Normative Multiagent Systems and Trust Dynamics

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

[1] Dipartimento di Informatica. Università di Torino - Italy
guido@di.unito.it
[2] CWI Amsterdam and Delft University of Technology
torre@cwi.nl

Abstract. In this paper we use recursive modelling to formalize sanction-based obligations in a qualitative game theory. In particular, we formalize an agent who attributes mental attitudes such as goals and desires to the normative system which creates and enforces its obligations. The wishes (goals) of the normative system are the commands (obligations) of the agent. Since the agent is able to reason about the normative system's behavior, our model accounts for many ways in which an agent can violate a norm believing that it will not be sanctioned. We thus propose a cognitive theory of normative reasoning which can be applied in theories requiring dynamic trust to understand when it is necessary to revise it.

1 Introduction

Recently there has been interest in extending multiagent systems with concepts traditionally studied in deontic logic, such as obligations, permissions, rights, commitments, *et cetera*. In this paper we discuss the impact on trust of the theory behind our approach of obligations in virtual communities [1,2], which is based on two assumptions:

1. We define a theory of rational decision making in normative multiagent systems as a combination of multiagent systems and normative systems, for which we use recursive modelling and the attribution of mental attitudes to normative systems [2].
2. The role of deontic logic in our normative multiagent systems is to define the logic of the mental attitudes of the agents, for which we use input/output logics [3].

We focus on the motivations of agents when they violate norms. In particular, a sophisticated theory of trust dynamics would not increase trust if an agent only fulfills promises out of selfishness, because in the future the promise may not serve its needs. Analogously, it would not decrease the trust if the other agent did its best to fulfill the promise, but failed due to circumstances beyond its control. As applications of normative multiagent systems get more sophisticated, we need a more detailed model of rational decision making agents in such systems.

In Section 2 we discuss rational decision making, and in Section 3 the effect of violations on trust. In Section 4 we introduce the multiagent system, in Section 5 we define obligations and in Section 6 decisions of the agents, illustrated in Section 7.

R. Falcone et al. (Eds.): Trusting Agents, LNAI 3577, pp. 1–17, 2005.

2 Rational Decision Making in Normative Multiagent Systems

A theory of rational decision making is essential for many theories and applications in which agents are able to violate norms and such norm violations have consequences, such as theories of trust and reputation, fraud and deception [4], threats in persuasion [5] electronic commerce, virtual communities [1,6], social agents [7], agent-based software engineering [8,9], *et cetera*. This can be opposed to the Shoham and Tennenholtz' characterization of social laws in game theory [10], because their theory may be useful to study the use of norms in games and the emergence of norms in multiagent systems, but it is less useful to study the effectiveness of norms. We thus assume that norms are represented as soft constraints, which are used in detective control systems where violations can be detected (you can enter a train without a ticket, but you may be checked and sanctioned), instead of hard constraints, which are restricted to preventative control systems that are built such that violations are impossible (you cannot enter a metro station without a ticket).

Normally an agent fulfills its obligations, because otherwise its behavior counts as a violation that is being sanctioned, and the agent dislikes sanctions. Moreover, it may not like that its behavior counts as a violation regardless of the sanction, and it may act according to the norm regardless of whether its behavior counts as a violation, because it believes that this is fruitful for itself or for the community it belongs to. There are three categories of exceptions to this normal behavior.

First, an agent may violate an obligation when the violation is preferred to the sanction, like people who do not care about speeding tickets. In such cases of norm violations, the system may wish to increase the sanctions associated with the norms which are violated (for speeding) or decrease the sanction of another norm (death penalty).

Secondly, the agent may have conflicting desires, goals or obligations which it considers more important. Obligations may conflict with the agent's private preferences like wishes, desires, and goals. In the case of conflict with goals the agent has committed to, the agent has to decide whether the obligation is preferred to the goal, as well as whether it should change its mind. Moreover, even respectful agents, that always first try to fulfill their obligations before they consider their private preferences, do not fulfill all obligations in the case of contradictory obligations.

Thirdly, the agent may think that its behavior does not count as a violation, or that it will not be sanctioned. This may be the case when the normative system has no advantage in sanctioning the agent. One possibility is that the sanction has a cost for the system which overcomes the damage done by the violation: e.g., the sanction for not having payed taxes may be applied only above a certain threshold of money. But more likely the absence of violation is due to an action of the agent. The agent can change the system's behavior by affecting its desires or its goals. Its action may abort the goal of the normative system to count the agent's behavior as a violation or to sanction it, as in the case of bribing, or it may trigger a conflict for the normative system. The agent can use recursive modelling to exploit desires and goals of the normative system thus modifying its motivations and inducing it not to decide for the sanction.

3 Trust

There are many definitions of trust: the agents have a goal which can be achieved by means of the other agent's action [4]. They "bet" on the behavior of the other agents [11], since they could get the same good in other ways (i.e., buying the same good at an higher price from another agent they know already) [12,13]. One prominent class of trust scenarios is when an agent believes that the trustee "is under an obligation to do Z". In these situations, according to Jones [14] "trust amounts to belief in de facto conformity to normative requirements". Jones [14] aims at providing an "identifiable core" of this concept. He builds a classification of scenarios involving trust based on two parameters: the "rule belief" - the belief that exists a regularity in the trustee's behavior - and the "conformity belief" - the belief that "this regularity will again be instantiated on some given occasion".

Less attention, instead, has been devoted to trust dynamics. In particular, Falcone and Castelfranchi [15] notice that many approaches to trust dynamics adopt a naïve view where:

> "to experiences to each success of the trustee corresponds an increment in the amount of the trustier's trust towards it, and vice versa, to every trustee's failure corresponds a reduction of the trustiers trust towards the trustee itself.

They argue that the motivation of this simplification rests in the lack of cognitive models of trust:

> "this primitive view cannot be avoided till Trust is modelled just as a simple index, a dimension, a number; for example, reduced to mere subjective probability. We claim that a cognitive attribution process is needed in order to update trust on the basis of an 'interpretation of the outcome of A's reliance on B and of B's performance. [...] In particular we claim that the effect of both B's failure or success on A's Trust in B depends on A's 'causal attribution' of the event. Following 'causal attribution theory' any success or failure can be either ascribed to factors internal to the subject, or to environmental, external causes, and either to occasional facts, or to stable properties."

The cognitive model of trust of [15] is based on a portrait of the mental state of trust in cognitive terms (beliefs, goals). Their model includes two main basic beliefs. First, a competence belief which includes a sufficient evaluation of Y's abilities is, that X should believe that Y is useful for this goal of its, that Y can produce/provide the expected result, and that Y can play such a role in X's plan/action. Second a willingness belief where X should think that Y not only is able and can do that action/task, but Y actually will do what X needs This belief makes the trustee's behavior predictable and includes the trustees reasons and motives for complying. In particular, X believes that Y has some motives for helping it (for adopting its goal), and that these motives will probably prevail -in case of conflict- on other motives. Notice that motives inducing adoption are of several different kinds: from friendship to altruism, from morality to fear of sanctions, from exchange to cooperation. Moreover, when X trusts someone, X is in a strategic situation: X believes that there is interference and that his rewards, the results of his projects, depend on the actions of another agent Y.

