Regulative and Constitutive Norms in Normative Multiagent Systems

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

Leendert van der Torre[*]
CWI Amsterdam & Delft University of Technology
The Netherlands
e-mail: torre@cwi.nl

Abstract

In this paper we introduce a formal framework for the construction of normative multiagent systems, based on Searle's notion of the construction of social reality. Within the structure of normative multiagent systems we distinguish between regulative norms that describe obligations, prohibitions and permissions, and constitutive norms that regulate the creation of institutional facts as well as the modification of the normative system itself. Using the metaphor of normative systems as agents, we attribute mental attitudes to the normative system. In particular, we formalize regulative norms as goals of the normative system, and constitutive norms as beliefs of the normative system. Agents reason about norm creation using recursive modelling.

Introduction

Normative multiagent 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" (Jones & Carmo 2001). Many theories and applications of multiagent systems such as electronic commerce, virtual communities, theories of fraud and deception, of trust dynamics and reputation, *et cetera*, can fruitfully employ the notion of a normative system regulating an agent society. For example, norms allow to regulate systems of heterogeneous agents where the absence of a central design prevents enforcing a given behavior by constraining the architecture. In earlier work we have studied the representation of norms and their role in reasoning. One question is whether norms require an explicit representation, and if they do, whether they should be represented as primitive entities or in terms of other explicitly represented notions like beliefs and goals of agents. Another question is how agents take decisions when they are subject to norms.

Most formalizations of normative systems, including our own previous work on normative multiagent systems, identify norms with obligations, prohibitions and permissions which specify the ideal behavior of agents. Searle (1995) observes that to describe the construction of social reality one needs, besides regulative norms like obligations, prohibitions and permissions, also what he calls constitutive norms, which define that something *counts as* something else for a given institution. We are interested in formal models of how the normative system creates institutional reality and regulates the changes that the agents of the system can perform by means of constitutive rules. The research questions we address in this paper are as follows.

1. How to define a formal framework for normative multiagent systems including regulative and constitutive norms?
2. How to reason about modifications of the normative system in this framework?
3. How to play games and other behaviors in this normative multiagent system, including violations of norms?

The challenge of the framework is to balance the first and the last question, that is, to balance on the one hand logical techniques to describe the agents, the norms, the institutional structures, *et cetera*, and on the other hand game theoretic techniques to describe the games the agents play. We proceed from the logical perspective to describe the static structure of a normative multiagent system, and only consider a limited set of games. We use input/output logic (Makinson & van der Torre 2000; 2001) to reason about mental attitudes in the normative multiagent system. We do not incorporate an action logic, but we define the games of the agents in terms of decision variables.

As a running example, we consider an agent who likes to cultivate some crop, and believes that it is sanctioned if it does not own the field. Since putting a fence around the field counts as making the field its property, its optimal decision is to build a fence and cultivate the crop. Moreover, if a second obligation to have an authorization of the land register is created when it becomes an estate owner, and the agent's applying for registration counts as being authorized according to some constitutive norm, then the optimal decision of the agent includes applying for the registration.

The paper is organized as follows. We first introduce the normative multiagent system, regulative and constitutive norms, and games. Thereafter we use constitutive norms to regulate modifications of the normative system.

[*] Supported by the ArchiMate research project.

The construction of social reality

We introduce a formal framework for the construction of normative multiagent systems, based on Searle's notion of the construction of social reality. Searle (1969) argues that there is a distinction between two types of rules.

> "Some rules regulate antecedently existing forms of behaviour. For example, the rules of polite table behaviour regulate eating, but eating exists independently of these rules. Some rules, on the other hand, do not merely regulate an antecedently existing activity called playing chess; they, as it were, create the possibility of or define that activity. The activity of playing chess is constituted by action in accordance with these rules. The institutions of marriage, money, and promising are like the institutions of baseball and chess in that they are systems of such constitutive rules or conventions" (Searle 1969, p. 131)

Within normative multiagent systems we distinguish between regulative norms that describe obligations, prohibitions and permissions, and constitutive norms that regulate the creation of institutional facts like property, marriage and money, as well as the modification of normative system itself. Constitutive norms are introduced in our normative multiagent systems for the following three reasons.

First of all, regulative norms are not categorical, but conditional: they specify all their applicability conditions. In case of complex and rapidly evolving systems new situations arise which should be considered in the conditions of the norms. Thus, new regulative norms must be introduced each time the applicability conditions must be extended to include new cases. In order to avoid changing existing norms or adding new ones, it would be more economic that regulative norms could factor out particular cases and refer, instead, to more abstract concepts only. Hence, the normative system should include some mechanism to introduce new institutional categories of abstract entities for classifying possible states of affairs. Norms could refer to this institutional classification of reality rather than to the commonsense classification (Breuker, Valente, & Winkels 1997): changes to the conditions of the norms would be reduced to changes to the institutional classification of reality.

Second, the dynamics of the social order which the normative system aims to achieve is due to the evolution of the normative system over time, which introduces new norms, abrogates outdated ones, and, as just noticed, changes its institutional classification of reality. So the normative system must specify how the normative system itself can be changed by introducing new regulative norms and new institutional categories, and specify by whom the changes can be done.

Third, the dynamics of a normative system includes the possibility that not only new norms are introduced by the agents playing a legislative role, but also that ordinary agents create new obligations, prohibitions and permissions concerning specific agents. This activity is particularly important in applications for e-commerce where it is necessary to model contracts which introduce new normative relations among agents, like the duty to pay a fee for a service (Dellarocas 2001; Neal *et al.* 2003).

Methodology

There are many formalisms to describe the evolution of normative multiagent systems. In principle, it can be formalized by for example state transition systems like dynamic deontic logic, action languages as developed in artificial intelligence, or generalizations of classical decision theory like Markov decision processes. In this paper we proceed from the logical perspective, and we only consider a limited set of games. More precisely, we model normative multiagent systems as detailed but static rule-based structures, we model games by a simple protocol that only contains a finite sequence of agents making a move, and we specify decision problems by an initial normative multiagent system and a protocol. The behavior is a sequence of decisions of the agents as specified by the game protocol, and the effects of these decisions are updated normative multiagent systems as well as other effects.

Moreover, to formalize games and other behaviors in the normative multiagent system, we attribute mental states to agents as well as to normative systems, thus we model them as agents too. This has been proposed by Boella and Lesmo (2002) and may be seen as an instance of Dennett's *intentional stance* (Dennett 1987). The main reason is that it facilitates the specification of games in which agents take the (autonomous!) normative system into account. For example, an agent considers whether its actions will lead to a sanction of the normative system. The advantage of the approach is that standard techniques developed in decision and game theory can be applied to normative reasoning. Moreover, the use of the agent metaphor is also useful to describe the structural relations between agents playing roles in a normative system like legislators creating norms, judges counting behavior as violations and associating sanctions, policemen enforcing sanctions, and citizens signing contracts. For example, the normative agent may contain the role of a legislator, a judge and a policeman. Finally, obligations of the agents can be formalized as desires or goals of the normative agent. This representation may be paraphrased as "Your wish is my command", because the desires or wishes of the normative agent are the obligations or commands of the other agents. The goals of the normative system describe the ideal behavior of the system. Likewise, constitutive norms can be formalized as beliefs of the normative agent. This is explained in detail later in this paper.

The application of the agent metaphor is in this paper only a useful technical trick, but we note that it can also be explained from a more philosophical point of view. In particular, it is inspired by the interpretation of normative systems as dynamic social order (Boella & van der Torre 2003a; 2004). According to Castelfranchi (2000), a social order is a pattern of interactions among interfering agents "such that it allows the satisfaction of the interests of some agent A". These interests can be a shared goal, a value that is good for everybody or for most of the members. But the agents attribute to the normative system, besides goals, also the ability to autonomously enforce the conformity of the agents to the norms. To achieve its goal the normative system forms subgoals to consider as a violation the behavior not conforming to obligations, and to sanction violations.

