initial state s_i^{i-2} from agent i's perspective. Likewise, we model persistency after decisions. Both the consequences of observations and the consequences of decisions are defined using a function $\max$. Intuitively, this function starts with either the observations or the decision, then applies a maximal set of rules with respect to the belief rule ordering $\geq^B$, using intermediate phases Q, Q' and Q''', and finally adds parameters from the previous state. We give also some conventions which facilitate the recursive definition. 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 $B_1^0 = B_n^{n+1} = \emptyset$.

Definition 4 (Respect) *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), f a set of literals, R a set of rules, and $\geq$ a transitive and reflexive relation on the powerset of R containing at least the superset relation, let $out(s, R) = \cup_0^\infty out^i(s, R)$ be the state obtained by $out^0(s, R) = s$ and $out^{i+1}(s, R) = out^i(s, R) \cup \{l \mid l_1 \wedge \ldots \wedge l_n \to l \in R$ and $\{l_1, \ldots, l_n\} \subseteq out^i(s, R)\}$, and let $\max(s, f, R, \geq, t)$ be the set of states obtained by:*

1. *Q is the set of subsets of R which can be applied to $s \cup f$ without leading to inconsistency:*

   $$Q = \{R' \subseteq R \mid out(s \cup f, R') \text{ consistent }\}$$

2. *Q' is the set of maximal elements of Q with respect to set inclusion:*

   $$Q' = \{R' \in Q \mid \nexists R'' \in Q \text{ such that } R' \subset R''\}$$

3. *Q'' is the set of maximal elements of Q' with respect to the $\geq$ ordering:*

   $$Q'' = \{R' \in Q' \mid \nexists R'' \in Q' \text{ and } R'' \geq R', R' \ngeq R''\}$$

4. *O is the set of new elements in $out(s \cup f, R')$:*

   $$O = \{(out(s \cup f, R') \cap Lit_{A_{t+1}P^{t+1}}) \mid R' \in Q''\}$$

5. *$\max(s, f, R, \geq, t)$ is the set of states in O plus some elements persisting from s:*

   $$\max(s, f, R, \geq, t) = \{G \cup s''' \mid G \in O \text{ and }$$
   $$s''' = \{l^{t+1} \mid l^t \in (P^t \cap s) \text{ and } \overline{l^{t+1}} \notin G\}\}$$

By convention $A_0 = d_{n+1} = B_1^0 = B_n^{n+1} = OP_1 = s_0^0 = \emptyset$. $\sigma_i = \langle B_i^{i-1}, B_i^i, B_i^{i+1}, s_i^{i-2}, s_i^{i-1}, s_i^i, s_i^{i+1}, \sigma_{i+1}\rangle$ respects the decision set $\delta = \langle d_1, \ldots, d_n\rangle$ together with the expected observations Obs_i of agent i if

1. *$s_i^{i-1} \in \max(s_i^{i-2}, Obs_i, B_i^{i-1}, \geq_i^B, i-2)$;*

2. *$s_i^i \quad \in \max(s_i^{i-2} \cup s_i^{i-1}, d_i, B_i^i, \geq_i^B, i-1)$;*

3. *$s_i^{i+1} \in \max(s_i^{i-2} \cup s_i^{i-1} \cup s_i^i, d_{i+1}, B_i^{i+1}, \geq_i^B, i)$;*

4. *If $i < n$, then σ_{i+1} respects the decision set $\delta = \langle d_1, \ldots, d_n\rangle$ together with the expected observations Obs_{i+1} of agent $i+1$.*

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 i towards a given state, depending on the context. Like belief rules, desire and goal rules can be conflicting. We express agent characteristics by a priority relation on the rules $\geq_i$ which encode, as detailed in Broersen *et al.* [9], how the agent resolves its conflicts.

Definition 5 (Motivational states) *The motivational state M_i of agent 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 i attributes to agent $i + 1$. The motivational state M_n of agent n is a tuple $\langle D_n, G_n, \geq_n\rangle$.*

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. Here for space reasons, we introduce only a selfish stable agent type, which bases its decisions only on its unsatisfied goals and desires. State s_1 is less preferred than s_2, denoted by $s_1 \leq s_2$, if the unfulfilled rules in s_1 are more preferred than the unfulfilled rules of s_2, denoted by $U(R, s) \geq U(R, s_2)$.