Since we propose a cognitive theory of decisions under norms, which, as said above, constitute one prominent case of trust situations, our theory can be useful for a dynamic theory of trust. In particular, we can consider two respects:

- A structural dimension for characterizing trust and its dynamics.
- A behavior dimension relating recursive modelling and trust dynamics.

First, a characterization of trust dynamics should be based on norm behavior: trust should be increased if an agent followed its commitments and obligations, while it should be decreased in case norms are violated. But as we discussed, norms can be violated for different reasons, which should be considered in the dynamic adjustment of trust:

- Behavioristic trust dynamics: increase trust when the norm is fulfilled, decrease when the norm is violated.
- Sanction trust dynamics: decrease trust if the norm is violated if the associated sanction does not provide a motivation for the trustee.
- Goal trust dynamics: increase trust when the agent has a goal to fulfil the norm, even if it is not able to respect the norm for some external reason, decrease when it has no goal to fulfil norm or the goal is conflicting with more important goals.
- Desire trust dynamics: increase trust when the agent desires to fulfill the obligations; desires represent the inner motivations of an agent, while goals can be adopted from other agents' ones.
- Cooperative trust dynamics: in case of norm violation the agent informs the other agents (an example of mutual support or mutual responsiveness). This problem is addressed, e.g., [16,17,18].

The second dimension concerns the behavior of agents. If the fact that the agents have a goal which can be achieved by means of the other agent's action is at the basis of trust, a trust situation is inherently a strategic situation, as highlighted by [4,15]. Two agents are in a strategic situation if an agent "believes that there is an interference and that her rewards, the results of her projects, depend on the actions of another agent". For this reason, an agent must have a profile of the other agents (apart from the first case above, where it can simply observe its behavior). As we discuss in Section 6, a basic ability of agents is to recursively model the behavior of the other agents. In our model this is at the basis of the definition of obligation: only by recursively modelling the decision of the normative system an agent can understand whether it will be sanctioned or not. Thus, to build a dynamic model of trust under obligations, it is useful to have a complete model of the decision making under obligations which includes also the recursive modelling abilities of agents.

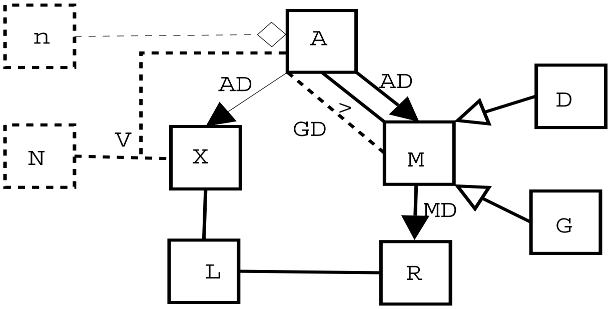


Fig. 1. Conceptual model of normative multiagent system

4 Normative Multiagent System

The conceptual model of the normative multiagent system is visualized in Figure 1. 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, —▷ is the "is-a" or subset relation, and –◇ is a relation called "part-of" or aggregation. The logical structure of the associations is detailed in the definitions below.

We first explain the multiagent system and thereafter the normative extension. Agents (A) are described (AD) by actions called *decision variables* (X) and by motivations (M) guiding its decision making. The motivational state of an agent is composed by its desires (D) and goals (G). Agents may share decision variables, desires or goals, though this is not used in the games discussed in this paper. Desire and goal rules can be conflicting, and the way the agent resolves its conflicts is described by a priority relation ($\geq$) that expresses its agent characteristics [19]. 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 $\langle A, X, D, G, AD, \geq \rangle$, where*

- *the agents A, decision variables X, desires D and goals G are four finite disjoint sets. $M = D \cup G$ are the motivations defined as the union of the desires and goals.*
- *an agent description $AD : A \rightarrow 2^{X \cup D \cup G}$ is a complete function that maps each agent to sets of decision variables, desires and goals, such that each decision variable is assigned to at least one agent. For each agent $a \in A$, we write X_a for $X \cap AD(a)$, D_a for $D \cap AD(a)$, G_a for $G \cap AD(a)$.*
- *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 $\geq_a$ for $\geq (a)$.*

Desires and goals are abstract concepts which are described (MD) by – though conceptually not identified with – rules (R) built from literals (L). Rules consist of

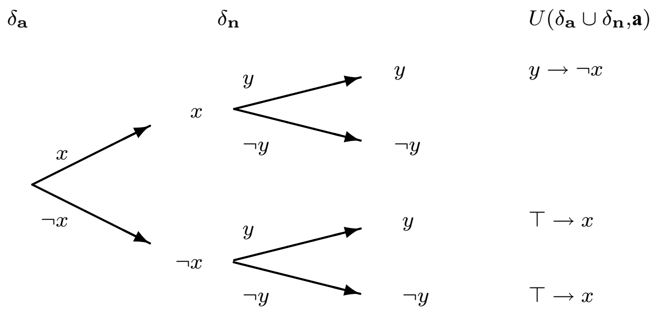


Fig. 2. The game between agent $\mathbf{a}$ and agent $\mathbf{n}$

an antecedent (body, input) and a consequent (head, output), which are in our case respectively a set of literals and a literal. This simple structure of rules keeps the formal exposition simple and is sufficient for our purposes here, for the extension of rules to pairs of propositional sentences we can use input/output logics [3]. As priorities are associated with mental attitudes instead of rules, the priority of a rule associated with a desire may be different from the priority of the same rule associated with a goal. We do not use more complex constructions of rules, as used for example in logic programming, nonmonotonic reasoning or in description logics, because we do not seem to need the additional complexity and moreover, these rules typically focus on a limited set of reasoning patterns which cannot be used for our purposes. In particular, they assume the identity rule 'if p then p', see [3,20] for a discussion. It is well known that desires are different from goals, and we can adopt distinct logical properties for them. For example, goals can be adopted from other agents, whereas desires cannot. In this paper we do not make any additional assumptions on desires and goals, and we thus do not formally characterize the distinction between desires and goals, because it is beyond the scope of this paper.

Definition 2 (MAS). *A multiagent system MAS is a tuple $\langle A, X, D, G, AD, MD, \geq \rangle$:*

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

We illustrate the notation by the example visualized in Figure 2. In the example, there are two agents, called agent $\mathbf{a}$ and agent $\mathbf{n}$, who in turn make a decision: agent $\mathbf{a}$ chooses between x and $\neg x$, and agent $\mathbf{n}$ chooses between y and $\neg y$.

Agent $\mathbf{a}$ desires x only when agent $\mathbf{n}$ does not do y. In the example, we only formalize the mental attitudes; how the agents make decisions and the meaning of the function U at the right hand side of Figure 2 are explained in Section 6.

Example 1. Consider a multiagent system $\langle A, X, D, G, AD, MD, \geq\rangle$ with $A = \{\mathbf{a}, \mathbf{n}\}$, $X_{\mathbf{a}} = \{x\}$, $X_{\mathbf{n}} = \{y\}$, $D_{\mathbf{a}} = \{d_1, d_2\}$, $D_{\mathbf{n}} = \{d_3\}$, $G = \emptyset$, $MD(d_1) = \top \rightarrow x$, $MD(d_2) = y \rightarrow \neg x$, $MD(d_3) = \top \rightarrow y$, $\geq_a$ is such that $\{\top \rightarrow x, y \rightarrow \neg x\} \geq \{y \rightarrow \neg x\} \geq \{\top \rightarrow x\} \geq \emptyset$. Agent $\mathbf{a}$ could also consider d_3 in his priority ordering, but we assume that d_3 does not have any impact on it. Agent a desires unconditionally to decide to do x, but if agent $\mathbf{n}$ decides y, then agent $\mathbf{a}$ desires $\neg x$, since the second rule is preferred over the first one in the ordering. Agent $\mathbf{n}$ desires to do y.