Structure of normative multiagent system

The conceptual model of the normative multiagent system is visualized in Figure 1, in which we distinguish the multiagent system (straight lines) and additions for the normative system (dotted lines). Following the usual conventions of, for example, class diagrams in the unified modelling language (UML), □ is a concept or set, — and → are associations between concepts, and —▷ is the "is-a" or subset relation. The logical structure of the associations is detailed in the definitions below.

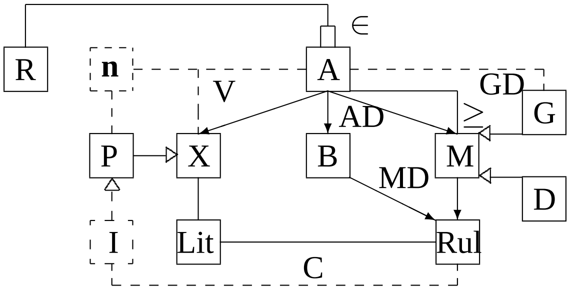

Figure 1: Conceptual model of a normative system.

The definition of the agents (A) is inspired by the rule based BOID architecture (Broersen *et al.* 2002). Beliefs (B), desires (D) and goals (G) are represented by different sets representing the epistemic and motivational states of the agent. We assume that the base language contains boolean variables and logical connectives. The variables (X) are either *decision variables* of an agent, which represent the agent's actions and whose truth value is directly determined by it, or *parameters* (P), which describe both the state of the world and *institutional facts*, and whose truth value can only be determined indirectly. Our terminology is borrowed from (Lang, van der Torre, & Weydert 2002). *Desires* (D_b) and *goals* (G_b) express the attitudes of the agent b towards a given state, depending on the context. Agents may share decision variables or mental attitudes, though this complication is not used in this paper.

Given the same set of mental attitudes, agents reason and act differently: when facing a conflict among their motivations, different agents prefer to fulfill different goals and desires. We express these agent characteristics by a priority relation ($\geq$) on the mental attitudes which encode, as detailed in Broersen *et al.* (2002), how the agent resolves its conflicts. The priority relation is defined on the powerset of the motivations such that a wide range of characteristics can be described, including social agents that take the desires or goals of other agents into account. The priority relation contains at least the subset-relation which expresses a kind of independence between the motivations.

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

- *the agents A, variables X, agent beliefs B, desires D and goals G are five finite disjoint sets. We write $M = D \cup G$ for the motivations defined as the union of the desires and goals.*
- *an agent description $AD : A \to 2^{X \cup B \cup M}$ is a total function that maps each agent to sets of variables (its decision variables), beliefs, desires and goals, but that does not necessarily assign each variable to at least one agent. For each agent $b \in A$, we write X_b for $X \cap AD(b)$, and B_b for $B \cap AD(b)$, etc. We write parameters $P = X \setminus \cup_{b \in A} X_b$.*
- *a priority relation $\geq$: $A \to 2^M \times 2^M$ is a function from agents to a transitive and reflexive relation on the powerset of the motivations containing at least the subset relation. We write $\geq_b$ for $\geq (b)$.*

The following example illustrates a single agent, who likes to cultivate crop, does not like to be sanctioned, and who can also build a fence around a field.

Example 1 $A = \{\mathbf{a}\}$, $X_{\mathbf{a}} = \{crop, fence\}$, $P = \{s\}$, $D_{\mathbf{a}} = \{d_1, d_2\}$, $\geq_{\mathbf{a}} = \{d_2\} \geq \{d_1\}$. *There is a single agent, agent* $\mathbf{a}$*, who can build a fence and grow crop. Moreover, it can be sanctioned. It has two desires, one to cultivate crop (d_1), another one not to be sanctioned (d_2). The second desire is more important than the first one.*

A multiagent system contains, besides an agent set, an organizational structure based on roles and hierarchical containment relations. Moreover, beliefs, desires and goals are abstract concepts which are described by rules (R) built from literals (L). A technical reason to distinguish mental attitudes from rules is to facilitate the description of the priority ordering. To keep the framework simple and to focus on the subject of this paper, we do not introduce nested mental attitudes, such as beliefs or desires of an agent about beliefs or desires about another agent. The consequence of the absence of such *agent profiles* is that we can formalize only a relatively simple kind of games, as is explained later in this paper.

Definition 2 (Multiagent system) *A multiagent system is a tuple* $\langle A, R, \in, X, B, D, G, AD, MD, \geq \rangle$*, where* $\langle A, X, B, D, G, AD, \geq \rangle$ *is an agent set, and:*

- *the roles R are a finite set disjoint from A, X, B, D and G.*
- *the containment relation $\in$: $R \to 2^{A \times A}$ is for each role an irreflexive transitive relation on the set of agents.*
- *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 mental description $MD : (B \cup M) \to$ Rul(X) is a total function from the sets of beliefs, desires and goals to the set of rules built from X. For a set of mental attitudes $S \subseteq B \cup M$, we write $MD(S) = \{MD(s) \mid s \in S\}$.*

Our running example illustrates the mental description; the roles and the related hierarchical structure of the agents are illustrated after the introduction of the normative system.

Example 2 (Continued) $MD(d_1) = \top \to crop$, $MD(d_2) = \top \to \neg s$.

In the description of the normative system, we do not introduce norms explicitly, but we represent several concepts which are illustrated in the following sections. Institutional facts (I) represent legal abstract categories which depend on the beliefs of the normative agent and have no direct counterpart in the world. $F = P \setminus I$ are what Searle calls "brute facts": physical facts produced by the actions of the agents. $V(x, b)$ represents the decision of agent $\mathbf{n}$ that recognizes x as a violation by agent b. The goal distribution $GD(b) \subseteq G_{\mathbf{n}}$ represents the goals of agent $\mathbf{n}$ the agent b is responsible for.

Definition 3 (Normative system) *A normative multiagent system, written as $NMAS$, is a tuple*

$$\langle A, R, \in, X, B, D, G, AD, MD, \geq, \mathbf{n}, I, V, GD \rangle$$

where the tuple $\langle A, R, \in, X, B, D, G, AD, MD, \geq \rangle$ is a multiagent system, and

- *the normative agent $\mathbf{n} \in A$ is an agent.*
- *the institutional facts $I \subseteq P$ are a subset of the parameters, and we write $F = P \setminus I$ for brute facts.*
- *the norm description V : Lit$(X_{\mathbf{a}} \cup P) \times A \to X_{\mathbf{n}} \cup P$ is a function from the literals and the agents to the decision variables of the normative agent together with the parameters.*
- *the goal distribution $GD : A \to 2^{G_{\mathbf{n}}}$ is a function from the agents to the powerset of the goals of the normative agent, such that if $L \to l \in MD(GD(b))$, then $l \in$ Lit$(X_{\mathbf{a}} \cup P)$.*

Our running example illustrates the role hierarchy and the normative agent. Agent $\mathbf{a}$ is a member of the normative system, and the normative agent has the goal that crop is only cultivated on property.

Example 3 (Continued) $A = \{\mathbf{a}, \mathbf{n}\}$, $R = \{$*member*$\}$, $\in$(*member*) $= \{\langle \mathbf{a}, \mathbf{n} \rangle\}$. *There is a new agent, agent $\mathbf{n}$, and a role called member. Agent $\mathbf{a}$ is a member of normative system $\mathbf{n}$.*

$X_{\mathbf{n}} = \{s\}$, $P = \{$*property*$\}$, $D_{\mathbf{n}} = G_{\mathbf{n}} = \{g_1\}$, $MD(g_1) = \{$*crop* $\to$ *property*$\}$, $GD(\mathbf{a}) = \{g_1\}$. *Agent $\mathbf{n}$ can sanction agent $\mathbf{a}$, because s is no longer a parameter. It has the goal that crop is build on property only, and it has distributed this goal to agent $\mathbf{a}$.*

Before we can define the regulative and constitutive norms, we have to introduce a logic of rules. We use a simplified version of the input/output logics introduced in (Makinson & van der Torre 2000; 2001). A rule set is a set of ordered pairs $p \to q$. For each such pair, the body p is thought of as an input, representing some condition or situation, and the head q is thought of as an output, representing what the norm tells us to be desirable, obligatory or whatever in that situation. We use input/output logics since they do not necessarily satisfy the identity rule. Makinson and van der Torre write (p, q) to distinguish input/output rules from conditionals defined in other logics, to emphasize the property that input/output logic does not necessarily obey the identity rule. In this paper we do not follow this convention. Following Makinson and van der Torre, we call operations that satisfy the identity rule *throughput operations*.

