

Deliberate Normative Agents

Guido Boella and Leonardo Lesmo
Dipartimento di Informatica and Centro di Scienza Cognitiva
Università di Torino
e-mail: {guido,lesmo}@di.unito.it

ABSTRACT

Agent theories have developed concepts and methodologies that can be applied for having a better understanding of reasoning about obligations. This work proposes an agent-based framework for modeling obligations and norms: agents are able to deal with norms and to decide autonomously whether to respect them or not. The key idea is that the addressee of the norm explicitly models the agent who watches on the norms and who can sanction him.

1. INTRODUCTION

The concepts of norm and obligation have achieved a renewed interest for their importance in multi-agent systems [19], [17]. In turn, the conceptual tools developed by agent theories may be useful for a better understanding of the concepts of norms and obligations. However, there seems to be some missing points about norms which have to be analysed in greater depth and that can receive a meaningful explanation inside a multi-agent framework. What is needed is a framework where the components of the distributed system are *deliberate normative agents* ([14]): that is, agents which have an explicit representation of norms and can reason about whether accepting and fulfilling them.

We propose an agent framework for reasoning about obligations and norms in which these concepts does not have a distinct ontological status, but are strictly integrated with goals and intentions: in this way, the sophisticated models developed for agents can be exploited for modeling deontic reasoning and, at the same time, agent models can be endowed with normative concepts.

In principle, an obligation is something an agent is obliged to do. In other words, given an initial situation, in any course of events produced by the agent chosen action(s) the obligation must be fulfilled. However, this need not be the most rational way for an agent to act. There can be situations where different obligations contrast with each other, or

situations where an obligation cannot be reconciled with the agent's personal desires or goals. In these cases, the agent must evaluate the situation carefully and must decide if the obligation (or, which obligation) must be pursued, and in which way.

Our attention is not devoted to moral assertions as “there should be no war” or “you should not kill” or to technical assertions as “in order to print a file, you should use the ‘lpr’ command”.¹ Our proposal is directed towards those obligations which are *personal* (i.e., they concern certain individuals), and which are issued by some entity which sanctions who violates some norm.

This assumption restricts the scope of the paper. But it can be noticed that the basic approach is only partially affected by this limitations. Also for moral norms, it holds that an agent breaking the norm can reach a state of negative desirability: perhaps not because of a re-action of a normative agent, but because it enters a negative mood (e.g. shame or fear), or because the entire community he lives within comes to play the role of the normative agent, making him an outcast.

The main aspects of norms which we consider in this paper are:

- the sanction waiting an agent who has violated the norm and the way it is applied,
- when it is rational for him to violate the norm,
- how he can do so reducing the probability of undergoing the sanction.

In particular, we focus our attention on the fact that a norm involves at least two individuals, both of which have to be modeled as (intelligent deliberative) agents: the *bearer* of the obligation, who must respect the norm, and the *normative agent* (an *authority*, in formal situations), which has posed the obligation, wants that the *bearers* of the norm fulfill it, and (possibly) will sanction the violators.²

¹Sometimes, works on deontic logic use the notion of obligation for modeling both kinds of assertions.

²In [17]'s terminology, we mostly consider situations where the *sovereign* (the agent who issued the norm) is a *defender* too (i.e., the agent who watches over the norm).

While “it is generally acknowledged that norms and normative action emphasize autonomy on the side of *decision*”³, no attention has been devoted to the fact that norms and obligation are enforced by the *normative* agent, who is an autonomous actor, too.

Up to now, the center of attention has been only the bearer of the norm. The remarkable exception [19] explicitly deals with sanction, but does not model the agent who is in charge of monitoring violations and applying the sanction.

In this work we will show the advantages of modeling as an agent also the counterpart of the *bearer* of the obligation.

The first consequence of this approach is that the *bearer* of the obligation has to consider explicitly the disadvantage of facing the sanction when he considers whether to fulfill the obligation. However, the sanction is not a granted exogenous event, but it is the result of an activity of the *normative* agent. He has the goal of checking the fulfillment of the norm and has a plan for doing so and eventually posing the sanction. But he also has other goals, preferences and obligations as any other agent.

The *bearer* of the obligation has to take into account all these facts when he considers the advantage of fulfilling or not the obligation: i.e. he has to model (recursively) also the *normative* agent as an agent.

Using recursive modeling of agents is a trend in agent theories which is developing in the last years [22], [28], [35] and [5]: these approaches are motivated by the fact that every action of an agent has an impact on the choices of other agents who can react to it. If the agent has enough information about the state of the other agents (i.e., their beliefs, goals, obligations, preferences and available plans), he can try to predict what they will do depending on what he decides to do. When this is possible, the agent has the opportunity to evaluate the goodness of his action not only from a local point of view, but from a state which includes the consequences due to the behavior of other agents. In particular, this form of reasoning has been proven useful in cases of cooperation among agents as in [23], [5] and [4].

On the other hand, the recursive modeling of the *normative* agent opens the way to another opportunity for the *bearer* besides a better evaluation of the resulting final state. The *bearer* agent can reason about how the *normative* agent will (decide to) check the fulfillment and will apply the sanction if he discovers a violation. This knowledge can allow an agent to predict when the *normative* agent possibly fails to become aware of a violation and how to induce him to this failure by means of some action.

