Attributing Mental Attitudes to Social Entities: Constitutive Rules are Beliefs, Regulative Rules are Goals

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

[1] Dipartimento di Informatica - Università di Torino- Italy. E-mail: guido@di.unito.it
[2] CWI Amsterdam and TU Delft - The Netherlands. E-mail: torre@cwi.nl

Abstract. In this paper, we propose a model of constitutive and regulative norms in a logical multiagent framework. We analyze the relationship between these two types of rules and explain similarities between them, using the metaphor of considering social entities - like normative systems, groups and organizations - as agents and of attributing them mental attitudes as well as an autonomous behavior. We argue that while constitutive norms expressing "counts-as" relations are modelled as the beliefs of social entities, regulative norms, like obligations, prohibitions and permissions, are modelled as their goals.

1 Introduction

Searle argues that there is a distinction between two types of norms:

> 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. Chess has no existence apart from 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).

According to Searle, institutional facts like marriage, money and private property emerge from an independent ontology of "brute" natural facts through constitutive norms 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 norms 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".

While the formalization of regulative norms, like obligations, prohibitions and permissions, is based in deontic logic on modal operators representing what

is obligatory, forbidden or permitted, the formalization of constitutive norms is rather different. An attempt to make the notion of constitutive norm more precise is [Jones and Sergot, 1996]'s formalization of the counts-as relation. For Jones and Sergot, the counts-as relation expresses the fact that a state of affairs or an action of an agent "is a sufficient condition to guarantee that the institution creates some (usually normative) state of affairs". As Jones and Sergot suggest, this relation can be considered as "constraints of (operative in) [an] institution", and express these constraints as conditionals embedded in a modal operator.

This view of the counts-as relation, however, leaves the explanation of regulative and constitutive rules totally unrelated. At first sight, the heterogeneity of these two meanings of the terms "rule" and "norm" can just be explained by the polysemy of the terms. It would be useful, however, to have a conceptual framework which explains the similarities and differences between the two concepts of regulative and constitutive rules. In this search for a common conceptual framework we are inspired by [Lakoff and Johnson, 1980]'s analysis of metaphorical reasoning, which, they argue, explains the different meanings for the terms by identifying the conceptual frame underlying them. Metaphors, as Lakoff and Johnson argue, are not only a form of figurative use of language, but they are at the basis of the cognitive ability of humans. Our minds use metaphors to understand and reason about concepts which we have no direct bodily experience of. For example, the domain of time is conceptualized and talked about by means of spatial notions and expressions: sentences like "the deadline is coming too fast" or "we are going towards the deadline without having finished this paper" witness that time is conceptualized in terms of mono-dimensional space (the time line) and instants of time are spaces (point) moving towards us or towards which we travel. In the "time-as-space" metaphor, space is the *source* domain which is mapped to the *target* domain of time: the first is better known to us so that we can attribute its properties to the less known domain of time. In this way they explain the common denominator of the different meanings of "come" and "go" as spatial and temporal verbs.

Social reality, which normative systems, groups and organizations belong to, is a complex phenomenon not directly accessible, so it is plausible that it is necessary to find a suitable source domain to explain, talk and reason about it.

In this paper, we address the following research questions:

- Which is the underlying the conceptual frame in which regulative and constitutive norms are defined?
- How these two concepts should be formally defined under the light of a metaphorical interpretation of social entities?

We address these questions using the logical multiagent framework for normative systems presented in [Boella and van der Torre, 2004e], [Boella and van der Torre, 2004g]. The basic assumptions of our model are that beliefs, goals and desires of an agent are represented by conditional rules and, that, when an agent takes a decision, it recursively models ([Gmytrasiewicz and Durfee, 1995]) the other agents interfering with it in order to predict their reaction to its decision.

2 Attributing mental attitudes to social entities

First of all, to proceed in our analysis, we must identify a suitable source domain for the metaphorical mapping whose properties allow to understand the target domain, in our case, an entity of the social reality like a normative system.

There is a candidate: the notion of human agenthood. At least two scholars support the use of agents as the source domain of a metaphorical mapping.