Definition 6 (Selfish stable agents) *Let $U(R, s)$ be the unfulfilled rules of state s,*

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

The unfulfilled mental state description *of agent i is $U_i = \langle U_i^D = U(D_i, s_i), U_i^G = U(G_i, s_i)\rangle$. For selfish stable agents, we have $s_i \leq s_i'$ iff*

1. *$U_i'^G = U(G_i, s_i') \geq_i U_i^G = U(G_i, s_i)$*

2. *if $U_i'^G \geq_i U_i^G$ and $U_i^G \geq_i U_i'^G$ then $U_i'^D \geq_i U_i^D$*

We finally define the optimal decisions. It is again a recursive definition, because optimality of a decision set for agent i is defined in terms of dominance, which is defined in terms of optimality of the decision set for the later agents. There are several definitions of optimality discussed in the literature, here we use a conservative one.

Definition 7 (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 i, and a mental state M_1. A decision set is optimal for a decision problem if it is optimal for each agent i. A decision set is* optimal *for agent i if there is no decision set that dominates it for agent 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 i iff $d_j = d_j'$ for $1 \leq j < i$, they are both optimal for agent 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 Formalization of obligations

The definition of obligation is inspired by Anderson's reduction of deontic logic to alethic modal logic [1], usually written as $O(p) = \Box(\neg p \rightarrow V)$, which says that 'what is obligatory is what necessarily leads to a violation', or less controversial, 'what leads to a bad state'.

Definition 8 (Conditional obligations with sanction)
For m agents, let NS be a set of unstructured norms $\{n, n', n'', \ldots\}$ and assume that some decision variables are so-called violation variables $V_i(n)$ for $n \in NS$ and $1 \le i \le m$, *to be read as 'agent i's behavior counts as a violation of norm n'.*

Agent i believes that it is obliged to decide to do x (a literal in $Lit_{A_i P^i P^{i+1}}$) with sanction s (a literal built from a decision variable in A_{i+1}) under condition q (a sequence of literals in $Lit_{A_i P^i P^{i+1}}$), denoted by $O_{i,i+1}(x, s|q)$, iff for some $n \in NS$:

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

2. $\top \rightarrow \overline{s} \in D_{i+1}$: *agent i believes that agent $i+1$ desires not to sanction.*

3. $\top \rightarrow \overline{s} \in D_i$: *agent i has the desire not to be sanctioned.*

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

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

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

The definition is based on three simplifications. First, there is no 'logic of rules' as for example proposed in [17]. Second, the conditions can be partitioned into preconditions of creating a norm (first three conditions), and conditions which follow from such a creation (last three conditions). This issue is explored in [7]. Third, sanctions are defined as decision variables only, not for parameters. In the latter case, we may add a condition that there is a way to apply the sanction.

The following example illustrates an obligation to achieve parameter p^1 of a stable agent 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. By convention we only give positive literals in states; all propositional variables not mentioned are assumed to be false.

Example 1 $O_{1,2}(p^1, s)$
$s_1^0 = \emptyset, B_1 = \{x \rightarrow p^1\}, \ge_1^B = \emptyset, x \in A_1, p^1 \in P^1,$
$G_1 = \emptyset, D_1 = \{\top \rightarrow \neg x, \top \rightarrow \neg s\},$
$\ge_1 = \{\top \rightarrow \neg s\} \ge \{\top \rightarrow \neg x\}$
$s_2^0 = \emptyset, OP_2 = A_1 \cup P^1, B_2 = \{x \rightarrow p^1\}, \ge_2^B = \emptyset,$
$V_1(n) \in A_2, s \in A_2, n \in NS,$
$G_2 = \{\top \rightarrow p^1, \neg p^1 \rightarrow V_1(n), V_1(n) \rightarrow s\},$
$D_2 = \{\top \rightarrow p^1, \neg p^1 \rightarrow V_1(n), V_1(n) \rightarrow s, \top \rightarrow \neg V_1(n), \top \rightarrow \neg s\},$
$\ge_2 \supseteq \{\neg p^1 \rightarrow V_1(n)\} > \{\top \rightarrow \neg V_1(n), \top \rightarrow \neg s\}$

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:
$U_1^D = \{\top \rightarrow \neg x\}, U_1^G = \emptyset, U_2^D = U_2^G = \emptyset$

If agent 1 decides to do x, $d_1 = x$, then we have $s_1^1 \in \max(s_1^0, d_1, B_1^1, \ge_1^B, 0) = \{\{x, p^1\}\}$ by Definition 4 of respecting mental states. Agent 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 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 4). 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 6).