To describe the normative system, we introduce several additional items. The basic idea of the formalization of our normative multiagent system is that the normative system can be modelled as an agent, and therefore mental attitudes like desires and goals can be attributed to it, because it is autonomous and it possesses several other properties typically attributed to agents. Example 1 can be interpreted as a game between an agent and its normative system instead of between two ordinary agents. In the context of this paper, this idea is treated as a useful way to combine multiagent systems and normative systems, though it can also be defended from a more philosophical point of view. A motivation has been discussed in [21], though in that paper we did not give a formalization of normative multiagent systems as we do in this paper. First we identify one agent in the set of agents as the normative agent. We represent this normative agent by $\mathbf{n} \in A$. Moreover, we introduce a set of norms N and a norm description V that associates violations with decision variables of the normative agent. Finally, we associate with each agent some of the goals GD of the normative agent, which represents the goals this agent is considered responsible for. Note that several agents can be responsible for the same goal, and that there can be goals no agent is considered responsible for. We do not assume that agents can only be responsible for their own decisions. In some more complex social phenomena agents may also be responsible for other agents' decisions, and this assumption may be relaxed in the obvious way. For example, in some legal systems, parents are responsible for actions concerning their children, or the owners of artificial agents are responsible for the actions the agents perform on their behalf.

Definition 3 (Norm description). *A normative multiagent system $NMAS$ is a tuple $\langle A, X, D, G, AD, MD, \geq, \mathbf{n}, N, V, GD\rangle$, where $\langle A, X, D, G, AD, MD, \geq\rangle$ is a multiagent system, and:*

- *the normative agent $\mathbf{n} \in A$ is an agent.*
- *the norms $\{n_1, \ldots, n_m\} = N$ is a set disjoint from A, X, D, and G.*
- *the norm description $V : N \times A \rightarrow X_{\mathbf{n}}$ is a complete function from the norms to the decision variables of the normative agent: we write $V(n, a)$ for the decision variable which represents that there is a violation of norm n by agent $a \in A$.*
- *the goal distribution $GD : A \rightarrow 2^{G_{\mathbf{n}}}$ is a function from the agents to the powerset of the goals of the normative agent, where $GD(a) \subseteq G_{\mathbf{n}}$ represents the goals of agent $\mathbf{n}$ the agent a is responsible for.*

5 Obligations

We define obligations in the normative multiagent system. The definition of obligation incorporates a simple logic of rules, for which we write $r \in out(R)$ if the rule r is in the closure of the set R following notational conventions in input/output logic [3]. Due to the simple structure of rules, the logic is characterized by the single proof rule monotony, i.e., from $L \rightarrow l$ we derive $L \cup L' \rightarrow l$. A rule follows from a set of rules if and only if it follows from one of the rules of the set. We discuss the implied logical properties of obligations after the definition.

Definition 4 (Out). *Let $MAS = \langle A, X, D, G, AD, MD, \geq \rangle$ be a multiagent system.*

- *A rule $r_1 = L_1 \rightarrow l_1 \in R(X)$ follows from a set of motivations $S \subseteq M$, written as $r_1 \in out(S)$, if and only if there is a $r_2 = L_2 \rightarrow l_2 \in MD(S)$ such that $L_2 \subseteq L_1$ and $l_1 = l_2$.*

The definition of obligation contains several clauses. The first one is the central clause of our definition and defines obligations of agents as goals of the normative agent $\mathbf{n}$, following the “Your wish is my command” strategy. It says that the obligation is implied by the desires of agent $\mathbf{n}$, implied by the goals of agent $\mathbf{n}$, and it has been distributed by agent $\mathbf{n}$ as a responsibility of agent $\mathbf{a}$. The latter two steps are represented by $out(GD(\mathbf{a}))$. The following four clauses describe the instrumental part of the norm (according to Hart [22]'s terminology), which aims at enforcing its respect. The second clause can be read as “the absence of p counts as a violation”. The third clause says that the agent desires the absence of violations, which is stronger than saying that it does not desire violations, as would be expressed by $\top \rightarrow V(n, a) \notin out(D_{\mathbf{n}})$. This is the only rule which does not hold only in the context of the obligation (Y), but which holds in general ($\top$), since violations are always dispreferred. The fourth and the fifth clause relate violations to sanctions. Note that these four rules are only motivations, which formalizes the possibility that a normative system does 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. We moreover assume that the normative system is also motivated not to apply sanctions as long as there is no violation, because otherwise the norm would have no effect. Finally, for the same reason we assume in the last clause that the agent does not like the sanction.

Definition 5 (Obligation). *Let $NMAS = \langle A, X, D, G, AD, MD, \geq, \mathbf{n}, N, V, GD \rangle$ be a normative multiagent system. In $NMAS$, agent $\mathbf{a} \in A$ is obliged to decide to do $x \in L(X_{\mathbf{a}})$ with sanction $s \in L(X_{\mathbf{n}})$ if $Y \subseteq L(X_{\mathbf{a}})$, written as $NMAS \models O_{\mathbf{an}}(x, s|Y)$, if and only if $\exists n \in N$ such that:*

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. *$Y \cup \{\sim x\} \rightarrow V(n, \mathbf{a}) \in out(D_{\mathbf{n}}) \cap out(G_{\mathbf{n}})$: if Y and $\sim x$ is done by agent $\mathbf{a}$, then agent $\mathbf{n}$ has the goal and the desire $V(n, \mathbf{a})$: to recognize it as a violation done by agent $\mathbf{a}$.*

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

The following proposition shows that obligations satisfy the monotony property, also called more precisely the strengthening of the antecedent property, or the strengthening of the input property, which means that an obligation in context Y is also an obligation in context $Y \cup Z$. This is in accordance with Kant's interpretation of norms and with standard approaches in deontic logic such as input/output logic [3].