According to [Dennett, 1987], attitudes like belief and desire are folk psychology concepts that can be fruitfully used in explanations of rational human behavior. For an explanation of behavior it does not matter whether an entity actually possesses these mental attitudes: we describe the behavior of an affectionate cat or an unwilling screw in terms of mental attitudes. Dennett calls treating a person or artifact as a rational agent the intentional stance.

> Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do.

This metaphor has pervaded Artificial Intelligent in the last decades and has lead also to the construction of a new programming language paradigm in computer science ([van der Hoek et al., 1999]). We apply this metaphor also to social entities, like groups, institutions, and, in particular, normative systems.

The second scholar supporting the agent metaphor is [Tuomela, 1995] with his analysis of collectives like groups, institutions and organizations:

> The possibility of ascribing goals, beliefs, and actions to collectives relies on the idea that collectives can be taken to resemble persons. [...] Following common-sense examples, I will accept [...] that both factual and normative beliefs can be ascribed (somewhat metaphorically) to groups, both formal and informal, structured and unstructured.

For Tuomela, in the case of formal and organized collectives such as corporations the statutes, by-laws and other relevant rules of the collective can be shown to connect goals (interests, purposes, and whatever subtypes of goals are at stake), beliefs (or views), and actions.

2.1 Attributing goals

In a similar way in [Boella and van der Torre, 2003a], [Boella and van der Torre, 2004g], [Boella and van der Torre, 2004e] we use the metaphor of attributing mental attitudes to normative systems in order to explain normative reasoning in autonomous agents. The normative system is considered as an agent playing a game with the bearer of the norms. Henceforth, we will call it the normative agent.

We start with a well known definition.

> *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 behaviour may at times deviate from the ideal, i.e. that violations of obligations, or of agents rights, may occur [Jones and Carmo, 2001].

This definition of Carmo and Jones does not seem to require that the normative system is autonomous, or that its behavior is driven by beliefs and desires.

Our motivation for using the agent metaphor in [Boella and van der Torre, 2003a] is inspired by the interpretation of normative *multiagent* systems as dynamic social orders. 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". These interests can be a shared goal, a value that is good for everybody or for most of the members; for example, the interest may be to avoid accidents. We say that agents attribute the mental attitude 'goal' to the normative system, because all or some of the agents have socially delegated goals to the normative system [Boella and van der Torre, 2004c]; these goals are the content of the obligations regulating it, we will call them *normative goals*.

Moreover, social order requires *social control*, "an incessant local (micro) activity of its units" [Castelfranchi, 2000], aimed at restoring the regularities prescribed by norms. Thus, the agents attribute to the normative system, besides goals, also the ability to autonomously enforce the conformity of the agents to the norms, because a dynamic social order requires a continuous activity for ensuring that the normative system's goals are achieved. To achieve the normative goal the normative system forms the subgoal to consider as a violation the behavior not conform to it and to sanction violations.

So, a normative system has all the properties requested by [Wooldridge and Jennings, 1995] for being an agent: the autonomy - it has control over its actions and internal state - the social ability to interact with other agents, the reactivity to the changes it observes in the environment, and the pro-activeness - it has a goal directed behavior and takes the initiative.

2.2 Attributing beliefs

In this paper, we attribute also beliefs to a normative agent. Its beliefs are distinguished, as in any other agent, between the *beliefs about the state of the world* (a set of atomic propositions) and the *belief rules* an agent uses to draw conclusions starting from the believed state of the world. The beliefs about the state of the world are distinguished in two categories:

1. What Searle calls "brute facts": natural facts and events produced by the actions of the agents that can be observed by them.
2. "Institutional facts": a legal classification of brute facts; they belong only to the beliefs of the normative agent and have no direct counterpart in the world.

How institutional facts are created? Belief rules connect beliefs representing the state of the world to other beliefs which are their consequences. They have a conditional character and are represented in the same rule based formalism as goals and desires. Some of the belief rules have as consequences not other beliefs of the worlds (e.g., if it is raining, it is cold), but new legal, institutional facts whose existence is related only to the normative system. These belief rules, moreover, can connect also institutional facts to other institutional facts.

This type of belief rules express the *counts-as* relations which are at the basis of constitutive norms. It is important that belief rules have a conditional character, since they must reflect the conditional nature of the counts-as relation as proposed by Searle: "such and such an X counts as Y in context C". For example, consider a society with the belief that a field has been fenced by an agent has as a consequence that the normative system believes that the field is property of that agent. The fence is a physical "brute" fact, while the fact that it is a property of someone is only an institutional fact attributed to the beliefs of the normative system.