Probably, many agent designers would think that the ability to deceive is not a desirable feature of agents, but, as also [17] states, there are some possible situations where this form of reasoning is useful; at least, for making agents aware about the possibility that rogue agents exploit them. In general, we claim that an agent is able to understand the behavior of another agent just in case he is able to build a model of his behavior in terms of possible plans and goals. So, no deceiving behavior can be understood (i.e. recognized) unless the agent has some knowledge about deceiving: any ‘honest’ normative agent needs some knowledge about dishonest

behavior if it is deemed to detect such a behavior. Moreover, as we will see, eluding the sanction is only one of the possible alternatives for agents who have to find a trade off among different factors ranging from their (material) utility and costs to social goals (as being sincere with other agents).

The structure of this paper is the following: first a reference example scenario is presented and discussed; in Section 3.1, the agent model is described and a formal definition of norm is presented in Section 3.2; then, we discuss how the agent model must be modified for dealing with obligations by means of recursive modeling. Finally, in Section 4, the different factors involved in the decision to fulfill an obligation are examined. Discussion, comparison with related work and conclusions end the paper.

2. A SAMPLE SCENARIO

As an example of obligation consider the assertion taken from a call for paper “authors should not submit their paper to other conferences”. This norm is personal (we are the agents obliged not to send this paper to other conferences), the SmartAgents2000 program committee is the institution who issued the norm and the possible sanction is the rejection of the paper by the SmartAgents2000 PC (and the less measurable, but still relevant, shame on the authors). The obligation has been posed for assuring the outstanding quality of the workshop for the convenience of all the participants.

The author agent (assume one for simplicity) has the goal of having as much publications as possible (for obvious academic purposes); however, he (currently) has just one article to submit: to overcome this shortcoming, he has the possibility to send the same paper to two prestigious conferences: SmartAgents2000 and DummyAgents2000. Since he has read the SmartAgents2000 call for papers, he knows that there is a norm concerning multiple submissions: multiple submissions are not infrequent in his scientific community, so, for a realistic scenario, the norm should (!) not be automatically accepted by the agent (otherwise, we could not model the less virtuous agents who populate conferences with cloned papers).

On the other hand, the sanction is not automatically applied when the author submits the paper: the SmartAgents2000 program committee (the other agent involved in the example) has to execute some ‘sensing’ actions for checking if authors have respected the norm. If he discovers a duplicated paper, he will apply the sanction by executing the ‘rejection’ action.

If the author sends the paper to both conferences and he is discovered, the SmartAgents2000 program committee rejects the paper: besides not achieving the advantage of having a further publication, he also incurs in the shame of the prestigious colleagues belonging to the program committee.

Therefore, the agent has to compare the alternatives of sending the paper to a single conference or to both, being aware of the possible reactions of the program committee.

³[17], p.100.

3. THE DEFINITION OF OBLIGATION

The description of the above simple normative situation allows us to highlight the similarities with other problems treated by agent theories; in particular, this example could be explained in terms of multi-agent interaction. We believe that what is relevant for the agent reasoning is when and how the sanction is applied and not only which is the ideal situation. Therefore, it is necessary a model where the players interact and reason about each others' behaviors.

In brief, we have an agent (the *normative* one) who has the power⁴ to accept or reject papers and has the goal that the papers he receives are not already submitted; in doing so, she decides to execute and executes a plan composed of checking and accepting/rejecting submissions.⁵

Second, there is another agent (the *bearer* of the obligation) who has the goal of submitting his paper to one or more conferences and who knows the goal of the program committee agent to accept only paper not published elsewhere and to reject the others: he has read the norm on the call for papers and now he has an internal representation of it.

Generalizing this example, we have that *an obligation holds when there is an agent A who has a goal G that another (or more than one) agent B satisfy it and who, in case the agent B has not adopted the goal G, has to decide whether to perform an action Act which (negatively) affects some aspect of the world which (presumably) interests B. Both agents know these facts.*

Differently from what appears at first sight, this definition covers not only 'institutional' cases, but also other situations like obligations in dialog (see [31] and [6]) which share the characteristic that new goals are acquired as a consequence of social inputs. Moreover, also inner rewards and punishments deriving from moral obligations can be considered.

We will discuss this topic in more depth in Section 5.

This definition allows BDI (Beliefs, Desires and Intentions) agents to deal with obligations since they are able to manage intentions, to take into account the goals of other agents and their behavior, to devise plans for satisfying goals, and to compare the alternative plans according to their preferences.

Some more words must be devoted to these different capabilities.

First of all, taking into account the goals of the other agents is, according to [13], one of the key capabilities for an agent to be social: social agents must be able to consider the goals of other agents and to have attitudes towards those goals, that is, to *adopt* those goals (i.e., "having a state of affairs as

⁴Even if we do not discuss these topics here, other features, beside power, should pertain to the *normative* agent, otherwise the definition risks to be undistinguishable from that of coercion (as [17] notice). For example, the normative agent should only pose norms which do not provide him with any personal advantage or that increase the overall utility of the society.