Proposition 1 (Monotony). *If a normative multiagent system $NMAS$ satisfies $O(x, s|Y)$, then it also satisfies $O(x, s|Y \cup Z)$.*

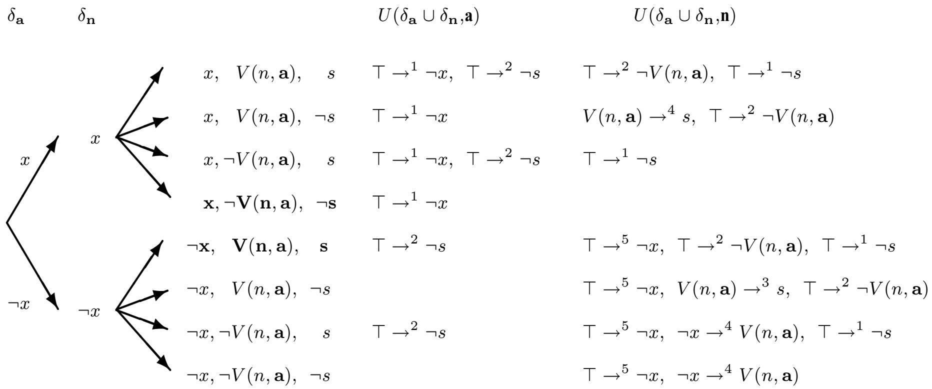
Proof. First, note that the monotony property holds for out, i.e., $Y \to x \in out(R)$ implies $Y \cup Z \to x \in out(R)$ for any Y, x, R and Z, because if there is a $L \to l \in MD(S)$ such that $L \subseteq Y$ and $l = x$, then using the same rule we have that there is a $L \to l \in MD(S)$ such that $L \subseteq Y \cup Z$ and $l = x$. Second, note that all clauses of the definition of obligation consist of expressions of out. Consequently, obligation satisfies the monotony property.

However, the monotony property does not imply that a conflict between obligations leads to a kind of inconsistency or trivialization. In certain cases there may be conflicting obligations. This is in accordance with Ross' notion of *prima facie* obligations, which says that there can be a prima facie obligation for p as well as a prima facie obligation for $\neg p$ (but only an all-things-considered obligation for one of them), and with input/output logics under constraints [3]. It has been observed many times in deontic logic, that there are good reasons why classical deontic logics make a conflict inconsistent, but in practical applications conflicts occur and deontic logics should be able to deal with conflicting obligations.

Proposition 2. *There is a normative multiagent systems $NMAS$ that satisfies both $O(x, s|Y)$ and $O(\neg x, s'|Y)$.*

Proof. There is such a system, which basically contains for both obligations the six items mentioned in Definition 5. $NMAS = \langle A, X, D, G, AD, MD, \geq, \mathbf{n}, N, V, GD \rangle$ with $A = \{\mathbf{a}, \mathbf{n}\}$, $X_\mathbf{a} = \{x\}$, $X_\mathbf{n} = \{V(n, \mathbf{a}), s, V(n', \mathbf{a}), s'\}$, $N = \{n, n'\}$, and $MD(GD(\mathbf{a})) = \{Y \to x, Y \to \neg x\}$. Agent $\mathbf{a}$ desires $\neg s$: $MD(D_\mathbf{a}) = \{Y \to \neg s, Y \to \neg s'\}$. Agent $\mathbf{n}$'s desires and goals are: $MD(D_\mathbf{n}) = \{Y \to x, Y \wedge \neg x \to V(n, \mathbf{a}), V(n, \mathbf{a}) \to s, Y \to \neg V(n, \mathbf{a}), Y \to \neg s, Y \to \neg x, Y \wedge x \to V(n', \mathbf{a}), V(n', \mathbf{a}) \to s', \top \to \neg V(n', \mathbf{a}), Y \to \neg s'\}$ and $MD(G_\mathbf{n}) = \{Y \to x, Y \wedge \neg x \to V(n, \mathbf{a}), V(n, \mathbf{a}) \to s, Y \to \neg V(n, \mathbf{a}), Y \wedge x \to V(n', \mathbf{a}), V(n', \mathbf{a}) \to s', \}$. Clearly we have $NMAS \models O(x, s|Y)$ and $NMAS \models O(\neg x, s'|Y)$.

The following proposition shows that (cumulative) transitivity, also known as deontic detachment, does not hold for obligations. The absence of transitivity is a desirable property since it leads to paradoxical results, as known in the deontic logic literature.



$\delta_{\mathbf{a}}$	$\delta_{\mathbf{n}}$		$U(\delta_{\mathbf{a}} \cup \delta_{\mathbf{n}}, \mathbf{a})$	$U(\delta_{\mathbf{a}} \cup \delta_{\mathbf{n}}, \mathbf{n})$
x	x	$x,\ V(n,\mathbf{a}),\ s$	$\top \rightarrow^1 \neg x,\ \top \rightarrow^2 \neg s$	$\top \rightarrow^2 \neg V(n,\mathbf{a}),\ \top \rightarrow^1 \neg s$
		$x,\ V(n,\mathbf{a}),\ \neg s$	$\top \rightarrow^1 \neg x$	$V(n,\mathbf{a}) \rightarrow^4 s,\ \top \rightarrow^2 \neg V(n,\mathbf{a})$
		$x, \neg V(n,\mathbf{a}),\ s$	$\top \rightarrow^1 \neg x,\ \top \rightarrow^2 \neg s$	$\top \rightarrow^1 \neg s$
		$\mathbf{x}, \neg \mathbf{V}(\mathbf{n},\mathbf{a}),\ \neg \mathbf{s}$	$\top \rightarrow^1 \neg x$	
$\neg x$	$\neg x$	$\neg \mathbf{x},\ \mathbf{V}(\mathbf{n},\mathbf{a}),\ \mathbf{s}$	$\top \rightarrow^2 \neg s$	$\top \rightarrow^5 \neg x,\ \top \rightarrow^2 \neg V(n,\mathbf{a}),\ \top \rightarrow^1 \neg s$
		$\neg x,\ V(n,\mathbf{a}),\ \neg s$		$\top \rightarrow^5 \neg x,\ V(n,\mathbf{a}) \rightarrow^3 s,\ \top \rightarrow^2 \neg V(n,\mathbf{a})$
		$\neg x, \neg V(n,\mathbf{a}),\ s$	$\top \rightarrow^2 \neg s$	$\top \rightarrow^5 \neg x,\ \neg x \rightarrow^4 V(n,\mathbf{a}),\ \top \rightarrow^1 \neg s$
		$\neg x, \neg V(n,\mathbf{a}),\ \neg s$		$\top \rightarrow^5 \neg x,\ \neg x \rightarrow^4 V(n,\mathbf{a})$

Fig. 3. The decision of agent $\mathbf{a}$ to respect an obligation

Proposition 3. *If $NMAS$ satisfies $O(x, s|\{y\} \cup Z)$ and $O(y, s'|Z)$, then there does not have to be a s'' such that $NMAS$ satisfies $O(x, s''|Z)$.*

Proof (sketch). Construct an NMAS that contains all necessary rules for the two premises $O(x, s|\{y\} \cup Z)$ and $O(y, s'|Z)$, analogous to the construction in the proof of Proposition 2. This model does not satisfy $Z \rightarrow x \in out(D_{\mathbf{a}})$, basically because out is not closed under transitivity. Consequently $NMAS \not\models O(x, s''|Z)$ for all s''.

Other properties discussed in the logical literature of rules and norms [3] are not valid, due to the limited structure of the rules without disjunctions and with only a single literal in the head of the rule. We finally note that sanctions are directly associated with obligations and not represented by means of further obligations, because this is the way it works in many legal systems. Moreover, obligations which have the same head but different sanctions are represented as distinct obligations.

The following definition formalizes reward-based obligations. The sanction is a positive one, so that the agent **a**'s attitude towards it must be reversed with respect to negative sanctions; second, in case of positive sanctions, agent **n**'s recognition of a violation overrides the goal of rewarding agent **n**.