In this paper, input and output are respectively a set of literals and a literal. We use a simplified version of input/output logics, since it keeps the formal exposition simple and it is sufficient for our purposes here. In Makinson and van der Torre's input/output logics, the input and output can be arbitrary propositional formulas, not just sets of literals and literal as we do here. Consequently, in input/output logic there are additional rules for conjunction of outputs and for weakening outputs.

Definition 4 (Input/output logic) *Let a rule set S be a set of rules $\{p_1 \to q_1, \ldots, p_n \to q_n\}$, read as 'if input p_1 then output q_1', etc., and consider the following proof rules strengthening of the input (SI), disjunction of the input (OR), cumulative transitivity (CT) and Identity (Id) defined as follows:*

$$\frac{p \to r}{p \wedge q \to r} SI \qquad \frac{p \wedge q \to r, p \wedge \neg q \to r}{p \to r} OR$$

$$\frac{p \to q, p \wedge q \to r}{p \to r} CT \qquad \frac{}{p \to p} Id$$

The following output operators are defined as closure operators on the set S using the rules above.

out_1: *SI*	*(simple-minded output)*
out_2: *SI+OR*	*(basic output)*
out_3: *SI+CT*	*(simple-minded reusable output)*
out_4: *SI+OR+CT*	*(basic reusable output)*

Moreover, the following four throughput operators are defined as closure operators on the set S.

- out_i^+: out_i*+Id (throughput)*

We write $out(S)$ for any of these output operations and $out^+(S)$ for any of these throughput operations. We also write $l \in out(S, L)$ iff $L \to l \in out(S)$, and $l \in out^+(S, L)$ iff $L \to l \in out^+(S)$.

The following definition of the so-called input/output and output constraints checks whether the derived conditional goals are consistent with the input.

Definition 5 *(Makinson & van der Torre 2001) Let S be a set of rules, and C a set of literals. S is consistent with C, written as $cons(S|C)$, iff there do not exist two contradictory literals in $C \cup out(S, C)$. We write $cons(S)$ for $cons(S|\emptyset)$.*

Due to space limitations we have to be brief on technical details with respect to input/output logics, see (Makinson & van der Torre 2000; 2001) for the semantics of input/output logics, further details on its proof theory, alternative constraints, and examples.

In the following sections, we use two input/output logics. First, to define whether a desire or goal implies another one, we use an output operation written as out. Moreover, to define the application of a set of belief rules to a set of literals, we use a throughput operation, written as out^+. We do not specify which output and throughput operations are used, but in the examples we assume the use of out_3 and out_3^+. Thus, in this paper we consider $out(MD(M))$ and $out^+(MD(B))$. To simplify the notation we write $out(M)$ and $out^+(B)$ instead.

Regulative norms

Regulative norms are based on the notion of conditional obligation with an associated sanction. Obligations are defined in terms of goals of the normative agent **n**, because regulative norms refer to states of affairs which are currently false or that can eventually be false. The rules in the definition of obligation are only motivations, and not beliefs, because a normative system may not recognize that a violation counts as such, or that it does not sanction it. Both the recognition of the violation and the application of the sanction are the result of autonomous decisions of the normative system that is modelled as an agent.

The definition of obligation contains several clauses. The first and central clause of our definition defines obligations of agents as goals of the normative agent, following the 'your wish is my command' metaphor. It says that the obligation is implied by the desires of the normative agent **n**, implied by the goals of agent **n**, and it has been distributed by agent **n** to the agent. The latter two steps are represented by $out(GD(\mathbf{a}))$.

The second and third clause can be read as "the absence of p is considered as a violation". The association of obligations with violations is inspired by Anderson's reduction of deontic logic to alethic logic (Anderson 1958). The third clause says that the agent desires that there are no violations, which is stronger than that it does not desire violations, as would be expressed by $\top \rightarrow V(\sim x, a) \notin out(D_{\mathbf{n}})$.

The fourth and fifth clause relate violations to sanctions. The fourth clause says that the normative system is motivated not to count behavior as a violation and apply sanctions as long as their is no violation, because otherwise the norm would have no effect. Finally, for the same reason the last clause says that the agent does not like the sanction.

Definition 6 (Obligation) *Let $NMAS = \langle A, R, \in, X, B, D, G, AD, MD, \geq, \mathbf{n}, I, V, GD \rangle$ be a normative multiagent system. Agent $\mathbf{a} \in A$ is obliged to see to $x \in Lit(X_{\mathbf{a}} \cup P)$ with sanction $s \in Lit(X_{\mathbf{n}} \cup P)$ if $Y \subseteq Lit(X_{\mathbf{a}} \cup P)$ in $NMAS$, written as $NMAS \models O_{\mathbf{an}}(x, s|Y)$, if and only if:*

1. $Y \rightarrow x \in out(D_{\mathbf{n}}) \cap out(GD(\mathbf{a}))$*: if Y then agent $\mathbf{n}$ desires and has as a goal that x, and this goal has been distributed to agent $\mathbf{a}$.*
2. $Y \cup \{\sim x\} \rightarrow V(\sim x, \mathbf{a}) \in out(D_{\mathbf{n}}) \cap out(G_{\mathbf{n}})$*: if Y and $\sim x$, then agent $\mathbf{n}$ has the goal and the desire $V(\sim x, \mathbf{a})$: to recognize it as a violation by agent $\mathbf{a}$.*
3. $\top \rightarrow \neg V(\sim x, \mathbf{a}) \in out(D_{\mathbf{n}})$*: agent $\mathbf{n}$ desires that there are no violations.*
4. $Y \cup \{V(\sim x, \mathbf{a})\} \rightarrow s \in out(D_{\mathbf{n}}) \cap out(G_{\mathbf{n}})$*: if Y and agent $\mathbf{n}$ decides $V(\sim x, \mathbf{a})$, then agent $\mathbf{n}$ desires and has as a goal that it sanctions agent $\mathbf{a}$.*
5. $Y \rightarrow \sim s \in out(D_{\mathbf{n}})$*: if Y, then agent $\mathbf{n}$ desires not to sanction. This desire of the normative system expresses that it only sanctions in case of violation.*
6. $Y \rightarrow \sim s \in out(D_{\mathbf{a}})$*: if Y, then agent $\mathbf{a}$ desires $\sim s$, which expresses that it does not like to be sanctioned.*

Since conditions of obligations are sets of decision variables and parameters, institutional facts can be among them. In this way it is possible that regulative norms refer to institutional abstractions of the reality rather than to physical facts only. Obligations are illustrated in our running example.