⁵We disregard the effect of the quality of the paper on the acceptance process.

a goal *because* another agent has the same state as a goal"); moreover, *goal adoption* is at the basis of the definition of cooperation among agents in [5] and [4], and of dialogical interaction in [2].

It must be noted that the satisfaction of a goal adopted from another agent *A* is not *per se* advantageous for an agent *B*. However, an adopted goal can be *instrumental* for *B*.

In the same way, an agent has normally a number of (private) goals which are only *instrumental* to other goals which appear among his preferences; the satisfaction of these goals is not by itself advantageous for the agent; consider the standard situation of adopting a subgoal for achieving the precondition of an action which satisfies the main goals of the agent.

The role of goals is to direct the planning process, but the fact that a goal is provided as an input to the planner does not assure that a plan for achieving it will become an intention of the agent: the candidate plans produced by the planning process are selected on the basis of the agent's preferences. In a similar way, the goals of other agents can be adopted (i.e., given as input to the planning process) not only because the agent has a preference towards them: there are also other *external* sources of goals like norms and obligations, requests by other agents or the need to provide help during cooperation, which can lead an agent to adopt another agent's goals.

In case of obligations, the *normative* agent *A* wants that the *bearer* of the obligation adopt the goal *G* concerning the obligation. Moreover, the *bearer* *B* knows the reaction of agent *A* if he does not adopt *G*; the resulting state can be less useful for *B*, so the goal *G* is really an instrumental goal for *B* (if he wants to preserve the current state of affairs).

Second, the agent must be able to foresee the reaction of the *normative* agent (both) in case he fulfills the obligation and in case he doesn't. This ability of *anticipatory coordination* is another fundamental feature of social agents, according to [13]. In the field of BDI agents there are already some proposals for this form of reasoning. [28] introduced the notion of "anticipation feedback loop", [22] the "recursive modeling" of agents and [5] a planning framework for anticipatory coordination.

Note that this form of reasoning is borrowed from the field of *Game Theory*; however, for what concerns the treatment of obligations, we do not resort to what has been discussed about norms and obligations in *Game Theory*. In fact, as noticed by [17], the proposals coming from *Game Theory* seem to forget the role of sanctions in the normative reasoning.

Moreover, since we would like to let the agent free not to fulfill an obligation, we need some mechanism for enabling him to choose among the various alternatives. As in all the proposals mentioned above, we exploit the notion of *utility* developed in *Decision Theory* in order to choose the best alternative of the agent (depending on the reaction of the partner). Utility is the formalization of the notion of *preference* of persons: therefore, it is possible to express the fact that the reaction of the *normative* agent leads to a less

preferred state for the agent together with the fact that the agent achieves some utility by satisfying his own goals.

3.1 The Agent Model

An agent C is a 6-tuple $\{IB, G, f, L, KP, \varphi\}$ where:

- IB are the agent beliefs (concerning also the beliefs about the possible *normative* agents);
- G is the set of private goals of C ;
- f is the utility function of C (a function from states to real numbers); it is used to evaluate the outcomes of C 's actions. f applies to states expressed as sets of ground predicates (implicitly conjoined). It embodies a definition of the basic desirability degrees of each predicate and a combination function used to obtain the overall evaluation of the state.
- L is a set of tuples representing the obligations known by C of which he is the *bearer* (see next Section).⁶
- KP is the set of plan recipes which C knows. The plan recipes are defined below.
- φ is a planner able to select the plans that may be possibly be executed by an agent (the agent's *candidate plans*) in any situation.

KP is a rather standard set of plan recipes (see, for example, [10]). However, in order to simplify the evaluation of the foreseen utility of a plan, we have assumed that each plan built out of the recipes is a two-level plan. At the first level, there are *complex* actions, which involve a decomposition into *simple* actions (which appear at the second level). The simple actions cannot be decomposed further.⁷ So, KP is a pair $\langle CAct, BAct \rangle$. $CAct$ is a set of Complex Action schemata, each of which is a 5-tuple $CActSch = \{ActA, ActV, ActP, ActB, ActE\}$,⁸ where:

- $ActA$ are the Action Arguments,
- $ActV$ are the Action local Variables,
- $ActP$ are the Action Preconditions,
- $ActB$ is the Action Body, i.e. is a sequence of elements $BA_i \in BAct$.
- $ActE$ are the Action Effects: they are conditioned to the preconditions $ActP$, so that different outcomes (with different probabilities) may derive from the execution of the action in different situations (a detailed definition of action schemata appears in [4]).

⁶For the sake of brevity, here we limit ourselves to one *normative* agent.

⁷Actually, a plan may be composed of a single simple action.

⁸In this paper, we do not consider the possibility of an abstraction hierarchy among actions. But see [2].

In turn, each element of $BAct$ is a 5-tuple $BActSch = \{ActA, ActV, ActP, Proc, ActE\}$, i.e. a Basic Action schema, which is a schema where the body (decomposition) is replaced with an executable procedure (for instance to activate a sensing action, or to execute a transaction on a data base). They represent executable actions.