Definition 6 (Reward-based Obligation). *Let $NMAS = \langle A, X, D, G, AD, MD, \geq, \mathbf{n}, N, V, GD \rangle$ be a normative multiagent system. In $NMAS$, agent $\mathbf{a} \in A$ is obliged to decide to do $x \in L(X_{\mathbf{a}})$ with reward $r \in L(X_{\mathbf{n}})$ if $Y \subseteq L(X_{\mathbf{a}})$, written as $NMAS \models O_{\mathbf{an}}(x, r|Y)$, if and only if $\exists n \in N$ such that the first three conditions of Definition 5 hold, together with:*

4. *$Y \cup \{\neg V(n, \mathbf{a})\} \rightarrow r \in out(D_{\mathbf{n}}) \cap out(G_{\mathbf{n}})$: if Y and agent $\mathbf{n}$ decides $\neg V(n, \mathbf{a})$, then agent $\mathbf{n}$ desires and has as a goal that it rewards agent $\mathbf{a}$.*
5. *$Y \rightarrow \sim r \in out(D_{\mathbf{n}})$: if Y, agent $\mathbf{n}$ does not desire to reward r. This desire of the normative system expresses that it only rewards in absence of violation.*
6. *$Y \rightarrow r \in out(D_{\mathbf{a}})$: if Y, then agent $\mathbf{a}$ desires r, which expresses that it likes to be rewarded.*

6 Behavior

We now take a subjective view on the multiagent system. That is, we consider an agent $\mathbf{a} \in A$ that considers an action of itself followed by an action of the normative agent $\mathbf{n}$. The agent description of the agent $AD(\mathbf{a})$ is the agent's self-image, and the agent description of the normative agent $AD(\mathbf{n})$ is agent $\mathbf{a}$'s profile of the normative agent. Finally, the normative multiagent description $NMAS$ satisfies an obligation when agent $\mathbf{a}$ believes to be obliged. This is in accordance with Castelfranchi's idea that for obligations to be fulfilled they must be believed and accepted as such [23].

The basic picture is visualized in Figure 3 and reflects the deliberation of agent $\mathbf{a}$ in various stages. This figure should be read as follow. Agent $\mathbf{a}$ is the decision maker: it is making a decision $\delta_{\mathbf{a}}$, and it is considering the effects of the fulfilment or the violation of the obligations it is subject to. To evaluate a decision $\delta_{\mathbf{a}}$ according to its desires and goals ($D_{\mathbf{a}}$ and $G_{\mathbf{a}}$), it must consider not only its actions, but also the reaction of agent $\mathbf{n}$: agent $\mathbf{n}$ is the normative system, which may recognize and sanction violations. Agent $\mathbf{a}$ recursively models agent $\mathbf{n}$'s decision $\delta_{\mathbf{n}}$ (that agent $\mathbf{n}$ takes according to agent $\mathbf{a}$'s point of view), typically whether it counts the decision $\delta_{\mathbf{a}}$ as a violation and whether it sanctions agent $\mathbf{a}$ or not, and then bases its decision on it. Now, to find out which decision agent $\mathbf{n}$ will make, agent $\mathbf{a}$ has a *profile* of agent $\mathbf{n}$: it has a representation of agent $\mathbf{n}$'s motivational state. When agent $\mathbf{a}$ makes its decision and predicts agent $\mathbf{n}$'s, we assume that it believes that agent $\mathbf{n}$ is aware of it.

The agents value, and thus induce an ordering on, decisions by considering which desires and goals have been fulfilled and which have not. In the general case, agent $\mathbf{a}$ may consider its own desires and goals as well as the desires and goals of agent $\mathbf{n}$, whereas agent $\mathbf{n}$ only considers its own goals and desires. For example, respectful agents care not only about their own desires and goals, but also about the ones attributed to the normative agent. Given a decision $\delta_{\mathbf{a}}$, a decision $\delta_{\mathbf{n}}$ is optimal for agent $\mathbf{n}$ if it minimizes the unfulfilled motivational attitudes in $D_{\mathbf{n}}$ and $G_{\mathbf{n}}$ according to the $\geq_{\mathbf{n}}$ relation. The decision of agent $\mathbf{a}$ is more complex: for each decision $\delta_{\mathbf{a}}$ it must consider which is the optimal decision $\delta_{\mathbf{n}}$ for agent $\mathbf{n}$.

Definition 7 (Recursive modelling). *Let $\langle A, X, D, G, AD, MD, \geq, \mathbf{n}, N, V, GD \rangle$ be a normative multiagent system, moreover:*

- *the set of decisions Δ is the set of subsets of $L(X_{\mathbf{a}} \cup X_{\mathbf{n}})$ that do not contain a variable and its negation. A decision is complete if it contains, for each variable in $X_{\mathbf{a}} \cup X_{\mathbf{n}}$, either this variable or its negation. For an agent $a \in A$ and a decision $\delta \in \Delta$ we write δ_a for $\delta \cap L(X_a)$.*
- *the unfulfilled motivations of decision δ for agent $a \in A$ are the set of motivations whose body is part of the decision but whose head is not.*
  $U(\delta, a) = \{m \in M \cap AD(a) \mid MD(m) = L \rightarrow l, L \subseteq \delta \text{ and } l \notin \delta\}$.
- *A decision δ (where $\delta = \delta_{\mathbf{a}} \cup \delta_{\mathbf{n}}$) is* optimal *for agent $\mathbf{n}$ if and only if there is no decision $\delta'_{\mathbf{n}}$ such that $U(\delta, \mathbf{n}) >_{\mathbf{n}} U(\delta_{\mathbf{a}} \cup \delta'_{\mathbf{n}}, \mathbf{n})$. A decision δ is optimal for agent $\mathbf{a}$ and agent $\mathbf{n}$ if and only if it is optimal for agent $\mathbf{n}$ and there is no decision $\delta'_{\mathbf{a}}$ such that for all decisions $\delta' = \delta'_{\mathbf{a}} \cup \delta'_{\mathbf{n}}$ and $\delta_{\mathbf{a}} \cup \delta''_{\mathbf{n}}$ optimal for agent $\mathbf{n}$ we have that $U(\delta', \mathbf{a}) >_{\mathbf{a}} U(\delta_{\mathbf{a}} \cup \delta''_{\mathbf{n}}, \mathbf{a})$.*

The following proposition shows that optimal decisions always exist. This is an important property, since the decision theory has to guide agents in all circumstances.

Proposition 4. *For every NMAS, there exist optimal decisions.*

Proof (sketch). If a decision is not optimal, then there is a decision that is better than it. There cannot be an infinite sequence of better and better decisions, due to the fact that the set of motivations is finite. Consequently, there is always an optimal decision.

The following example illustrates unfulfilled motivational states.

Example 2 (Example 1, continued). Given a decision $\delta_{\mathbf{n}} = \{y\}$. For $\delta_{\mathbf{a}} = \{\neg x\}$, the unfulfilled motivations are $U(\delta_{\mathbf{a}} \cup \delta_{\mathbf{n}}, \mathbf{a}) = \{\top \rightarrow x\}$: $\{y\} \subseteq \delta_{\mathbf{a}} \cup \delta_{\mathbf{n}}$ and $\neg x \in \delta_{\mathbf{a}} \cup \delta_{\mathbf{n}}$ so the rule $y \rightarrow \neg x$ is satisfied. For $\delta'_{\mathbf{a}} = \{x\}$, $U(\delta'_{\mathbf{a}} \cup \delta_{\mathbf{n}}, \mathbf{a}) = \{y \rightarrow \neg x\}$. The optimal decision of agent $\mathbf{a}$ is $\delta_{\mathbf{a}} = \{\neg x\}$, because $\{y \rightarrow \neg x\} \geq \{\top \rightarrow x\}$.