Example 4 (Continued)
$MD(g_2) = \{\mathit{crop}, \sim\!\mathit{property}\} \rightarrow V(\sim property, \mathbf{a})$
$MD(g_3) = \top \rightarrow \neg V(\sim\!\mathit{property}, \mathbf{a})$
$MD(g_4) = \{\mathit{crop}, V(\sim\!\mathit{property}, \mathbf{a})\} \rightarrow s$
$MD(g_5) = \mathit{crop} \rightarrow \sim s$
$\{g_1, g_2, g_4\} = G_{\mathbf{n}}$, $G_{\mathbf{n}} \cup \{g_3, g_5\} = D_{\mathbf{n}}$, $\{g_1\} = GD(\mathbf{a})$

$NMAS \models O_{\mathbf{an}}(\mathit{property}, s \mid \mathit{crop})$*, since:*

1. $\mathit{crop} \rightarrow \mathit{property} \in out(D_{\mathbf{n}}) \cap out(GD(\mathbf{a}))$
2. $\{\mathit{crop}, \sim\!\mathit{property}\} \rightarrow V(\sim \mathit{property}, \mathbf{a}) \in out(D_{\mathbf{n}}) \cap out(G_{\mathbf{n}})$
3. $\top \rightarrow \neg V(\sim\!\mathit{property}, \mathbf{a}) \in out(D_{\mathbf{n}})$
4. $\{\mathit{crop}, V(\sim\!\mathit{property}, \mathbf{a})\} \rightarrow s \in out(D_{\mathbf{n}}) \cap out(G_{\mathbf{n}})$
5. $\mathit{crop} \rightarrow \sim s \in out(D_{\mathbf{n}})$
6. $\mathit{crop} \rightarrow \sim s \in out(D_{\mathbf{a}})$

One has to be careful when defining multiple obligations with the same sanction. For example, when both for speeding and for parking in a no parking street there is a penalty of 100 euros, then it is implicitly assumed that one can also be sanctioned 200 euros for violating both obligations at the same time. We do not discuss this problem any further in this paper, since it has to do with the formalization of resources which is beyond the scope of this paper. We simply assume that there is a separate sanction for each obligation.

Other regulative norms like prohibitions and permissions can be defined in an analogous way. Prohibitions are obligations concerning negated variables.

Definition 7 (Prohibition) *Agent $\mathbf{a} \in A$ is prohibited to see to $x \in Lit(X_{\mathbf{a}} \cup P)$ with sanction $s \in Lit(X_{\mathbf{n}} \cup P)$ if $Y \subseteq Lit(X_{\mathbf{a}} \cup P)$ in $NMAS$, written as $NMAS \models F_{\mathbf{an}}(x, s|Y)$, if and only if $NMAS \models O_{\mathbf{an}}(\sim x, s|Y)$*

Permissions are defined as exceptions to obligations. A permission to do x is an exception to a prohibition to do x if agent **n** has the goal that x does not count as a violation under some condition. The permission overrides the prohibition if the goal that something does not count as a violation ($Y \wedge x \rightarrow \neg V(x, \mathbf{a})$) has higher priority in the ordering on goal and desire rules $\geq_{\mathbf{n}}$ with respect to the goal of a corresponding prohibition that x is considered as a violation ($Y' \wedge x \rightarrow V(x, \mathbf{a})$):

Definition 8 (Permission) *Agent $\mathbf{a} \in A$ is permitted by agent $\mathbf{n}$ to see to $x \in Lit(X_{\mathbf{a}} \cup P)$ under condition $Y \subseteq Lit(X_{\mathbf{a}} \cup P)$, written as $NMAS \models P_{\mathbf{an}}(x \mid Y)$, iff*

- $Y \cup \{x\} \rightarrow \neg V(x, \mathbf{a}) \in out(G_{\mathbf{n}})$*: if Y and x then agent $\mathbf{n}$ wants that x is not considered a violation by agent $\mathbf{a}$.*

In this paper we do not consider the problem of how the normative system is constructed by the sources of norms such as governments. See for example (Boella & van der Torre 2003f) for a discussion of the problem of the legal sources of norms.

Constitutive norms

Constitutive norms introduce new abstract classifications of existing facts and entities, called institutional facts, or they describe the legal consequences of actions on the normative system. According to Searle, institutional facts like marriage, money and private property emerge from an independent ontology of "brute" physical facts through constitutive rules of the form "such and such an X counts as Y in context C" where X is any object satisfying certain conditions and Y is a label that qualifies X as being something of an entirely new sort. Examples of constitutive rules are "X counts as a presiding official in a wedding ceremony", "this bit of paper counts as a five euro bill" and "this piece of land counts as somebody's private property".

We formalize the counts-as conditional as a belief rule of the normative agent $\mathbf{n}$. Since the condition x of the belief rule is a variable it can be an action of an agent, a brute fact or an institutional fact. So, the counts as relation can be iteratively applied. An additional condition of the counts-as conditional is that if it is triggered by an agent, then this agent must participate in the normative system.

Definition 9 (Counts-as relation) *Let* $NMAS = \langle A, R, \in, X, B, D, G, AD, MD, \geq, \mathbf{n}, I, V, GD\rangle$ *be a normative multiagent system. A literal* $x \in Lit(X)$ *counts-as* $y \in Lit(I)$ *in context* $C \subseteq Lit(X)$, $NMAS \models$ *counts-as*$(x, y|C)$*, iff:*

1. $C \cup \{x\} \rightarrow y \in out^{+}(B_{\mathbf{n}})$*: if agent* $\mathbf{n}$ *believes* C *and* x *then it believes* y.
2. *If* $x \in Lit(X_b)$*, then there is* $r \in R$ *such that* $\langle b, \mathbf{n}\rangle$ *in* $\in(r)$*: if the condition is a decision of an agent, then it must play a role in the normative system.*

The constitutive rules are illustrated by our example.

Example 5 (Continued) $B_{\mathbf{n}} = \{e_1\}$ *and* $MD(e_1) =$ *fence* $\rightarrow$ *property. Consequently we have* $NMAS \models$ *counts-as*(*fence*, *property*$|\top$)*, because we also have* $\in(member) = \{\langle \mathbf{a}, \mathbf{n}\rangle\}$*. This formalizes a society which believes that a field fenced by an agent who is member of the normative system counts as the fact that the field is a property of that agent. The presence of the fence is a physical "brute" fact, while being a property is an institutional fact. A regulative norm which forbids trespassing refers to the abstract concept of property rather than to fenced fields:* $O_{b\mathbf{n}}$(*trespass*, s | *property*)*. As the system evolves, new cases are added to the notion of property by means of new constitutive rules, without changing the regulative norms about property. E.g., if a field is inherited, then it is property of the heir: inherit* $\rightarrow$ *property* $\in MD(B_{\mathbf{n}})$.

From a knowledge representation point of view, constitutive norms behave as *data abstraction* in programming languages: types are gathered in new abstract data types; new procedures are defined on the abstract data types to manipulate them. So it is possible to change the implementation of the abstract data type without modifying the programs using those procedures. In our case, it is possible to change the constitutive norms defining the institutional facts without modifying the regulative norms which refer to those institutional facts.

Games

The games we consider in this paper are based on recursive modelling, in which an agent chooses an optimal decision by assuming that other agents make optimal decisions too. For example, in the running example agent $\mathbf{a}$ makes an optimal decision from its point of view, assuming that the normative agent thereafter makes an optimal decision from its point of view. We call the order of the agents making decisions the *protocol* of the game. When an agent imagines the decision of another agent, it must have a profile of the other agent's mental state. More precisely, if we consider a recursive model with protocol of n agents $b_1 \dots b_n$, then each agent b_i has to have a profile of each sequence of agents $b_{i+1} \dots b_n$.

agent b_1 deliberates about optimal decision

→ considers optimal decision of agent b_2

agent b_2 deliberates about optimal decision

→ considers optimal decision of agent b_3

agent b_3 deliberates about optimal decision

→ considers optimal decision of agent b_4

...

agent b_n deliberates

Figure 2: Recursive modelling

However, in this paper we have not defined agent profiles. Moreover, it is unrealistic to assume that agents have such detailed agent profiles. We therefore assume in our games that each agent has the same profile of the other agents. More precisely, we assume in our games that the mental state $AD(b_i)$ is the profile of agent b_i according to the other agents. There are two main complications:

- If an agent makes two decisions in the recursive modelling, then this assumption is unrealistic. For example, when the normative agent creates a new norm, considers the reaction of an agent, and thereafter may sanction the agent (Boella & van der Torre 2003f). In this paper we exclude this kind of games.
- If an agent can observe the effects of decisions of other agents, then the assumption is unrealistic. This problem normally does not occur with institutional facts, since they cannot be observed, but it occurs for brute facts. In this paper we assume that agents do not observe the effects of decisions of other agents, the only effects of decisions are derived by the agents' belief rules.