Since f considers the set of predicates, each of which is associated with a value (desirability degree), and, by means of a combination function, produces the overall desirability of a state on the basis of its description, it is clear that just the actions including effects that involve some predicates appearing in f can affect the evaluation of the state resulting from the execution of the action. So, just these predicates can provide an utility for the agent; the predicate may affect the utility of the resulting state in a positive or negative way; in particular, negative utilities are used for representing the costs of executing the action in terms of time and resource consumption.

In the following, we introduce *situated agents*. An agent, as defined above, includes general beliefs about the world, about what is good in it, and about what can be done in it. Now, what is needed is knowledge about the particular context where the agent has to move in a specific situation. A *situated agent* SA is a 5-tuple $SA = \{A, S, CG, CP, CI\}$ where:

- A is an agent, i.e. a tuple $\{IB, G, f, L, KP, \varphi\}$, as defined above,
- S is a set of beliefs about the current situation (i.e., a state),
- CG is a set of Current Goals, i.e. predicates to be possibly made true via an action executable in the current situation,
- CP is the set of candidate plans produced by the planner φ in the situation S (that is, CP is a set of *potential intentions* in the sense of [24]),
- CI is the Current Intention to execute a plan (a newly planned plan or the remaining part of the previous part).

We do not aim at providing a formal specification of an intention, but it may be observed that one of its main properties (according to [16]), i.e. the *persistence* of intentions, is achieved from a computational point of view by making φ take always into account the current intention (i.e. the previously chosen plan). In fact, φ continuously tries to expand or update the current plan, unless new information makes it believe that the intended goal has already been achieved, or it is not worth being achieved any more. Notice that the presence of utilities can lead φ to believe that a different (totally new) plan can enable the agent to reach higher utility. So, the previous plan is discarded.

Since the planning framework has been described elsewhere ([5], [4]), in this paper we focus on the criteria for the selection of the actions to execute in the current situation. In the same way we do not present here the agent architecture for the reactive execution of plans described in those papers.

In [5], a hierarchical decision theoretic planner is employed which is inspired to [25].⁹ In [7] we describe how the planner deals with obligations.

The planner takes as input goals consisting in states or actions: in case the goal is a state, it is considered as a state to be achieved, so that \wp must find all actions which can contribute to achieving the state; in case the goal is an action, \wp assumes that it is a complex action which needs to be executed, so that its (easier) task is to find all possible decompositions of (i.e. ways to carry out) the task. The latter activity is called *refinement* of the action.

In case no obligations exist (L is empty), the set CP is produced by \wp starting from the initial state S , and inspecting the KP to find all the recipes of actions which have as among their effects a predicate in CG and the recipes which refer to (expand) an action in CG . Then, on the basis of f , the possible alternatives are examined and the best one (P), which becomes the current intention of the agent, is chosen. The best plan is the one which maximizes the expected utility:

$$P = \text{Max}_{\{P_i \in CP\}} f(P_i(S))$$

where $P_i(S)$ is the state resulting from the execution of the plan P_i in the state S .

In [5], we have shown that in a multi-agent context, it is not sufficient to take into account the resulting state $P_i(S)$, but it is also necessary to consider the possible subsequent behavior of the other agents starting from $P_i(S)$. For instance, in a cooperative setting, it may happen that a state very positive for the agent endanger the activity of the partners, so that the overall (group) goal is harder to achieve. Our solution has been to base the evaluation not on $P_i(S)$, but on the states achievable from the partners starting from $P_i(S)$ (a kind of one-level lookahead in the spirit of min-max search).

In the next Section, we aim at showing that the same approach can be adopted to handle obligations, where the *partner*, in this case, is not a generic member of a group, but he is the agent in charge of checking that obligations are respected.

3.2 Formal Definition of Obligations

In the L component of an agent, an obligation Ω is represented as a 4-tuple $\{O, B, N, R\}$ where:

- O is the content of the obligation, i.e., the state or action goal which N wants to be adopted by B ,
- B is an agent who is called the *bearer* of the obligation,
- N is an agent called the *normative* agent,

⁹In [5], the planner prunes suboptimal plans during the refinement of non-primitive plans: therefore, the number of plans considered in CP is smaller than the set of possible primitive plans for achieving the set of goals.

- R is an action (called sanction) which N will presumably bring about in case he detects a violation of the obligation.

The content of the obligation Ω , O , is not necessarily a state (e.g., “the font of the submitted paper should be courier”), but it can be also an action where C is the agent (e.g., “the author should send a signed copyright form”) or not (e.g., “your children should go to school”). Finally, it can be the prescription of not executing an action: “you should not send the submitted paper to other conferences”.

It must be observed that in our multi-agent framework the behavior of the partners of an agent C is influenced by the actions of C , just insofar the effects of his actions can be noticed by the partners. In other words, any action can have a side-effect on the partners’ behavior just in case they are able to detect that something relevant for them has happened. This means that when C carries out his ‘lookahead’ he must start not from $P(S)$ (the resulting state how C sees it), but from $P(S')$, i.e. from the state S' that (according to C ’s knowledge) his partners will see.