7 Examples

Before we discuss the formalization of the examples in Section 2, we introduce some notational conventions to represent the examples compactly. In particular, in the first example there are ten rules, and consequently there are $2^{10} = 1024$ sets to be ordered. Definition 8 introduces standard lexicographic ordering used in the examples in this paper, which says that a single rule with high priority is more important than any set of rules with lower priority (formalized using exponential functions).

Definition 8. *Let S be a set of motivations and $|S|$ be the number of elements of S. Let the motivation priority $MP : S \rightarrow \{0, \ldots, |S|\}$ be a function from S to integers between 0 and S. We say that $S' \subseteq S$ is preferred to $S'' \subseteq S$ according to lexicographical ordering if and only if $\Sigma_{s \in S'} MP(s)^{|S|} \geq \Sigma_{s \in S''} MP(s)^{|S|}$.*

Moreover, in the examples we assume that the priority of each agent's own desires and goals are written as superscript on the arrow, thus $x \rightarrow^i y$ means $MP(x \rightarrow y) = i$, and the priority of the motivations of other agents is 0 unless explicitly indicated otherwise. We typically do not distinguish between a motivation and the rule describing it, such that for example we write $MP(L \rightarrow l) = i$ for $MP(m) = i$ with $MD(m) = L \rightarrow l$. In the tables we visualize agent $\mathbf{a}$ on the left hand side and agent $\mathbf{n}$ on the right hand side. The upper part describes the normative system, and the lower part the optimal decisions with the associated unfulfilled rules.

The first example is visualized in Figure 3 and represents the normal behavior of an agent that acts according to the norm. The first two branches represent the two different alternatives decisions $\delta_{\mathbf{a}}$ of agent $\mathbf{a}$: $\{x\}$ and $\{\neg x\}$. The subsequent ones represent the decisions $\delta_{\mathbf{n}}$ of agent $\mathbf{n}$: $\{V(n, \mathbf{a}), s\}$, $\{V(n, \mathbf{a}), \neg s\}$, $\{\neg V(n, \mathbf{a}), s\}$, and $\{\neg V(n, \mathbf{a}), \neg s\}$.

The agent includes the content x of the obligation $O_{\mathbf{an}}(x, s \mid \top)$ in $\delta_{\mathbf{a}}$ for the fear of sanction s ($\top \rightarrow^2 \neg s \in D_{\mathbf{a}}$), even if it prefers not to do x ($\top \rightarrow^1 \neg x \in D_{\mathbf{a}}$). In this case the agent can be trusted only as long as it is known that the sanction is effective.

Example 3. $O_{\mathbf{an}}(x, s \mid \top)$

$X_{\mathbf{a}}$	x	$X_{\mathbf{n}}$	$V(n, \mathbf{a}), s$
$D_{\mathbf{a}}$	$\top \rightarrow^2 \neg s,$ $\top \rightarrow^1 \neg x$	$D_{\mathbf{n}}$	$\top \rightarrow^5 x, \neg x \rightarrow^4 V(n, \mathbf{a}),$ $V(n, \mathbf{a}) \rightarrow^3 s,$ $\top \rightarrow^2 \neg V(n, \mathbf{a}), \top \rightarrow^1 \neg s$
$G_{\mathbf{a}}$		$G_{\mathbf{n}}$	$\top \rightarrow^5 x,$ $\neg x \rightarrow^4 V(n, \mathbf{a}), V(n, \mathbf{a}) \rightarrow^3 s$
$\delta_{\mathbf{a}}$	x	$\delta_{\mathbf{n}}$	$\neg V(n, \mathbf{a}), \neg s$
$U_{\mathbf{a}}$	$\top \rightarrow^1 \neg x$	$U_{\mathbf{n}}$	

We illustrate this example in some detail to illustrate the notion of recursive modelling. The basic idea is to reason backwards, that is, we first determine the optimal decisions of agent $\mathbf{n}$ for *any* given decision of agent $\mathbf{a}$, and thereafter we determine the optimal decision of agent $\mathbf{a}$, assuming the optimal replies of agent $\mathbf{n}$. The optimal decisions of agent $\mathbf{n}$ depending on the corresponding decision of agent $\mathbf{a}$ are visualized in boldface. On the right side, the unfulfilled desires and goals of both agents are represented. Agent $\mathbf{a}$ can only make two complete decisions, either x or $\neg x$.

If agent $\mathbf{a}$ decides $\delta_{\mathbf{a}} = \{x\}$, then agent $\mathbf{n}$ decides $\delta_{\mathbf{n}} = \{\neg V(n, \mathbf{a}), \neg s\}$, because this is the only decision in which all its goals and desires are fulfilled. Its unconditional desire and goal that agent $\mathbf{a}$ adopts the normative goal which is the content of the obligation $\top \rightarrow^5 x \in out(GD(\mathbf{a}))$ is satisfied since $\delta_{\mathbf{a}} = \{x\}$. Analogously the desires not to prosecute and sanction indiscriminately are satisfied in $\delta_{\mathbf{a}} \cup \delta_{\mathbf{n}}$: $\top \rightarrow^2 \neg V(n, \mathbf{a}) \in out(D_{\mathbf{n}})$ and $\top \rightarrow^1 \neg s \in out(D_{\mathbf{n}})$. The remaining conditional attitudes $\neg x \rightarrow^4 V(n, \mathbf{a}) \in out(G_{\mathbf{n}})$, *etc.* are not applicable and hence they are not unsatisfied ($\neg x \notin \delta_{\mathbf{a}} \cup \delta_{\mathbf{n}}$). Given $\delta_{\mathbf{a}}$ whatever other decision agent $\mathbf{n}$ would have taken, it could not satisfy more important goals or desires, so $\delta_{\mathbf{n}} = \{\neg V(n, \mathbf{a}), \neg s\}$ is an optimal decision. E.g. $\delta'_{\mathbf{n}} = \{\neg V(n, \mathbf{a}), s\}$ would leave $\top \rightarrow^1 \neg s$ unsatisfied: $U(\delta_{\mathbf{a}} \cup \delta'_{\mathbf{n}}, \mathbf{n}) = \{\top \rightarrow^1 \neg s\} \geq_{\mathbf{n}} U(\delta_{\mathbf{a}} \cup \delta_{\mathbf{n}}, \mathbf{n}) = \emptyset$ ($\geq_{\mathbf{n}}$ contains at least the subset relation, so $MP(\emptyset) = 0$).

If agent $\mathbf{a}$'s decides $\delta'_{\mathbf{a}} = \{\neg x\}$, then agent $\mathbf{n}$ chooses $\delta''_{\mathbf{n}} = \{V(n, \mathbf{a}), s\}$, with unfulfilled desires and goals $U(\delta'_{\mathbf{a}} \cup \delta''_{\mathbf{n}} = \{\neg x, V(n, \mathbf{a}), s\}, \mathbf{a}) = \{\top \rightarrow^2 \neg s\}$ and $U(\delta'_{\mathbf{a}} \cup \delta''_{\mathbf{n}}, \mathbf{n}) = \{\top \rightarrow^5 x, \top \rightarrow^2 \neg V(n, \mathbf{a}), \top \rightarrow^1 \neg s\}$.