A fact p counts as an institutional fact q for normative agent a_j $\mathit{counts\text{-}as}_j(p, q)$, iff:

– agent a_j believes that p has q as a consequence.

Why are constitutive rules needed in a normative system? First, 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 et al., 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.

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 defined as beliefs inside the overall frame of the attribution of the state of agent to the normative system. 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.

If the normative system can be considered as an agent, then it takes its decisions on the basis of which of its conditional goals which applicable according to its beliefs are satisfied and which remain unsatisfied.

Continuing the example above, assume that the normative agent has as goals that if a field is a property, no one enters it and that if a property is entered, this action is considered as a violation. These two goals, which are part of an obligation not to trespass property, have among their conditions the fact that the field is a property: the field being a property is an institutional fact believed by the normative agent, while entering the field is a brute fact.

The possibility that institutional facts appear as conditions in the goals of the normative system or as goals themselves explains the following puzzling assertion of [Searle, 1995]: "constitutive rules constitute (and also regulate) an activity the existence of which is logically dependent on the rules" (p.34). How can constitutive rules *regulate* an activity, if this is the role played by regulative rules? E.g., [Hindriks, 2002] argues that constitutive rules consist of also regulative ones.

In contrast our model explains this assertion in a straightforward way: constitutive rules regulate a social activity since they create institutional facts that are conditions or objects of regulative rules. In our metaphorical mapping regulative rules are goals, and goals base their applicability in a certain situation on the beliefs of the agent; in the previous example, being a property indirectly regulates the behavior of agents, since entering a field is a violation only if it is a property; if a field is not a property, the goal of considering trespassing a violation does not apply to it.

[Searle, 1995] interprets the creation institutional facts also in terms of what he calls "status functions": "the form of the assignment of the new status function can be represented by the formula 'X counts as Y in C'. This formula gives us a powerful tool for understanding the form of the creation of the institutional fact, because the form of the collective intentionality is to impose that status and its function, specified by the Y term, on some phenomenon named by the X term", p.46. Where "the ascription of function ascribes *the use to which we intentionally put* these objects" (p.20). In our model, this teleological aspect of the notion of function depends again on the fact that institutional facts make conditional goals relevant as they appear in the conditions of regulative norms or as goals themselves. The end of fencing a field is to prevent trespassing by considering it a violation: the obligation is part of the function of property.

Hence, Searle's assertion that "the institutions [...] are systems of such constitutive rules" is partial: institutions are systems where constitutive (i.e., beliefs) and regulative (i.e., goals) rules interacts as goals and beliefs do in agents.

Searle claims that "the creation of a status-function is a matter of conferring some new power" [Searle, 1995, p.95]. The kind of power at stake, for Searle, is conventional or deontic power, which can be negative (for regulative rules) or positive (for constitutive rules). Positive deontic power is a matter of enabling someone to do something and comes with a right to do it. Negative deontic power consists of a requirement to do something, which comes with an obligation to do it.

We believe, as [Artosi, 2002] does, that this is a weak point in Searle's discussion. The status function is indeed a matter of conferring a new power, but it has nothing to do with rights. As [Makinson, 1986] noted it is possible to have the power to create an institutional fact without having the right or permission to do that: a priest can have the power to marry people without the permission or right to do so. We interpret the creation of institutional facts as a form of power, but in the sense of [Castelfranchi, 2003]'s *power-of*: an agent has the ability and is in the condition of making true some fact which is among its goals. Differently from the individual power in this form of institutional power the agent depends on the other agents: "the imposition of a status function [...] has to be collectively recognized and accepted or the function will not be performed." As discussed in the example below, in our model normative systems work only as far as they are collectively accepted by all the agents.

What Searle calls negative power, i.e., obligations, in our model has nothing to do with positive one. Obligations are the goals of the normative system. They have not necessary a "negative" connotation, since only the collective respect by a society of the goals of the normative system - which are the content of the obligations - allows the society to achieve its ends. As again [Castelfranchi, 2003] notices, only in the context of groups and institutions an agent can overcome the limitations of its individual power.