This is particularly relevant in the case of obligations. In fact, the N counterparty of an agent C who is the *bearer* of an obligation cannot be assumed to become immediately acquainted with the (possible) violation of the obligation. According to C ’s knowledge, there is some probability that this happens: in fact C is assumed to know that N has some actions available to check the fulfillment of O , that these actions may fail, and that just in case of their success, N will consider (not necessarily decide) to apply the sanction.

Finally, any agent knows that any action may fail; also the action of applying the sanction may fail. So, even if the violation has been detected, and N has decided to apply the sanction (which he may not, in case the cost of applying it is greater than the gained utility), the sanctioning action may fail. C must (or, at least, we claim that rational agents do) weigh all of these possibilities when he chooses the best way of acting.

3.3 The Behavior of the Normative Agent

In general, the sanction is an action of the *normative* agent (e.g. check out the driver license of the violator of a norm), but it can also involve an action to be executed by the *bearer*. For instance, the sanction could be:

‘Request($N, C, \text{Pay}(C, \text{Money}, \text{State})$)’

It is up to the *normative* agent, however, to issue the request, i.e. to communicate to C the content of the request. Notice also that the (successful) execution of a request in a formal context has the effect to make true another obligation (“you must pay the sum of money to the state”). Therefore, the sanctioning of a violation results in the fact that a new obligation arises, which is treated again in the same way in the course of C ’s reasoning.

According to the model outlined above, the bearer C should foresee the possible reactions of N . As we have seen, there must occur some sensing action enabling N to detect the vi-

olation. If C assumes that this action succeeded, and so that N knows that a violation occurred, he must try to imagine which action N will do next. Although the sanctioning action is a possibility, C should take into account that N has to balance it against other alternatives. So, C must reason about the motivations of N for executing the sanction.

As for any other action, there are two factors that contribute to action choice.

First, since N is an agent, his model includes a utility function. So, if the predicates appearing in the effects of the sanction have a positive desirability degree for N , then N can select the sanction as his preferred action. This would mean that N can gain an advantage if the violation of the norm is sanctioned.¹⁰

But it is also possible that the sanction does not provide N with any personal utility. For instance, there is no utility for a policeman to sanction the breaking of a norm. In this case, the execution of R by N may be due to the existence of another norm, where the policeman acts as C and the local administration acts as N . In other words, it is a duty of the policeman to sanction a driver who parked outside the allowed areas: this is a duty established by the administration for which the policeman works and a sanction should be applied to the policeman in case he does not respect the norm.

The need of having some knowledge about the *normative* agent's utility function and goals is a strong requirement. However, some defaults can be applied. So that a set of definitions for the 'standard policeman', or the 'standard program committee' can be used. But in some cases, more detailed user models can be available as the 'policeman I meet everyday in front of my office', or the 'program committee of a prestigious workshop'.

3.4 The Deliberate Normative Agent Model

If L (the set of obligations) is not empty, then the planning phase and the selection process of the best alternative must be modified for two reasons:

- besides the built-in goals of the agent, there are other goals that must be examined (even if not necessarily satisfied): C should examine whether to satisfy the O component in the obligations in L (for the sake of brevity, only one obligation will be assumed in the following).
- the agent knows that the world resulting after his action is then modified by the reaction of the *normative* agent: the expected utility must be evaluated after the reaction (if there is one).

The first modification is that the planning phase must be given as input not only the goals in CG together with the

¹⁰The advantage gained by sanctioning should be justified by a more sophisticated form of reasoning: the *normative* agent has posed the norm for achieving a state, e.g., that the taxes are paid by everyone; such a state is useful for her or for the community: the respect of the norm provides an indirect utility since it is a means for achieving the desired state.

current intentions CI , but also a different alternative CG' which consists in the union of $CG \cup CI$ together with the goal(s) O of the obligation(s) in L .¹¹

As we said in Section 3.1, the planner takes in input both state and action goals. The difference is that in case of state-goals, the actions which can achieve the goal are selected and passed as the real input of the planner; on the other hand, in case of action goals, these are added directly to the plans identified for satisfying the goals in CG . Instead, in case of negations of actions, we have chosen a different strategy: if an action which is forbidden by the obligation is inserted in a plan (while planning how to achieve goals from CG') during the planning phase it is canceled from the plan, leaving a possibly incomplete plan whose (in)utility will be computed at the end of the planning process (remember that the elimination of steps occurs only in the plans deriving from CG' , so that a copy of the complete plan is examined by the planner anyway).

The resulting CP will be the union of the results of planning a solution for CG and then CG' .

The second modification concerns the selection of the plan to be executed among the P in CP . It is not sufficient to consider the utility of the resulting state $P(S)$ since the reaction of the *normative* agent N must be simulated first.

N is modeled as an agent $\{IB', G', f', L', KP', \varphi\}$, the sanctioning procedure R of Ω in L is a goal of N in G' (together with other goals known by C), f is the (presumed) utility function, O' may be empty. N may have the same knowledge KP' about plans as C or not.