For what concerns agent 2's attitudes, its unconditional desire and goal that agent 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 unfulfilled.

Whatever other decision agent 2 would have taken, it could not satisfy more goals or desires, so $d_2 = \emptyset$ is a minimal and optimal decision - Definition 7. E.g. $d_2'' = \{s\}$ leaves $\top \rightarrow \neg s$ unsatisfied: $\{\top \rightarrow \neg s\} \ge_2 \emptyset$ (in fact, $\ge_2$ contains the subset relation) and then $U''_2^D = \{\top \rightarrow \neg s\} \ge U_2^D = \emptyset$ for a stable agent.

Had agent 1's decision been $d_1' = \emptyset$, agent 2 would have chosen $d_2' = \{V_1(n), s\}$. The unfulfilled desires and goals in state $s_1' = s_2' = \{V_1(n), s\}$: $U'_1^D = \{\top \rightarrow \neg s\}$, $U'_1^G = \emptyset$, $= U'_2^D = \{\top \rightarrow p^1, \top \rightarrow \neg V_1(n), \top \rightarrow \neg s\}$, $U'_2^G = \{\top \rightarrow p^1\}$.

How does agent 1 take a decision between d_1 and d_1'? Since its agent type is *stable* (Definition 6) it compares which of its goals and desires remain unsatisfied: $U_1^G = U'_1^G = \emptyset$ but $U'_1^D = \{\top \rightarrow \neg s\} \ge U_1^D = \{\top \rightarrow \neg x\}$. And hence, the optimal state (Definition 7) is s_1: $s_1 = \{x, p^1, p^2\} \le s_1' = \{V_1(n), s\}$.

5 Formalization of defender agents

We now introduce the distinction between a defender agent $i+1$ who has the duty to enforce a norm n and a normative agent $i+2$ who imposes by means of norms n' and n'' to $i+1$ the duty to watch over a norm n. The definition contains the same six conditions as the definition in the previous section. The crucial distinction between the normative system and the defender agent, is that the defender agent does not desire to count the agent's behavior as a violation and to sanction it.

Definition 9 (Delegated obligations) *Agent i believes that it is obliged to decide to do x (a literal in $Lit_{A_i P^i P^{i+1}}$) with sanction s, a literal built from a decision variable of A_{i+1} performed by defender $i+1$, and sanctions s' and s'' for the defender, literal built from a decision variable of A_{i+2}, on behalf of the normative agent $i+2$ (where $V_i(n) \in A_{i+1}$), denoted by $O_{i,i+1,i+2}(x,s,s',s''|q)$, iff for some $n \in NS$:*

1. $q \rightarrow x \in D_{i+2} \cap G_{i+2}$*: agent i believes that agent $i+2$ desires and has as a goal that x.*

2. $\top \rightarrow \overline{s} \in D_{i+2}$*: agent i believes that agent $i+2$ desires not to sanction.*

3. $\top \rightarrow \overline{s} \in D_i$*: agent i has the desire not to be sanctioned.*

4. $\top \rightarrow \neg V_i(n) \in D_{i+2}$*: agent i believes that agent $i+2$ desires that there is no violation.*

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

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

Item 5 and 6 imply by the definition of obligation in Definition 8: Given that these two goals are normative goals for agent $i+1$, if it adopts them, then i is in the same situation as before the introduction of the defender agent.

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

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

Further definitions with multiple levels of defenders are possible since obligations at Items 5 and 6 can be delegated to a second defender agent and so on.

Moreover, item 5 and 6 imply by the definition of obligation in Definition 8 also that introducing a defender agent leads to the addition of two new norms.

In the second example we show a three agent situation where agent 2 is the defender of the obligation to do x on behalf of the normative system 3 (Definition 9). However, 1 prefers to violate the obligation with respect to not being sanctioned.