The attribution of mental attitudes goes in parallel with the attribution to the normative agent of a behavior directed by its beliefs and goals. When an agent has to take a decision and it is subject to some norm, he must consider in the outcome of its decision also the possibility that its actions are considered as institutional facts and that its behavior is considered as a violation and, thus, that it will be sanctioned. These possibilities can be predicted by recursively modelling (using [Gmytrasiewicz and Durfee, 1995]'s terminology) the normative agent: i.e., considering which decision the normative agent will take as a reaction to the agent's decision. As any other agent, the normative agent is assumed to take a decision basing on its beliefs, desires and goals. Its decision is performed indirectly by the agents composing the normative system, e.g., judges and policemen. In [Boella and van der Torre, 2003b] we address the issue of how the behavior of these roles is specified.

It is not sufficient, however, that a single agent attributes mental attitudes to the normative system. It is necessary, as argued by [Tuomela, 2000], that the attribution of mental states is commonly accepted by the members of the society. In our running example about property, which is the point of fencing the field? An agent will fence the field only if it believes that the other members of the society attribute to the normative system the same goals and beliefs which represent the two norms: only in this case, the agent believes that fencing the property achieves the goal of keeping other agents off the field. When the agent recursively models the other agents, it knows that they will, in turn, recursively model the normative agent. If they do so, they come to the conclusion that the normative system believes the institutional fact that the field is property of that agent; moreover, they come to the conclusion that the normative system will punish their trespassing of the field.

In summary, the situation is depicted in Figure 1. We have a world composed of agents (boxes 1 and 2) and brute facts (the box f). Agents observe facts and, basing on their observations and of the consequences of their observations, defined by the belief rules (oval B), they have beliefs about the state of the world (s). They control some of these facts of the world by means of their actions (d); decisions about which actions to perform are taken on the basis of their goals and desires (expressed by the goal and desire rules in ovals G and D).

If there were only agents of this sort, there would be no groups, organizations, or normative systems. The social ability of agents depend on the fact that they (collectively) attribute to social entities the status of agents (box 3). In the specific case of normative systems, the agents attribute to them regulative and constitutive norms: institutional facts are just beliefs of the normative system (i). As concerns regulative norms, this is possible since obligations are defined in terms of goal rules attributed to the normative agent: goals regarding what is the ideal situation, what is considered as a violation (oval V), and how violations must be sanctioned. As concerns constitutive norms, this is possible since they are defined in terms of belief rules expressing counts-as relations between brute or institutional facts and new institutional facts.

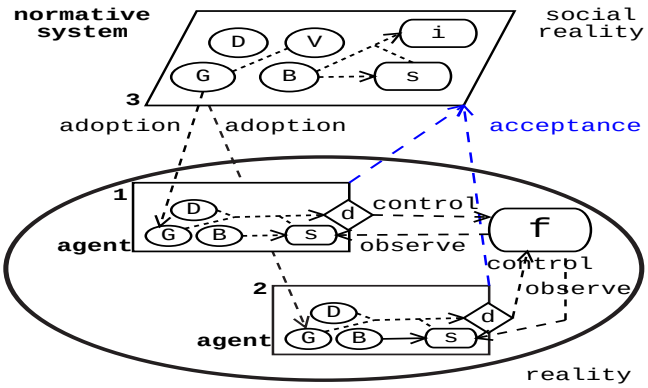

Fig. 1. The normative system as an agent.

3 The structure of norm governed social entities

The conceptual model of a social entity like a normative multiagent system is visualized in Figure 2, 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), $\square$ is a concept or set, — and $\rightarrow$ are associations between concepts, and $\longrightarrow\!\triangleright$ is the "is-a" or subset relation. The logical structure of the associations is detailed in the definitions below.

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 et al., 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.

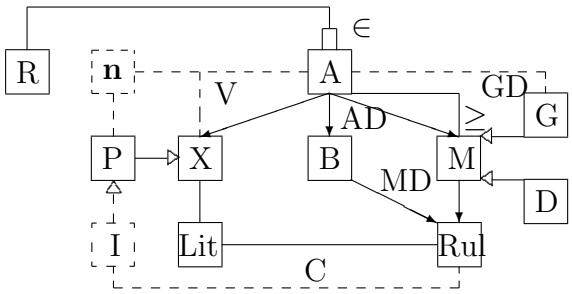

Fig. 2. Conceptual model of a normative system.