N is situated by creating $\{N, S'', CG', CP', CI'\}$. Given S' , the initial state from N 's point of view (according to C 's beliefs S): the set of intentions CI' must be created by N (in C 's simulation) by planning how to achieve the goals G' not from state S' , but from the state S''_P which follows the execution of each plan P in the set of candidate plans CP . We will call S''_P the state concerning what N believes about a state $P(S)$: only the effects of the plan P which affect N 's beliefs (according to the definition of P in the knowledge base KP) are considered. State S''_P is created by propagating from S' those propositions which are not affected (in N 's beliefs) by the plan P .¹²

Therefore, for each plan P in the set of candidate plans CP , given the situated agent $\{\{S''_P, G', f', L, KP', \varphi\}, S'', CG', CP', CI'\}$, the set of current intentions CI' is produced by means of the planner φ , with inputs S''_P and CG' (the set of current goals).

¹¹The union of G with the set of the powerset of the O in L in the general case.

¹²Note that we assume that C and N have the same knowledge about the very initial situation. This is clearly a simplification since the topic of belief revision in a multi-agent setting is not the focus of the article. For a more sophisticated framework for reasoning about other agents' belief change see [26].

By applying the formula:

$$CI' = P_P^{best} = \text{Max}_{\{P'_P \in CP'_P\}} f'(P'_P(S''_P))$$

the reaction of N in each situation $P(S)$ is computed: $P_P^{best}(P(S))$ will be the real outcome of plan P of C , that is, the state containing the possible sanction for his behavior (recall that the different P s in the set of candidate plans CP are plans which may or may not fulfill the obligation of C together with achieving his own goals).

As we said in Section 3.1, N will select a plan P for sanctioning C only if it is rational for him to do so (P has a greater utility for him than other options).

C will select the plan P^{best} in the set of candidate plans CP such that:

$$P^{best} = \text{Max}_{\{P \in CP\}} f(P^{best}(P(S)))$$

where P_P^{best} is the plan selected by N when C executes P (note that the plan P is executed from state S instead of S''_P , since S is C 's point of view).

A further modification is needed when actions may have non-deterministic outcomes. In this case, $P(S)$ is a set of states with associated probabilities.

When N plans her reaction, she will be in a specific state of $P(S)$ (since C will have already executed the action he chose). Therefore, C has to simulate N 's reaction in each of these states. In this situation, P_P^{best} will be a set of (*state, probability, plan*) tuples (the probability is the one of the state in $P(S)$ from which the associated plan has been planned); the above formula must be modified in:¹³

$$P^{best} = \text{Max}_{\{P \in CP\}} \sum_{(S_i, p, P^N) \in P_P^{best}} p * f(P^N(S_i))$$

Note that the described framework does not model the fact that N , as the modeled agent C does, may examine the future reactions of other agents. It is possible to extend the theoretic model and the corresponding implementation by allowing a further level of recursion (C considers that N considers the subsequent reaction of C or some other agent). But, as noticed also by [22], recursion must be blocked somewhere since the resources of the planning agent are limited. A possible application of a further level of recursion is the modeling of nested obligations.

As an example, take the obligation of a policeman discussed in Section 3.

That situation would be modeled by means of an obligation Ω_1 , where the *bearer* is our agent and N is the policeman; O would be not to park and R the action of checking and

sanctioning C ; R is included in the KP' of N .¹⁴ In turn, N is modeled as an agent where L' includes the obligation Ω_2 . The *normative* agent of N is the administration, and O' is Ω_1 ; R' would be a suitable action of the administration for checking the policeman and sanctioning him.

4. WHY TO FULFILL AN OBLIGATION

The bearer of an obligation has to decide whether to (try to) fulfill the obligation: that is, he has to decide whether it is worth adopting a plan in CP which derives from the planning of the goals $G \cup CI \cup \{O\}$. As described above, he will select P^{best} according to the utility of the state following the reaction of the partner; in this way, no direct utility is (in general) achieved from the fulfillment of the obligation (rather, he would get just costs), but a state where the obligation is fulfilled may have a greater utility for the agent, due to the sanction effect. Therefore, he will possibly choose the plan which also includes the fulfillment obligation. This decision, however, is a trade off between the cost (in terms of time or resources consumed) of doing something for achieving the obligation (plus the cost of postponing his own goals), and the effect of the reaction of the *normative* agent.

The trade off of costs and sanctions is only one of the factors which can lead an agent not to do anything for the obligation. As appears in the definition, the *normative* agent has to check whether the obligation has been fulfilled before applying the sanction. But, checking the fulfillment and applying the sanction have a cost for her, so she may decide not to do anything.¹⁵ Finally, the action of checking the fulfillment may fail with a certain probability. In this case, the decrease in the final utility due to the sanction must be weighed according of the probability of success of the *normative* agent (if she fails to discover the violations, she cannot apply the sanction).

The last observation opens the way to a different possible way for avoiding the sanction while not respecting the obligation: the *bearer* of the obligation may do something for misleading the *normative* agent in his task of checking the fulfillment of the obligation or for making the sanction impossible to be applied.

In other words, the *bearer* C can make the *normative* agent N believe that he has fulfilled the obligation or that he is not liable.

As noticed by [12] it is not sufficient that an obligation is fulfilled only in a subjective manner (as, e.g., [21] proposes). That is, the satisfaction of an obligation is not defined in terms of what N believes. He notices, in fact, that the *normative* agent could discover the violation later or a third party may be aware of the violation.