Example 2 $O_{1,2,3}(x,s,s',s''|\top)$ *and thus there is a norm n such that $O_{2,3}(V_1(n),s'|\neg x)$ and $O_{2,3}(s,s''|V_1(n))$. We call the norms related to the latter obligations n' and n''.*

$s_1^0 = \emptyset, B_1 = \emptyset, \geq_1^B = \emptyset, x \in A_1,$
$G_1 = \emptyset, D_1 = \{\top \rightarrow \neg x, \top \rightarrow \neg s\},$
$\geq_1 = \{\top \rightarrow \neg x\} \geq \{\top \rightarrow \neg s\}$
$s_2^0 = \emptyset, OP_2 = A_1 \cup P^1, B_2 = \emptyset, \geq_2^B = \emptyset,$
$V_1(n) \in V \cap A_2, s \in A_2,$
$G_2 = \emptyset, D_2 = \{\top \rightarrow \neg s', \top \rightarrow \neg s''\},$
$s_3^0 = \emptyset, OP_3 = A_2 \cup P^2, B_3 = \emptyset, \geq_3^B = \emptyset,$
$V_2(n'), V_2(n'') \in A_3, s', s'' \in A_3, n, n', n'' \in NS,$
$G_3 = \{\top \rightarrow x, \neg x \rightarrow V_1(n), V_1(n) \rightarrow s, \neg x \wedge \neg V_1(n) \rightarrow V_2(n'),$
$V_2(n') \rightarrow s', V_1(n) \wedge \neg s \rightarrow V_2(n''), V_2(n'') \rightarrow s''\},$
$D_3 = \{\top \rightarrow x, \neg x \rightarrow V_1(n), V_1(n) \rightarrow s,$
$\neg x \wedge \neg V_1(n) \rightarrow V_2(n'), V_2(n') \rightarrow s',$
$V_1(n) \wedge \neg s \rightarrow V_2(n''), V_2(n'') \rightarrow s'', \top \rightarrow \neg V_1(n),$
$\top \rightarrow \neg s, \top \rightarrow \neg V_2(n'), \top \rightarrow \neg s', \top \rightarrow \neg V_2(n''),$
$\top \rightarrow \neg s''\},$
$\geq_3 \supseteq \{\neg x \wedge \neg V_1(n) \rightarrow V_2(n'), V_1(n) \wedge \neg s \rightarrow V_2(n'')\} \geq$
$\{\top \rightarrow \neg V_2(n'), \top \rightarrow \neg s', \top \rightarrow \neg V_2(n''), \top \rightarrow \neg s''\}$

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 1 prefers not to comply with its obligation, agent 2 has to choose to determine that $\neg x$ is a violation and thus to sanction it. Agent 2 has no direct motivation to do so apart from the fact that if it decides otherwise, then it can be sanctioned by agent 3. In fact, agent 3 has the goal $V_2(n')$ in a context where $\neg x$ is true but agent 2 does not decide for $V_1(n)$. To deal with this reasoning, agent 1 has to recursively model agent 2's decision process: in doing so agent 1 assumes that agent 2 recursively models agent 3 since it depends on agent 3's decisions for what concerns the obligation to determine violations by agent 1 and to punish it.

6 Summary and concluding remarks

In this paper we show how agents make decisions involving norms, and in particular, when they fulfill or violate norms, in 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. the agents make decisions in this framework by recursively modelling the reaction of their behavior by the normative system.

Moreover, we show how the role of monitoring and sanctioning violations can be delegated by the normative system to autonomous agents called defender agents [10]. With this extension, the agents model the defender agents, which recursively model the autonomous normative systems. Our approach can be used to control distributed systems by delegating the task of monitoring and sanctioning violations to defender agents, even when these agents are not assumed to be fully cooperative. Since, as [12] 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 [11] 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* [19] 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. Moreover, in [6] our framework is extended with permissions. While permissions are usually modelled as the dual of obligations, we 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.

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 *Procs. of AAMAS'03*, 2003. ACM Press.

[4] G. Boella and L. van der Torre. Decentralized control: Obligations and permissions in virtual communities of agents. In *Procs. of ISMIS'03*, 2003. Springer Verlag.

[5] G. Boella and L. van der Torre. Local policies for the control of virtual communities. In *Procs. of IEEE/WIC WI'03*, 2003. IEEE Press.

[6] 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.

[7] G. Boella and L. van der Torre. Rational norm creation. In *Procs. of ICAIL'03*, pages 81–82, Edinburgh, 2003. ACM Press

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

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

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

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

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

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

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

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

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

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

[18] J.-J. Meyer and R. Wieringa. *Deontic Logic in Computer Science: Normative System Specification*. John Wiley & Sons, 1993.

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