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* $AD : A \rightarrow 2^{X \cup B \cup M}$ *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* $\geq$*:* $A \rightarrow 2^M \times 2^M$ *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 **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 **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 **n**, and a role called *member*. Agent **a** is a member of normative system **n**.

$X_{\mathbf{n}} = \{s\}$, $P = \{property\}$, $D_{\mathbf{n}} = G_{\mathbf{n}} = \{g_1\}$, $MD(g_1) = \{crop \rightarrow property\}$, $GD(\mathbf{a}) = \{g_1\}$. Agent **n** can sanction agent **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 **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 and van der Torre, 2000,Makinson and van der Torre, 2001]. A rule set is a set of ordered pairs $p \rightarrow q$. 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 $\{p_1 \rightarrow q_1, \ldots, p_n \rightarrow q_n\}$, 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:*

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

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

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* $L \rightarrow l \in out(S)$*, and* $l \in out^+(S, L)$ *iff* $L \rightarrow l \in out^+(S)$*.*

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 and 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 and van der Torre, 2000], [Makinson and van der Torre, 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.

3.1 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).
$MD(g_2) = \{crop, \sim property\} \rightarrow V(\sim property, \mathbf{a})$
$MD(g_3) = \top \rightarrow \neg V(\sim property, \mathbf{a})$
$MD(g_4) = \{crop, V(\sim property, \mathbf{a})\} \rightarrow s$
$MD(g_5) = crop \rightarrow \sim s$
$\{g_1, g_2, g_4\} = G_{\mathbf{n}}$, $G_{\mathbf{n}} \cup \{g_3, g_5\} = D_{\mathbf{n}}$, $\{g_1\} = GD(\mathbf{a})$

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

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 $\mathbf{n}$ 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 and van der Torre, 2003c] for a discussion of the problem of the legal sources of norms.

3.2 Constitutive norms

Constitutive norms introduce new abstract classifications of existing facts and entities, called institutional facts. 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 asymmetry in the definition of counts-as with respect to the definition of obligation is only due to the fact that the other items are subgoals aiming at the enforcement of the obligation: they are subgoals of the main normative goal in item 1 which represents what is the ideal situation for the normative system.

The constitutive rules are illustrated by our example.

Example 5 (Continued).
$B_{\mathbf{n}} = \{e_1\}$ and $MD(e_1) = \mathit{fence} \rightarrow \mathit{property}$. Consequently we have $NMAS \models \mathit{counts\text{-}as}(\mathit{fence}, \mathit{property} | \top)$, because we also have $\in (\mathit{member}) = \{\langle \mathbf{a}, \mathbf{n} \rangle\}$. 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: $O_{b\mathbf{n}}(\mathit{trespass}, s' \mid \mathit{property})$. 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: $\mathit{inherit} \rightarrow \mathit{property} \in MD(B_{\mathbf{n}})$.

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.

3.3 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 $b_1 \ldots b_n$, then each agent b_i has to have a profile of each sequence of agents $b_{i+1} \ldots b_n$.

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:

agent b_1 deliberates about optimal decision
$\rightarrow$ considers optimal decision of agent b_2
agent b_2 deliberates about optimal decision
$\rightarrow$ considers optimal decision of agent b_3
agent b_3 deliberates about optimal decision
$\rightarrow$ considers optimal decision of agent b_4
...
agent b_n deliberates

Fig. 3. Recursive modelling

- 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 and van der Torre, 2003c]. 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 (see [Boella and van der Torre, 2003b] for observations).

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 and 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* $U(S, b) =$

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

The protocols are illustrated by our running example.

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

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

3.4 Example

In this section we conclude our example showing the recursive modelling of the normative system's decision on the basis of its beliefs and goals defining, respectively, constitutive and regulative norms.