In fact, in our agent model, the acknowledgement that an obligation is fulfilled depends on the the beliefs of C that specify whether it is fulfilled in the real world (according to C 's belief space) and in the beliefs of the normative agent N .

¹⁴In this situation, the *defender* of the obligation is different from the *sovereign* who issued it, in [17]'s terminology.

¹⁵For a public administration, checking fiscal evasion has sometimes a cost which does not cover the returns gained from the payment of monetary sanctions.

¹³For the details of the planning algorithm see [4].

So it is possible that C believes that O has not been fulfilled, and, at the same time, that he believes that N does not (or will not or cannot) realize that fact, or, viceversa, that he believes that O has in fact been fulfilled but that N could fail to realize that. Finally, C may believe that he fulfilled the obligation, while it is not actually the case.

Our definition does not give such a subjective notion of satisfaction: we are just pointing out that an agent has this opportunity if there is no third party or the *normative* agent has no other way of subsequently checking the fulfillment.

Symmetrically, if N is not aware that the obligation has been fulfilled, she may apply the sanction anyway; therefore, the *bearer*, besides fulfilling the obligation (from the objective point of view), has to make the *normative* agent aware of this fact.

For what concerns the first issue, assume that, in the example of Section 2, the SmartAgents2000 PC agent has only one way for checking the obligations of authors: she sends the title of the submitted papers to other conferences and receives a response from them. The malicious author may have a plan for submitting articles which ensures that each submission has a slightly different title. In this way, he will not be sanctioned since the PC agent will fail (perhaps with a certain probability) to detect the multiple submission.¹⁶

In summary, there are various motivations for an agent to decide not to fulfill an obligation Ω :

1. The agent has adopted the obligation but he cannot do anything for it (i.e., he has no feasible plan in KB).
2. The possible plans which include some actions for fulfilling Ω achieve a lower utility than some other plan (due to the cost of fulfilling the obligation). In particular, this may happen if some of the actions do not ensure that the *normative* agent becomes aware of the fulfillment so that she will probably apply the sanction anyway (decreasing furthermore the utility of the plan).
3. There is some plan which does not fulfill the obligation but which induces the *normative* agent to believe otherwise.
4. There is some plan which does not fulfill the obligation but which make the sanction impossible to be applied.
5. The *bearer* of the obligation can bribe the *normative* agent so that he does not apply the sanction.

Obligations have been discussed in the field of multi-agent systems mostly in order to build agents which respect a certain behavior. Hence, the analysis of the possible deviations from the norm seems at first sight misleading. Instead, there are a number of reasons for the present work. First of all,

¹⁶ A further interesting problem to analyse is how to devise these misleading plans: note that they can be built only by including in the definition of norms the knowledge about how the *normative* agent carries out checking.

obligations must be distinguished from other propositional attitudes as goals and intentions. If the only possible deviation were of kind 1 an obligation would be similar to an intention (as happens, e.g., in [20]). Second, as [27] notices, there could be cases of “wrong” obligations which the agent designer would like to avoid [9].

But, most importantly, possible deviations should be analysed in order to let agents reason about the behavior of other agents (and human users), which are not necessarily built to respect obligations. In particular, in some domains, agents must be able to judge whether the other ones are *trusted* and maintain obligations concerning *security* and *privacy* (see [15]).

If agents who respect (if they can) obligations are needed, there are some ways to enforce the fulfillment of norms:

- The content of a certain obligation Ω can occur also as a preference of the agent: in this way, when adopted, it becomes similar to an intention (reinforced by the possible sanction): the agent directly achieves an utility from the satisfaction of the obliged state (the content of the obligation is a *value* for the agent.)
- The agent may have the preference not to mislead the *normative* agent: the former agent does not do anything to induce false beliefs in the *normative* agent, e.g., that the obligation is fulfilled when it is not the case.
In this way, the agent does not exploit the possibility described at point 3 and 4 above.
- The agent has some *social goal* which makes him not prefer situations where he is liable (for example because he does not want that other agents decrease the trust they have on him).

5. DISCUSSION

For what concerns the classification of norm-abiding systems proposed in [17], our framework is classifiable as with *built-in obligations*. In fact:

- *reliability*: the agent is not as reliable as in a model where norms are treated as constraints, but the goals deriving from norms are more persistent than standard ones, since the sanction is computed in the evaluation of the agent's utility. On the other hand, the autonomy of the agent is increased.
- *learning*: due to the declarative representation of norms, they can be acquired and discharged while the system is on-line.
- *novelty*: both prescriptions and prohibitions may be the content of norms.
- *repair*: since norms are treated as any other goals, they are subject to the standard (reactive) planning process.
- *social control*: that is, the *bearer* agents should be interested on the monitoring of the respect of the norms they follow; The ability of reasoning about what other agents do is a necessary precondition for enabling *social control*.

As we show in [6], in our model obligations can arise even if not explicitly stated, but as a result of common knowledge and social goals. In that work, obligations arise as a result of speech acts as requests and questions in non-cooperative contexts. In order not to offend the requester (a social goal), the requestee is led to adopt the requested goal if it is easy to accomplish (say, showing understanding or telling the time).