Moreover, to define games we have to consider how we define the effects of decisions (by applying belief rules), and how we evaluate the effects of the decisions. For the belief rules any kind of logic can be plugged into our framework, to keep our system simple we use a an input/output logic that does not deal with exceptions. Consequently, the rules are monotonic, there are no constraints, and there is no belief revision. See (Makinson & van der Torre 2001) for a discussion on these issues in the present setting, and an approach to introduce them.

However, the absence of exceptions in our logic of belief rules introduces the problem that it may be the case that an agent makes a decision, but then agents recursively modelled believe that the earlier decision is not possible, and they therefore cannot define a response to the first decision. In this paper we do not further consider this problem, but we simply exclude such games. There are several ways in which it can formally be forced that such situations do not occur, for example by assuming that if an agent recursively models another agent, then the belief rules of the former agent are a superset of the belief rules of the latter: since the former agent knows the latter agent's beliefs, it believes them too. Clearly this property only holds for a particular kind of beliefs (of the type usually identified with knowledge), but this is exactly the case of constitutive rules we are discussing. Constitutive rules do not concern reality, but they are established by the normative system, so they cannot be wrong.

A decision profile for a decision problem is a sequence of decisions, one for each agent. We thus do not consider simultaneous decisions. Agents evaluate states of affairs according to which motivational attitudes remain unfulfilled: their body is part of the expected effects of the decision, but their head is not.

Definition 10 *Let NMAS be a normative multiagent system.*

- *A protocol is a sequence of distinct agents $\langle b_1, \ldots b_n \rangle$. A decision problem $\langle nmas, protocol \rangle$ consists of a normative multiagent system and a protocol.*
- *A decision profile for a protocol is a sequence $\langle \delta_{b_1}, \ldots, \delta_{b_n} \rangle$ such that $cons(B_{b_i} \mid \delta)$ for $i = 1 \ldots n$ and $\delta = \cup_{i=1 \ldots n} \delta_{b_i}$. We also write Δ for the set of all decisions profiles.*
- *Agent b prefers a state of affairs $S_1 \subseteq Lit(X)$ to another one $S_2 \subseteq Lit(X)$ iff $U(S_2, b) >_b U(S_1, b)$, where $U(S, b) =$*

  $$\{m \in M_b \mid MD(m) = L \rightarrow l, L \subseteq S \text{ and } l \not\in S\}$$

The protocols are illustrated by our running example.

Example 6 (Continued) *Assume the protocol $\langle \mathbf{a}, \mathbf{n} \rangle$, in which first agent $\mathbf{a}$ takes a decision, and thereafter agent $\mathbf{n}$ reacts on it. The decision profile $\langle \{crop\}, \{s\} \rangle$ represents that first agent $\mathbf{a}$ cultivates crop, and thereafter agent $\mathbf{n}$ sanctions agent $\mathbf{a}$.*

The games the agents can play in this extended game theory are based on a recursive definition. Due to the fact that the protocol is finite, the definition is well founded.

Definition 11 *A decision profile δ_1 dominates decision profile δ_2 for agent b_i if they have the same set of decisions $\delta_{b_1} \ldots \delta_{b_i - 1}$, and for every decision profile agent δ'_1 and δ'_2 that coincide with δ_1 and δ_2 for b_1 to b_i and that are optimal for agent $b_{i+1} \ldots b_n$, b_i prefers $out^+(B_{b_i}, \delta'_1)$ to $out^+(B_{b_i}, \delta'_2)$, i.e., $U(out^+(B_{b_i}, \delta'_2), b_i) >_{b_i} U(out^+(B_{b_i}, \delta'_1), b_i)$.*

A decision profile is optimal for agent b_i if it is not dominated by another decision profile, and it is optimal for all agents b_j with $j > i$. A decision profile is optimal if it is optimal for agent b_1.

The games are introduced in our running example. The optimal decision for agent $\mathbf{a}$ is to build a fence and cultivate crop, since it fulfills all the agent's desires and goals, and thereafter the optimal decision of the normative agent is not to sanction.

Example 7 (Continued) *The decision profile $\langle \{crop, fence\}, \{s\} \rangle$ is optimal, but the decision profile $\langle \{crop\}, \{s\} \rangle$ is not.*

Modifying the normative system

Searle's analysis of constitutive rules has focused mainly on the attribution of a new functional status to entities, as, for examples, weddings, money, property. Searle's idea is that constitutive rules "create the possibility or define that activity". However, we believe that the role of constitutive rules is not limited to the creation of an activity and the construction of new abstract categories. Constitutive norms specify both the behavior of a system and the evolution of the system: the normative system $\mathbf{n}$ itself specifies by means of its belief rules how its beliefs, desires and goals can be changed, who can change them, and the limits of the possible changes depending on the role played by an agent. In this section we only define actions that create new beliefs, desires and goals; actions that modify or delete mental attitudes can be defined analogously. Technically, we have defined the beliefs, desires and goals as part of the structure of the normative multiagent system. To change these mental attitudes, we have to introduce a more general structure that also contains actions that change the normative structure.

Despite their name, create actions are modelled as institutional facts, since they belong to social reality. As institutional facts are parameters, they cannot be directly controlled by any agent. But the normative system itself, by means of constitutive rules, assigns the power of executing create actions to agents playing roles in it: agents execute actions which count as the execution of creation actions, and, thus, they change the normative system following the agent's decision.

Definition 12 *An extended normative multiagent system $ENMAS$ is a tuple*

$$\langle A, R, \in, X, B, D, G, AD, MD, \geq, \mathbf{n}, I, V, GD, C \rangle$$

that consists of a normative multiagent system

$$\langle A, R, \in, X, B, D, G, AD, MD, \geq, \mathbf{n}, I, V, GD \rangle$$

together with a set of actions that can change the normative system:

- *the belief create actions $C_{B_\mathbf{n}}$: Rul$(X) \rightarrow I$, is a mapping from belief rules to institutional fact, where $C_{B_\mathbf{n}}(r)$ stands for the creation of $m \in B_\mathbf{n}$, together with the update of MD such that $r = MD(m)$.*
- *the desire create actions $C_{D_\mathbf{n}}$: Rul$(X) \rightarrow I$, where $C_{D_\mathbf{n}}(r)$ stands for the creation of $m \in D_\mathbf{n}$, together with the update of MD such that $r = MD(m)$.*
- *the goal create actions $C_{G_\mathbf{n}}$: Rul$(X) \rightarrow I$, where $C_{G_N}(r)$ stands for the creation of $m \in G_\mathbf{n}$, together with the update of MD such that $r = MD(m)$.*
- *the goal create actions C_{GD} : Rul$(X) \times A \rightarrow I$, where $C_{G_N}(r, b)$ stands for the update of GD such that $m \in GD(b)$ and $r = MD(m)$.*

Games in ENMAS are defined analogously to games in NMAS, with the only exception that all references to the agents' mental states are made to the normative multiagent system that results after the normative multiagent system of the decision problem has been updated with the decisions.

Definition 13 *Let ENMAS be an extended normative multi-agent system. A protocol is defined in the same way as for NMAS. Moreover:*

- *A decision profile for a protocol is a sequence $\langle \delta_{b_1}, \ldots, \delta_{b_n} \rangle$ such that $cons(B'_{b_i} \mid \delta)$ for $i = 1 \ldots n$ and $\delta = \cup_{i=1 \ldots n} \delta_{b_i}$, where B'_{b_i} contains the beliefs from B_{b_i} together with the beliefs that occur in the create actions in $out^+(B_{b_i}, \delta)$.*
- *Agent b prefers a state of affairs $S_1 \subseteq Lit(X)$ to another one $S_2 \subseteq Lit(X)$ iff $U(S_2, b) >_b U(S_1, b)$, where $U(S, b) =$*

  $$\{m \in M'_b \mid MD(m) = L \rightarrow l, L \subseteq S \text{ and } l \notin S\}$$

  where M'_b contains the motivations from M_b together with the motivations that occur in the create actions in $out^+(B_b, \delta)$.