The example presents a game played by an agent **a** who, in order to cultivate some crop (*crop*), fences (*fence*) a field to make that it is its property (*property*), as requested by the obligation $O_{\mathbf{an}}(property, s \mid crop)$ (see Example 4); fencing counts as being the owner *counts-as*$(fence, property \mid \top)$, as discussed in Example 5. This means that $B_{\mathbf{n}} = \{b_1\}$ and $MD(b_1) = fence \rightarrow property$, and $\in(member) = \{\langle \mathbf{a}, \mathbf{n} \rangle\}$ (see Example 3).

Agent **a** has to take a decision to fulfill its desire of cultivating crop ($MD(d_1) = \top \rightarrow crop$ and $d_1 = MD(D_{\mathbf{a}})$). At the same time, it does not desire to be sanctioned with s if it grows crop ($MD(d_2) = crop \rightarrow \neg s$ and $d_2 \in D_{\mathbf{a}}$) by the normative agent **n**, and the latter desire is stronger than the former ($\geq_{\mathbf{a}} \supseteq \{\{d_2\} \geq \{d_1\}\}$).

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.

Hence, agent **a** has not only to consider the effects of its behavior, but also to consider that the second agent in the protocol will act.

The possible decisions of agent **a** are doing nothing, fencing the field, cultivating crop, and all the possible combinations of these actions. The possible decisions of agent **n** are doing nothing, considering some violation, sanctioning and all the possible combinations of these actions.

The optimal decision of agent **a** has to take into account which is the reaction of agent **n**. Hence, agent **a** has to consider not the state immediately following its decision $\delta_{\mathbf{a}}$, rather the final state. So, even if the decision $\{crop\}$ satisfies all its desires in state $\top \rightarrow crop$, it is not the optimal decision, since in the subsequent state the effect includes the sanction s as a result of the normative agent's decision that the obligation has been violated. The optimal decision is, instead, $\{fence, crop\}$: agent **a** knows that fencing counts as being a property for agent **n**, hence, the obligation is not violated.

$NMAS_1 = \langle A, R \in, X, B, D, G, AD, MD, \geq, \mathbf{n}, I, V, GD \rangle$
$A = \{\mathbf{a}, \mathbf{n}\}, R = \{member\}, X_{\mathbf{a}} = \{fence, crop\},$
$B_{\mathbf{a}} = B_{\mathbf{n}}, G_{\mathbf{a}} = \emptyset, D_{\mathbf{a}} = \{d_1, d_2\}$

$X_{\mathbf{n}} = \{V(\neg property, \mathbf{a}), s\}$
$B_{\mathbf{n}} = \{b_1\}, G_{\mathbf{n}} = \{g_1, \ldots, g_5\}, D_{\mathbf{n}} = \{g_1, \ldots, g_5\}, GD(\mathbf{a}) = \{g_1\}$

$I = \{property\}$

$\delta_{\mathbf{a}} = \{fence, crop\}$
$\delta_{\mathbf{n}} = \emptyset$
$F_1 = out^+(B_{\mathbf{a}}, \delta_{\mathbf{a}} = \{fence, crop\}) = \{fence, crop, property\}$
$F_1 \cap I = \{property\}$
$F_2 = out^+(B'_{\mathbf{n}}, \delta_{\mathbf{a}} = \{fence, crop\} \cup \delta_{\mathbf{n}}) = \{fence, crop, property\}$

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

---

4 Conclusions

In this paper we define a formal framework for social entities like normative multiagent systems. By attributing mental attitudes to normative systems we model both constitutive and regulative norms as conditional rules representing, respectively, the beliefs and goals of the normative agent. Constitutive norms play the role of creating new abstract categories which compose the institutional reality. Roles are used to specify the powers of agents to create institutional facts or to modify the norms and obligations of the normative system.

The framework is used by means of recursive modelling to define games among agents 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.

In [Boella and van der Torre, 2004g] we extend this framework to model the problem of how the normative system itself specifies who can change the normative system itself. This specification is made by means of constitutive rules specifying what facts count as the creation of new regulative and constitutive rules in the normative system. This work is at the basis of the definition of contracts we make in [Boella and van der Torre, 2004b]. The negotiation of the distribution of obligations is described in [Boella and van der Torre, 2004d]. The agent metaphor is extended to model roles in [Boella and van der Torre, 2004a] and organizational structures in [Boella and van der Torre, 2004f].