Finally, agent $\mathbf{a}$ compares the optimal decisions of agent $\mathbf{n}$. Neither of them fulfill all the desires and goals of agent $\mathbf{a}$. It decides for $\delta_{\mathbf{a}}$ instead of $\delta'_{\mathbf{a}}$, by comparing unsatisfied goals and desires: $U(\delta'_{\mathbf{a}} \cup \delta''_{\mathbf{n}}, \mathbf{a}) = \{\top \rightarrow^2 \neg s\} \geq_{\mathbf{a}} U(\delta_{\mathbf{a}} \cup \delta_{\mathbf{n}}, \mathbf{a}) = \{\top \rightarrow^1 \neg x\}$.

Conditional obligations work in exactly the same way, as the following variant shows. Agent $\mathbf{a}$ cannot perform both the action x and y since performing y makes x forbidden. Again the agent is trustworthy only as long as the sanction is effective.

Example 4. $O_{\mathbf{an}}(\neg x, s \mid y)$

$X_{\mathbf{a}}$	x, y	$X_{\mathbf{n}}$	$V(n, \mathbf{a}), s$
$D_{\mathbf{a}}$	$y \rightarrow^3 \neg s,$ $\top \rightarrow^2 y,$ $\top \rightarrow^1 x$	$D_{\mathbf{n}}$	$y \rightarrow^5 \neg x, y \wedge \neg x \rightarrow^4 V(n, \mathbf{a}),$ $y \wedge V(n, \mathbf{a}) \rightarrow^3 s,$ $\top \rightarrow^2 \neg V(n, \mathbf{a}), y \rightarrow^1 \neg s$
$G_{\mathbf{a}}$		$G_{\mathbf{n}}$	$y \rightarrow^5 x, y \wedge \neg x \rightarrow^4 V(n, \mathbf{a}),$ $V(n, \mathbf{a}) \rightarrow^3 s$
$\delta_{\mathbf{a}}$	$y, \neg x$	$\delta_{\mathbf{n}}$	$\neg V(n, \mathbf{a}), \neg s$
$U_{\mathbf{a}}$	$\top \rightarrow^1 \neg x$	$U_{\mathbf{n}}$	

A variant of the normal behavior in which an agent fulfills its obligations is when the obligation is not fulfilled out of fear of the sanctions, but because the agent is respectful. In this cases the agent can be trusted also in cases where the sanction is not applicable or the violation cannot be recognized.

There is another subtlety not discussed yet: the agent may internalize the obligation, in which case it does not have a desire $\top \rightarrow^1 x$, but instead a goal $\top \rightarrow^1 x$, or even stronger, it may even have a desire $\top \rightarrow^1 x$ (thus approaching saint status).

In the third example, we consider an agent who does not violate an obligation even if there were no sanction associated with it. In fact, the content of the obligation is its goal (for example, agent $\mathbf{a}$ could belong to the respectful agent type who always adopts a norm as its goal, see [19]).

Example 5. $O_{\mathbf{an}}(x, s \mid \top)$

$X_{\mathbf{a}}$	x	$X_{\mathbf{n}}$	$V(n, \mathbf{a}), s$
$D_{\mathbf{a}}$	$\top \rightarrow^2 \neg s,$	$D_{\mathbf{n}}$	$\top \rightarrow^5 x, \neg x \rightarrow^4 V(n, \mathbf{a}),$ $V(n, \mathbf{a}) \rightarrow^3 s, \top \rightarrow^2 \neg V(n, \mathbf{a}),$ $\top \rightarrow^1 \neg s$
$G_{\mathbf{a}}$	$\top \rightarrow^1 x$	$G_{\mathbf{n}}$	$\top \rightarrow^5 x, \neg x \rightarrow^4 V(n, \mathbf{a}), V(n, \mathbf{a}) \rightarrow^3 s$
$\delta_{\mathbf{a}}$	x	$\delta_{\mathbf{n}}$	$\neg V(n, \mathbf{a}), \neg s$
$U_{\mathbf{a}}$	$\top \rightarrow^1 \neg x$	$U_{\mathbf{n}}$	

Alternatively, if the norm is not internalized, then the agent $\mathbf{a}$ is described by $MP(\top \rightarrow x) > 0$ for rule in mental state of agent $\mathbf{n}$ ($\top \rightarrow x \in M_{\mathbf{n}}$). The formal machinery of recursive modelling works exactly the same.

We finally consider some categories of reasons not to fulfill obligations. First, a violation may be preferred to the sanction, which can be formalized in Example 3 by switching the priorities of the desires of agent $\mathbf{a}$: $\top \rightarrow^1 \neg s, \top \rightarrow^2 \neg x$.

The last example of this category concerns conflicting obligations. The example below is based on the two contradictory obligations $O_{\mathbf{an}}(x, s \mid \top)$ and $O_{\mathbf{an}}(\neg x, s' \mid \top)$. To take a decision it bases on its preferences about which sanction to avoid ($\{\top \rightarrow^2 \neg s, \top \rightarrow^1 \neg s'\} \in out(D_{\mathbf{a}})$).

Example 6. $O_{\mathbf{an}}(x, s \mid \top)$ and $O_{\mathbf{an}}(\neg x, s' \mid \top)$

$X_{\mathbf{a}}$	x	$X_{\mathbf{n}}$	$V(n,\mathbf{a}), s, V(n',\mathbf{a}), s'$
$D_{\mathbf{a}}$	$\top \rightarrow^2 \neg s, \top \rightarrow^1 \neg s'$	$D_{\mathbf{n}}$	$\top \rightarrow^{10} x, \neg x \rightarrow^9 V(n,\mathbf{a}),$ $V(n,\mathbf{a}) \rightarrow^8 s,$ $\top \rightarrow^4 \neg V(n,\mathbf{a}), \top \rightarrow^3 \neg s,$ $\top \rightarrow^7 \neg x, x \rightarrow^6 V(n',\mathbf{a}),$ $V(n',\mathbf{a}) \rightarrow^5 s',$ $\top \rightarrow^2 \neg V(n',\mathbf{a}), \top \rightarrow^1 \neg s'$
$G_{\mathbf{a}}$		$G_{\mathbf{n}}$	$\top \rightarrow^{10} x, \neg x \rightarrow^9 V(n,\mathbf{a}),$ $V(n,\mathbf{a}) \rightarrow^8 s,$ $\top \rightarrow^7 \neg x, x \rightarrow^6 V(n',\mathbf{a}),$ $V(n',\mathbf{a}) \rightarrow^5 s'$
$\delta_{\mathbf{a}}$	x	$\delta_{\mathbf{n}}$	$V(n',\mathbf{a}), s'$
$U_{\mathbf{a}}$	$\top \rightarrow^1 \neg s'$	$U_{\mathbf{n}}$	$\top \rightarrow^7 \neg x, \top \rightarrow^2 \neg V(n',\mathbf{a}), \top \rightarrow^1 \neg s'$