The dominance relation and optimal decisions are defined in the same way as for NMAS.

Since regulative norms are defined in terms of goals of the normative agent and constitutive norms in terms of its beliefs, by means of creation actions it is possible to create new regulative and constitutive rules. The following example illustrates the modification of the normative system by creating a constitutive norm.

Example 8 *An example of a constitutive norm for creating new constitutive rules in a normative agent is counts-as$(y, C_{B_\mathbf{n}}(x \rightarrow p) \mid \top)$ where $y \in X_b$ is an action of agent $b \in A$, $x \in X_c$ is an action of agent $c \in A$ and $p \in I$ is an institutional fact. In the normative system agent b must play, e.g., the role of legislator, who has the power to create norms; agent c is in a role that has the power to make the institutional fact p true by doing x.*

In the following section, we illustrate the modification of the normative system by the creation of a regulative norm in our running example. As discussed in (Boella & van der Torre 2003e) creating regulative norms is more complicated, because it does not make sense to create all the items of Definition 6. Some of the conditions of obligation are preconditions for norm creation, and others are postconditions of such an action.

Norm creation in the running example

In this section we conclude our example showing the other role played by constitutive rules: not only as abstractions for an institutional classification of reality, but also as specifications of the possible changes to the normative agent. Thus far, we have considered a game played by an agent $\mathbf{a}$ who, in order to cultivate some crop (*crop*), fences (*fence*) a field to show that it is its property (*property*), as requested by the obligation $O_{\mathbf{an}}(\textit{property}, s \mid \textit{crop})$ in Example 4. Fencing counts as being the owner *counts-as*$(\textit{fence}, \textit{property} \mid \top)$, as discussed in Example 5.

In the normative system $\textit{ENMAS}_1$ considered in this section, having a property creates an obligation that if the owner $\mathbf{a}$ of the field grows crop, the agent must be authorized ($autho$) by the land registry $O_{\mathbf{an}}(\textit{autho}, s' \mid \textit{crop})$. For example, the registry has to keep track of the crops for taxation purposes. To be registered it is sufficient to apply (*apply*), since the application counts as being authorized: *counts-as*$(\textit{apply}, \textit{autho} \mid \top)$. This means that $B_\mathbf{n} = \{e_1, e_2\}$ and $MD(e_1) = \textit{fence} \rightarrow \textit{property}$, $MD(e_2) = \textit{apply} \rightarrow \textit{autho}$, and $\in(\textit{member}) = \{\langle \mathbf{a}, \mathbf{n} \rangle\}$, see Example 3.

The normative agent $\mathbf{n}$ specifies how the normative system can be changed. In the example, it specifies how a new obligation $O_{\mathbf{an}}(autho, s' \mid crop)$ is created by means of create actions. Since this obligations is defined in terms of the goals and desires of the normative agent, the create actions must add these goals to the normative system. Note that the desire of agent $\mathbf{a}$ not to be sanctioned ($\textit{crop} \rightarrow \neg s'$) must be already true in $\textit{ENMAS}_1$ for the obligation to be defined correctly. Moreover, to create the obligation $\textit{ENMAS}_1$ must specify the following create actions:

$c_1 = C_{G_\mathbf{n}}(\textit{crop} \rightarrow \textit{autho})$
$c_2 = C_{D_\mathbf{n}}(\textit{crop} \rightarrow \textit{autho})$
$c_3 = C_{G_\mathbf{n}}(\textit{crop} \wedge \neg\textit{autho} \rightarrow V(\neg\textit{autho}, \mathbf{a}))$
$c_4 = C_{D_\mathbf{n}}(\textit{crop} \wedge \neg\textit{autho} \rightarrow V(\neg\textit{autho}, \mathbf{a}))$
$c_5 = C_{D_\mathbf{n}}(\textit{crop} \rightarrow \neg V(\neg\textit{autho}, \mathbf{a}))$
$c_6 = C_{G_\mathbf{n}}(\textit{crop} \wedge V(\neg\textit{autho}, \mathbf{a}) \rightarrow s')$
$c_7 = C_{D_\mathbf{n}}(\textit{crop} \wedge V(\neg\textit{autho}, \mathbf{a}) \rightarrow s')$
$c_8 = C_{D_\mathbf{n}}(\textit{crop} \rightarrow \neg s')$
$c_9 = C_{GD}(\textit{crop} \rightarrow \textit{autho}, \mathbf{a})$

Since these create actions are parameters in I, they are not directly controlled by the normative agent: it cannot perform them by itself. Rather the normative agent specifies who is able to do these changes by counts-as rules: it specifies who is able to execute those create actions by means of its decision variables. In our example, we have that *property* counts as the creation of the obligation: *counts-as*$(\textit{property}, c_i \mid \top)$ where $1 \leq i \leq 9$. Let us assume that these constitutive norms are based on $\overline{e_i} \in B_\mathbf{n}$ for $1 \leq i \leq 9$ with $MD(\overline{e_i}) = \textit{property} \rightarrow c_i$.

We consider now for the decision problem $\langle \textit{ENMAS}_1, \langle \mathbf{a}, \mathbf{n} \rangle \rangle$ the decision profile $\langle \textit{ENMAS}_1, \langle \delta_\mathbf{a}, \delta_\mathbf{n} \rangle \rangle$ The possible decisions of agent $\mathbf{a}$ are doing nothing, to fence the field, to cultivate crop, to apply for an authorization and all the possible combinations of these actions. The possible decisions of agent $\mathbf{n}$ are doing nothing, considering some violation, sanctioning and all the possible combinations of these actions. Agent $\mathbf{a}$ has to take a decision to fulfill its desire of cultivating crop ($MD(d_1) = \top \rightarrow \textit{crop}$ and $d_1 = MD(D_\mathbf{a})$). At the same time, it does not desire to be sanctioned with s and s' if it grows crop ($MD(d_2) = \top \rightarrow \neg s$ and $d_2 \in D_\mathbf{a}$, and $MD(d_3) = \top \rightarrow \neg s'$ and $d_3 \in D_\mathbf{a}$) by the normative agent $\mathbf{n}$, and the latter desires are stronger than the former ($\geq_\mathbf{a} \supseteq \{\{d_2\} \geq \{d_1\}, \{d_3\} \geq \{d_1\}\}$).

To achieve its desire to grow crop, it has also to take into account the consequences of this action. First, it knows that the normative agent will consider crop without property a violation ($crop \wedge \neg property \rightarrow V(\neg property, \mathbf{a}) \in out(G_{\mathbf{n}})$, from Example 4). For this reason it has also to fence the field. Second, it knows that the existence of a property makes the normative agent $\mathbf{a}$ create a new obligation: that it gets an authorization to grow crop. Hence, agent $\mathbf{a}$ has not only to consider the effects of its behavior, but also to consider that the second agent in the protocol, when it will act, could be in a different normative system, due to the effects of agent $\mathbf{a}$'s actions.

The optimal decision of agent $\mathbf{a}$ has to take into account which is the reaction of agent $\mathbf{n}$. Hence, agent $\mathbf{a}$ has to consider not the state immediately following its decision $\delta_{\mathbf{a}}$, rather the final state. So, even if the decision $\{\mathit{crop}\}$ satisfies all its desire in state $\top \rightarrow crop$, it is not the optimal decision, since in the subsequent state the effect includes the sanctions s and s' as a result of the normative agent's decision that the obligation $O_{\mathbf{an}}(autho, s \mid crop)$ has been violated. The optimal decision is, instead, $\{\mathit{fence}, \mathit{crop}, \mathit{apply}\}$: agent $\mathbf{a}$ knows that fencing counts as property and applying for authorization counts as being authorized for agent $\mathbf{n}$, hence, both the existing obligation and newly created obligation of *ENMAS*$_2$ is not violated.