Future work is considering 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], these conditionals must have a nonmonotonic character since it is possible that under some circumstances a certain state of affairs does not count as something else; e.g., it is not possible to fence a public garden to make it a property. Moreover, 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, 1958] Anderson, A. (1958). The logic of norms. *Logic et analyse*, 2.

[Artosi, 2002] Artosi, A. (2002). On the notion of an empowered agent. In *Procs. of LEA 2002 workshop*, Bologna.

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

[Boella and van der Torre, 2003b] Boella, G. and van der Torre, L. (2003b). Norm governed multiagent systems: The delegation of control to autonomous agents. In *Procs. of IEEE/WIC IAT'03*, pages 329– 335. IEEE Press.

[Boella and van der Torre, 2003c] Boella, G. and van der Torre, L. (2003c). Rational norm creation: Attributing mental attitudes to normative systems, part 2. In *Procs. of ICAIL'03*, pages 81–82. ACM Press.

[Boella and van der Torre, 2004a] Boella, G. and van der Torre, L. (2004a). Attributing mental attitudes to roles: The agent metaphor applied to organizational design. In *Procs. of ICEC'04*. IEEE Press.

[Boella and van der Torre, 2004b] Boella, G. and van der Torre, L. (2004b). Contracts as legal institutions in organizations of autonomous agents. In *Procs. of AAMAS'04*, pages 948–955, New York.

[Boella and van der Torre, 2004c] Boella, G. and van der Torre, L. (2004c). Δ: The social delegation cycle. In *LNAI n.3065: Procs. of ΔEON'04*, pages 29–42, Berlin.

[Boella and van der Torre, 2004d] Boella, G. and van der Torre, L. (2004d). The distribution of obligations by negotiation among autonomous agents. In *Procs. of ECAI'04*, pages 13–17. IOS Press.

[Boella and van der Torre, 2004e] Boella, G. and van der Torre, L. (2004e). Groups as agents with mental attitudes. In *Procs. of AAMAS'04*, pages 964–971, New York.

[Boella and van der Torre, 2004f] Boella, G. and van der Torre, L. (2004f). Organizations as socially constructed agents in the agent oriented paradigm. In *Procs. of ESAW'04*, Berlin. Springer Verlag.

[Boella and van der Torre, 2004g] Boella, G. and van der Torre, L. (2004g). Regulative and constitutive norms in normative multiagent systems. In *Procs. of KR'04*, pages 255–265.

[Breuker et al., 1997] Breuker, J., Valente, A., and Winkels, R. (1997). Legal ontologes: A functional view. In *Procs. of 1st LegOnt Workshop on Legal Ontologies*, pages 23–36.

[Broersen et al., 2002] 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, 2000] Castelfranchi, C. (2000). Engineering social order. In *Procs. of ESAW'00*, pages 1–18, Berlin. Springer Verlag.

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

[Dennett, 1987] Dennett, D. C. (1987). *The Intentional Stance*. The MIT Press, Cambridge, MA.

[Gelati et al., 2002] 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.

[Gmytrasiewicz and Durfee, 1995] Gmytrasiewicz, P. J. and Durfee, E. H. (1995). Formalization of recursive modeling. In *Procs. of ICMAS'95*, pages 125–132.

[Hindriks, 2002] Hindriks, F. (2002). The constitutive rule revisited. In *Procs. of 3rd Conference on Collective Intentionality*, Rotterdam.

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

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

[Lakoff and Johnson, 1980] Lakoff, G. and Johnson, M. (1980). *Metaphors we live by.* The University of Chicago Press, Chicago.

[Lang et al., 2002] Lang, J., van der Torre, L., and Weydert, E. (2002). Utilitarian desires. *Autonomous Agents and Multiagent Systems*, pages 329–363.

[Makinson, 1986] Makinson, D. (1986). On the formal representation of rights relations. *Journal of philosophical Logic*, 15:403–425.

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

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

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

[Searle, 1969] Searle, J. (1969). *Speech Acts: an Essay in the Philosophy of Language.* Cambridge University Press, Cambridge (UK).

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

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

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

[van der Hoek et al., 1999] van der Hoek, W., Hindriks, K., de Boer, F., and Meyer, J.-J. C. (1999). Agent programming in 3APL. *Autonomous Agents and Multi-Agent Systems*, 2(4):357–401.

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