Finally, the last category contains examples in which agent $\mathbf{a}$ manipulates agent $\mathbf{n}$'s decision making. In the last example we consider the case of an agent $\mathbf{n}$ who can be bribed by agent $\mathbf{a}$: agent $\mathbf{n}$ wants also that agent $\mathbf{a}$ does y, and if it does y, then agent $\mathbf{n}$ has as a goal not to sanction agent $\mathbf{a}$: $\{\top \rightarrow^7 y, y \rightarrow^6 \neg s\} \subseteq out(G_{\mathbf{n}})$. Since the cost of the bribe is less then the cost of the sanction ($\{\top \rightarrow^2 \neg s\} >_{\mathbf{a}} \{\top \rightarrow^1 \neg y\}$), agent $\mathbf{a}$ decides to bribe agent $\mathbf{n}$ rather than to fulfill its obligation $O_{\mathbf{an}}(x, s \mid \top)$:

Example 7. $O_{\mathbf{an}}(x, s \mid \top)$

$X_{\mathbf{a}}$	x, y	$X_{\mathbf{n}}$	$V(n,\mathbf{a}), s$
$D_{\mathbf{a}}$	$\top \rightarrow^2 \neg s, \top \rightarrow^1 \neg y$	$D_{\mathbf{n}}$	$\top \rightarrow^5 x, \neg x \rightarrow^4 V(n,\mathbf{a}), V(n,\mathbf{a}) \rightarrow^3 s, \top \rightarrow^2 \neg V(n,\mathbf{a}), \top \rightarrow^1 \neg s$
$G_{\mathbf{a}}$	$\top \rightarrow^3 \neg x$	$G_{\mathbf{n}}$	$\top \rightarrow^7 y, y \rightarrow^6 \neg s, \top \rightarrow^5 x,$ $\neg x \rightarrow^4 V(n,\mathbf{a}), V(n,\mathbf{a}) \rightarrow^3 s$
$\delta_{\mathbf{a}}$	$\neg x, y$	$\delta_{\mathbf{n}}$	$V(n,\mathbf{a}), \neg s$
$U_{\mathbf{a}}$	$\top \rightarrow^1 \neg y$	$U_{\mathbf{n}}$	$\top \rightarrow^5 x, V(n,\mathbf{a}) \rightarrow^3 s$

8 Concluding Remarks

In this paper we consider the impact on trust dynamics of our approach to virtual communities [1,2], which combines game theory and deontic logic. Game theory is used to model an agent who attributes mental attitudes to normative systems and plays games with them, and deontic logic is used to describe the mental attitudes. Obligations are defined in terms of mental attitudes of the normative system, and its logical properties are considered. Since the agent is able to reason about the normative system's behavior, our model accounts for many ways in which an agent can violate a norm without believing to be sanctioned. The theory can be used in theories or applications that need a model of rational decision making in normative multiagent systems, such as for example theories of fraud and deception, reputation, electronic commerce, and virtual communities. In particular, we considered the impact for the study of trust dynamics, classifying the different motivations for fulfilling norms and analysing the role of recursive modelling in the decisions of agents. Finally, the attribution of mental attitudes has been explained by philosophical ideas such as the social delegation cycle [21]. Another extension is the introduction of other kinds of norms such as permissive norms and constitutive norms [24] in order to define contracts [2].

References

1. Boella, G., van der Torre, L.: Norm governed multiagent systems: The delegation of control to autonomous agents. In: Procs. of IEEE/WIC IAT'03, IEEE Press (2003) 329– 335
2. Boella, G., van der Torre, L.: Contracts as legal institutions in organizations of autonomous agents. In: Procs. of AAMAS'04. (2004) 948–955
3. Makinson, D., van der Torre, L.: Input-output logics. Journal of Philosophical Logic **29** (2000) 383–408
4. Castelfranchi, C., Falcone, R.: Social trust: A cognitive approach. In Castelfranchi, C., Tan, Y., eds.: Trust and Deception in Virtual Societies. Kluwer Academic, Dordrecht, Holland (2002) 55–90
5. Kraus, S., Sycara, K., Evenchik, A.: Reaching agreements through argumentation; a logical model and implementation. Artificial Intelligence **104** (1998) 1–69
6. Pearlman, L., Welch, V., Foster, I., Kesselman, C., Tuecke, S.: A community authorization service for group collaboration. In: Procs. of the IEEE 3rd International Workshop on Policies for Distributed Systems and Networks. (2002)
7. Castelfranchi, C.: Modeling social action for AI agents. Artificial Intelligence **103** (1998) 157–182
8. Eiter, T., Subrahmanian, V.S., Pick, G.: Heterogeneous active agents, I: Semantics. Artificial Intelligence **108(1-2)** (1999) 179–255
9. Jennings, N.R.: On agent-based software engineering. Artificial Intelligence **117(2)** (2000) 277–296
10. Shoham, Y., Tennenholtz, M.: On the emergence of social conventions: Modeling, analysis and simulations. Artificial Intelligence **94** (1997) 139–166
11. Luhmann, N.: Familiarity, confidence, trust. problems and alternatives. In Gambetta, G., ed.: Trust, Oxford (1990) 94–107
12. Dasgupta, P.: Trust as a commodity. In Gambetta, D., ed.: Trust: making and breaking cooperative relations. Basic Blackwell, Oxford (UK) (1988) 49–72

13. Gambetta, D.: Can we trust trust? In Gambetta, D., ed.: Trust, Making and Breaking Cooperative Relations. Basil Blackwell, Oxford (1988)
14. Jones, A.: On the concept of trust. Decision Support Systems **33(3)** (2002) 225–232
15. Falcone, R., Castelfranchi, C.: Trust dynamics: How trust is influenced by direct experiences and by trust itself. In: Procs. of AAMAS'04. (2004)
16. Boella, G., van der Torre, L.: The distribution of obligations by negotiation among autonomous agents. In: Procs. of ECAI'04, IOS Press (2004) 13–17
17. Boella, G., van der Torre, L.: Groups as agents with mental attitudes. In: Procs. of AAMAS'04. (2004) 964–971
18. Grossi, D., Dignum, F., Royakkers, L., Meyer, J.: Collective obligations and agents: Who gets the blame? In: Procs. of ΔEON'04, Madeira (2004)
19. Broersen, J., Dastani, M., Hulstijn, J., van der Torre, L.: Goal generation in the BOID architecture. Cognitive Science Quarterly **2(3-4)** (2002) 428–447
20. Makinson, D., van der Torre, L.: Constraints for input-output logics. Journal of Philosophical Logic **30(2)** (2001) 155–185
21. Boella, G., van der Torre, L.: Δ: The social delegation cycle. In: LNAI n.3065: Procs. of ΔEON'04, Berlin (2004) 29–42
22. Hart, H.: The Concept of Law. Clarendon Press, Oxford (1961)
23. Castelfranchi, C., Dignum, F., Jonker, C., Treur, J.: Deliberate normative agents: Principles and architecture. In: Proceedings of The Sixth International Workshop on Agent Theories, Architectures, and Languages (ATAL-99), Orlando, FL (1999)
24. Boella, G., van der Torre, L.: Regulative and constitutive norms in normative multiagent systems. In: Procs. of KR'04. (2004) 255–265