$enmas_1 =$
$\langle A, R \in, X, B, D, G, AD, MD, \geq, \mathbf{n}, I, V, GD, C\rangle$
$A = \{\mathbf{a}, \mathbf{n}\}, R = \{\mathit{member}\}, X_{\mathbf{a}} = \{\mathit{fence}, \mathit{crop}, \mathit{apply}\},$
$B_{\mathbf{a}} = B_{\mathbf{n}}, G_{\mathbf{a}} = \emptyset, D_{\mathbf{a}} = \{d_1, d_2, d_3\}$

$X_{\mathbf{n}} = \{V(\neg \mathit{property}, \mathbf{a}), s, V(\neg \mathit{autho}, \mathbf{a}), s'\}$
$B_{\mathbf{n}} = \{e_1, e_2, \overline{e_1}, \overline{e_1}, \ldots, \overline{e_9}\}, G_{\mathbf{n}} = \{g_1, g_2, g_4\},$
$D_{\mathbf{n}} = \{g_1, \ldots, g_5\}, GD(\mathbf{a}) = \{g_1\}$

$I = \{\mathit{property}, \mathit{autho}, c_1, \ldots, c_9\}$

$\delta_{\mathbf{a}} = \{\mathit{fence}, \mathit{crop}, \mathit{apply}\}$
$F_1 = out^{+}(B_{\mathbf{a}}, \delta_{\mathbf{a}} = \{\mathit{fence}, \mathit{crop}, \mathit{apply}\}) = \{\mathit{fence}, \mathit{crop}, \mathit{apply}, \mathit{property}, \mathit{autho}, c_1, \ldots, c_9\}$
$F_1 \cap I = \{\mathit{property}, \mathit{autho}, c_1, \ldots, c_9\}$

$enmas_2 =$
$\langle A, R, \in, X, B, D', G', AD, MD, \geq, \mathbf{n}, I, V, GD', C\rangle$
$B'_{\mathbf{n}} = \{e_1, e_2, \overline{e_1}, \overline{e_1}, \ldots, \overline{e_9}\},$
$G'_{\mathbf{n}} = \{g_1, g_2, g_4, g_6, g_7, g_9\}, D'_{\mathbf{n}} = \{g_1, \ldots, g_{10}\},$
$GD'(\mathbf{a}) = \{g_1, g_6\}$

$MD(g_6) = \mathit{crop} \rightarrow \mathit{autho}$
$MD(g_7) = \mathit{crop} \wedge \neg \mathit{autho} \rightarrow V(\neg \mathit{autho}, \mathbf{a})$
$MD(g_8) = \top \rightarrow \neg V(\neg \mathit{autho}, \mathbf{a})$
$MD(g_9) = \mathit{crop} \wedge V(\neg \mathit{autho}, \mathbf{a}) \rightarrow s'$
$MD(g_{10}) = \mathit{crop} \rightarrow \neg s'$

$\delta_{\mathbf{n}} = \emptyset$
$F_2 = out^{+}(B'_{\mathbf{n}}, \delta_{\mathbf{a}} = \{\mathit{fence}, \mathit{crop}, \mathit{apply}\} \cup \delta_{\mathbf{n}}) = \{\mathit{fence}, \mathit{crop}, \mathit{apply}, \mathit{property}, \mathit{autho}, c_1, \ldots, c_9\}$

$U(F_1, \mathbf{a}) = \emptyset$
$U(F_2, \mathbf{n}) = \emptyset$

Regulative or constitutive norms?

One relevant problem in encoding norms is whether to use many regulative norms and a few constitutive norms, or a few regulative norms and many constitutive norms. In our framework, the question is whether to use many goals and a few beliefs, or a few goals and many beliefs. Interestingly, a similar trade-off can be found in knowledge-based systems. Traditional planning systems are based on a single goal, but modern agent systems typically contain many goals and a goal selection mechanism. Other inspirations for this trade-off can be found in legal theory. Traditionally, law scholars like Hart (Hart 1961) distinguish between primary laws, whose purpose is to direct the behavior of citizens, and secondary laws, which, among other functions, serve to the maintenance and dynamic management of the normative system. These rules form a "subsystem of rules for change" (Biagioli 1997): rules which have juridical effects and which are instrumental to the primary system, in that they regulate the regulation (e.g., art. 2 of Italian Civil Code: "the creation of laws [...] is regulated by constitutional laws" Cost. 70). This subsystem, according to Hart, does not include only the rules of change which specify how new laws are introduced or old ones removed, but it also includes rules about "powers for the private citizen". These rules are at the basis of civil code and allow testaments and contracts; for Hart they allow the exercise of limited legislative powers by the citizens. These rules do not create or remove general laws but they introduce and remove individual obligations and permissions: e.g., in the Italian Civil Code art. 1173 (sources of obligations) specifies that obligations are created by contracts (a contract being an agreement among two or more parties to regulate a juridical relationship about valuables by art. 1321).

The normative agent metaphor allows also using our framework to support legislative drafting. One of the issues in writing laws is that also regulative norms can be expressed by means of assertions, i.e. in the same way as constitutive norms, rather than by means of deontic statements concerning what is obligatory or permitted: e.g. "murder is punished with ten years of jail." This sentence does not describe a constitutive norm. It is a description of the fact that murder is prohibited and murderers are sanctioned

The normative system as agent metaphor provides us with a criterion for distinguishing the two types of norms. If an assertion refers to actions performed by the normative agent, the norm is a regulative one: it is a description of the behavior of the normative agent, and, since agent behavior is described in terms goals, what the sentence really describes is the goal of the agent: in case of murder the normative agent has the goal of sanctioning the murderer with a ten year term. This is the goal contained in the fourth clause of the definition of obligation.

In contrast, assertions describing facts are constitutive norms. For example the sentence "a contract is an agreement among two or more parties" (art. 1321 of Italian Civil Code) is not a regulative norm in that it does not describe an action of the normative agent: it describes the belief of the normative agent that an agreement (a brute fact among agents) is considered as an institutional fact which does not exist without the normative system: a contract.

Analogously for constitutive norms concerning norm changing. For example "obligations derive by contract" (*ibid.* art. 1372): the sentence does not refer to any action of the normative system. Rather it specifies that the goals of the normative agents which define an obligation are modified by a contract. I.e., the sentence describes the beliefs of the normative agent about the consequences of a contract made among agents.

Related work

This work is part of a wider research on normative reasoning of autonomous agents based on the attribution of mental attitudes to the normative system (Boella & van der Torre 2003a). In (Boella & van der Torre 2004) we consider the social delegation of goals to the normative system, in (Boella & van der Torre 2003f) we introduce permissions, in (Boella & van der Torre 2003d) the definition of the role of a defender which fulfills the task of identifying violations and sanctioning them on behalf of the normative system. In (Boella & van der Torre 2003c) we apply the framework to the regulation of virtual communities of agents based on the grid infrastructure. Finally, in (Boella & van der Torre 2003b) we explore how to formalize the *trias politica* using the standard BDI$_{CTL}$ logic (Rao & Georgeff 1998) for agent verification.

Other related work is (Lopez y Lopez, Luck, & d'Inverno 2002). They propose a model of obligation compliance of agents that emphasizes their autonomy. They classify different motivations for which agents decide to stick to obligations or to violate them. In (Lopez y Lopez, Luck, & d'Inverno 2001), the same authors stress the importance of having a model of the other agents in order to reason about the dependance relations with them. But, while, as in our work, in (Lopez y Lopez, Luck, & d'Inverno 2002) sanctions are associated with the definition of an obligation, there is no agent in charge of sanctioning violations since "the application of punishments and rewards is taken for granted". So they are not able to model unpunished violations when the addressee of the obligation exploits the recursive modelling of the normative system's decisions.

Also (Dignum *et al.* 2000) are interested in integrating norms and obligations in a BDI approach to multiagent systems. We focus on obligations since in their model obligations and not norms are intended as associated with an explicit agent which is responsible for enforcing the penalty. In doing so they adopted van der Torre and Tan's approach (1999) on preference based dyadic deontic logic built on Kripke models. On the contrary our approach is entirely based on input/output logic.

As in (Castelfranchi *et al.* 2000) we consider norms as "mental objects entering the mental processing" which interact with beliefs, goals and decisions. Moreover, they also claim that norms cannot be hard constraints, but they just can influence an agent to a certain behavior if they directly or indirectly satisfy some of his goals.

In Artificial Intelligence, the modelling of counts-as relations has been introduced by Jones and Sergot (1996). We depart from their model in that constitutive norms are not modelled as operative constraints of an institution but as beliefs of the normative agent. We distinguish also from the subsequent work of Sadighi Firozabadi and Sergot (1999) who propose a model for power and permission in security policies; in our work, powers are defined in terms of how the normative system allows agents which have a role in the normative system to change its beliefs and motivations.

What distinguishes our approach from other models of counts-as relations is that we can connect goals, and obligations defined as goals, to institutional facts inside the overall frame of the attribution of the status of agent to the normative system: institutional facts are beliefs of the normative agent as any other belief. Here, we take full advantage of the metaphor, as also Tuomela (1995) argues:

> The notions of goal, belief, and action are linked in the case of a group to approximately the same degree as in the individual case. In the latter case their interconnection is well established; given that the person-analogy applies to groups [. . .], these notions apply to groups as well.

The ontology presented by (Gangemi, Sagri, & Tiscornia 2003) adopts a perspective similar to our view of constitutive rules as beliefs: institutional facts are considered human facts depending on consciousness (and not on will) and are legally constituted by (satisfying) constitutive rules.

Conclusions

In this paper we introduce a formal framework for normative multiagent systems including regulative and constitutive norms. We model constitutive and regulative norms as conditional rules representing, respectively, the beliefs and goals of the normative system, i.e., by attributing mental attitudes to normative systems.

We show how to reason about modifications of the normative system in this framework. Constitutive norms play not only the role of creating new abstract categories which compose the institutional reality, but they also specify how the normative system evolves over time by introducing or removing norms and obligations. Roles are used to specify the powers of agents to create institutional facts or to modify the norms and obligations of the normative system.

We also show how to play games in this normative multiagent system, including violations of obligations. The games are based on recursive modelling, concerning, e.g., the decision to fulfill or violate a norm, and the decision of which norms to create in order to achieve the desired social order.

An issue for further research is a formal analysis of the balance between regulative and constitutive norms. We would like to derive guidelines in which circumstances a normative system should create mainly regulative norms, and in which circumstances it should create mainly constitutive norms. Moreover, we would like to formally characterize the set of normative systems that can be generated by a normative system, given the rules it contains to modify itself.

Another issue for future research is an analysis of the properties of the counts-as relation based on the properties of the belief rules. For example, an ordering on beliefs can be used to achieve non-monotonicity of beliefs and thus of counts-as, as required by (Gelati *et al.* 2002). It is possible that under some circumstances a certain state of affairs does not count as something else. For example, it is not possible to fence a public garden to make it ones property.

Moreover, we like to study the application of our framework in legal reasoning. Constitutive rules are at the basis of legal institutions: "systems of [regulative and constitutive] rules that provide frameworks for social action within larger rule-governed settings" (Ruiter 1997).

References

Anderson, A. R. 1958. The logic of norms. *Logic et analyse* 2.

Biagioli, C. 1997. Towards A legal rules functional micro-ontology. In *Procs. of 1st LegOnt Workshop on Legal Ontologies*.

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

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

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

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

Boella, G., and van der Torre, L. 2003d. Norm governed multiagent systems: The delegation of control to autonomous agents. In *Procs. of IEEE/WIC Intelligent Agent Technology Conference*, 329– 335. IEEE Press.

Boella, G., and van der Torre, L. 2003e. Permissions and obligations in hierarchical normative systems. In *Procs. of ICAIL'03*, 81–82. ACM Press.

Boella, G., and van der Torre, L. 2003f. Rational norm creation: Attributing mental attitudes to normative systems, part 2. In *Proceedings of ICAIL'03*, 109–118. ACM Press.

Boella, G., and van der Torre, L. 2004. Δ: The social delegation cycle. In *Procs. of DEON'04 Workshop*.

Breuker, J.; Valente, A.; and Winkels, R. 1997. Legal ontologes: A functional view. In *Procs. of 1st LegOnt Workshop on Legal Ontologies*, 23–36.

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

Castelfranchi, C.; Dignum, F.; Jonker, C. M.; and Treur, J. 2000. Deliberate normative agents: Principles and architecture. In *Intelligent Agents VI (ATAL-99)*. Springer-Verlag. 364–378.

Castelfranchi, C. 2000. Engineering social order. In *Procs. of ESAW'00*, 1–18. Springer Verlag.

Dellarocas, C. 2001. Negotiated shared context and social control in open multi-agent systems. In Conte, R., and Dellarocas, C., eds., *Social Order in MAS*. Kluwer.

Dennett, D. C. 1987. *The Intentional Stance*. Cambridge, MA: The MIT Press.

Dignum, F.; Morley, D.; Sonenberg, E. A.; and Cavedon, L. 2000. Towards socially sophisticated BDI agents. In *Procs. of ICMAS'00*, 111–118. IEEE press.

Gangemi, A.; Sagri, M.; and Tiscornia, D. 2003. Metadata for content description in legal information. In *Procs. of LegOnt Workshop on Legal Ontologies*.

Gelati, J.; Governatori, G.; Rotolo, N.; and Sartor, G. 2002. Declarative power, representation, and mandate. A formal analysis. In *Procs. of JURIX 02*. IOS press.

Hart, H. L. A. 1961. *The Concept of Law*. Oxford: Clarendon Press.

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

Jones, A., and Sergot, M. 1996. A formal characterisation of institutionalised power. *Journal of IGPL* 3:427–443.

Lang, J.; van der Torre, L.; and Weydert, E. 2002. Utilitarian desires. *Autonomous Agents and Multi-agent Systems* 329–363.

Lopez y Lopez, F.; Luck, M.; and d'Inverno, M. 2001. A framework for norm-based inter-agent dependence. In *Procs. of 13rd Mexican International Conference on Computer Science*, 31–40.

Lopez y Lopez, F.; Luck, M.; and d'Inverno, M. 2002. Contraining autonomy through norms. In *Procs. of AAMAS'02*, 674–681. ACM press.

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

Makinson, D., and van der Torre, L. 2001. Constraints for input-output logics. *Journal of Philosophical Logic* 30(2):155–185.

Neal, S.; Cole, J.; Linington, P. F.; Milosevic, Z.; Gibson, S.; and Kulkarni, S. 2003. Identifying requirements for business contract language: a monitoring perspective. In *Procs. of EDOC'03*, 50–62. IEEE press.

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

Ruiter, D. 1997. A basic classification of legal institutions. *Ratio Juris* 10(4):357–371.

Sadighi Firozabadi, B., and Sergot, M. 1999. Power and permission in security systems. In *Security Protocols*. Springer Verlag. 48–53.

Searle, J. 1969. *Speech Acts: an Essay in the Philosophy of Language*. Cambridge, England: Cambridge University Press.

Searle, J. 1995. *The Construction of Social Reality*. New York: The Free Press.

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

van der Torre, L., and Tan, Y. 1999. Contrary-to-duty reasoning with preference-based dyadic obligations. *Annals of Mathematics and Artificial Intelligence* 27:49–78.