Speech acts are modeled as having the standard illocutionary effect (e.g., making mutually believed believed the illocutive purpose) in both cooperative and non-cooperative situations. The illocutionary effects of speech acts make mutually known between the speaker A and the addressee B that A wants G to be adopted by B (e.g., $G = \text{Inform}(B, A, \text{time})$). Moreover, it is mutually known that refusing ‘simple’ requests may be offensive for the requester: the offense of the requester plays the role of the sanction if B has the *social goal* not to offend anyone (i.e., he prefers states where A is not offended, other things being equal). Moreover, the offended agent may express public protest and receive the support and consent of the community. The sanction R , from A ’s point of view corresponds to the action of interpreting and evaluating negatively the reply of B . [29] show how the interpretation process can affect not only the beliefs but also the attitudes of speakers.

Recall that, as stated in Section 3.4, the bearer of the obligation compares both the plans which do something for the requester and those where he goes on with his activity without changing his behavior. Therefore, after a request the addressee compares the result of replying to the partner with the result of ignoring him.

The strict relation between respecting obligations and other behaviors like fulfilling requests has been highlighted by [17]:

Norms, in other words, are but a special case of a general social law: in order for autonomous agents to accept others’ requests (including normative ones) they ought to find some convenience for doing so. That such a convenience coincides or not with the request’s reasons and objectives is irrelevant.

6. RELATED WORK

Since [33] *deontic logic* has been proposed as a formalism for reasoning about obligations, normative concepts and what “should be” (or happen) in the world. The main assumption in most proposals is that the verbs as “ought”, “should” can be modeled in the same way as other modalities as necessity or belief by means of a possible world framework. Modals operators as O have been introduced in order to express formulas as Op which are true in a world w if the proposition p is true in a world in all the ‘ideal’ (possible) worlds which are accessible from w . The ideal worlds represent how the reality should be according to some normative system or preferences.

However, the aim of deontic logic is different from the way obligations are used in agent theories: the main goal of the former is to examine how obligations follow from each other and which are the paradoxes of deontic reasoning (see [32]). In contrast, agent theories aim more at examining the relationship between intentions and obligations, i.e. how the

agents decide or not to fulfill an obligation. In our work, this includes reasoning about the application of sanctions and, also, about how to avoid them.

While deontic logic has devoted attention to the possibility of violating norms, less attention has been addressed to the role of sanctions, even if in one of the first works about obligations, [1], they are reduced to the alethic modality of necessity via the idea of the occurrence of a sanction s :¹⁷

$$Op \equiv NEC(\neg p \supset s)$$

Moreover, one of the main goals of agent theories is to model why agents follow or violate norms and when it is rational to do so; in deontic logic terms, we do not want the T axiom ($Op \supset p$, which holds for the *necessity* and *knowledge* operators).

In recent works on deontic logic, the importance of taking into account sanctions has been explicitly recognized:

The threat of punishment might be taken into account when the agent designer considers building into his agent the capability of adhering [to obligations]. [...] When a rule is violated, and the violation is detected, a sanctioning act (or an act of recovery) is effectuated.¹⁸

In [19], deontic logic is applied in an agent framework for dealing with norms and conventions. This work explicitly models sanctions consequent to violations and relates the fulfillment of obligations to preferences.

Moreover, recent works as [11] have addressed interesting issues in a deontic logic framework as the management of norms in case of collective agency.

For what concerns agent theories, the notion of obligation has been exploited for the goal of directing the behavior of agents; as an example, in [30] (as well as in similar approaches) there is a different view of obligations, as [27] has noticed: in [30] obligations are used for *regimenting* agents, that is, for assuring that they will behave in a certain way. Because of this goal, the actions of the agent repertoire are constrained by the norms and the axiom T is adopted for modeling obligations, and obligations are constrained to be consistent. In a similar way, [8] proposes to constrain the evaluation module for enforcing norms.

On the contrary, our approach leans towards another view of obligations which is inspired to [18], where obligations can be violated, *normative* agents can be deceived in order to avoid sanctions, and the fulfillment is motivated by some instrumental relation with some goal or preference. The main difference with [18] is in the role given to the recursive modeling of the *normative* agent, a difference which is more apparent in [14] where an implementation with the DESIRE agent architecture is proposed. In our work too, obligations

¹⁷ s should be better defined as *liability* since a sanction does not necessarily occurs, as noticed by [34].

¹⁸ [27], p. 163.

lead to goal adoption; but here, those goals becomes intentions only after the evaluation of the effects of the agent's alternatives, obtained via the recursive modeling of the reaction of the normative agent.

On the other hand, with respect to [18] we do not consider here the problem of accepting a norm as such.

7. CONCLUSION

Our proposal constitutes a step forward in the understanding of deontic reasoning in that we include in the decision process the prediction of the *normative* agent's autonomous behavior. This is the basis not only for discovering when it does not worth to fulfill an obligation, but also for enabling agents to reason about how to deceive the *normative* agent. Predicting the possible failures and deceptions of obligations is fundamental for building agent communities regulated by norms.

Finally, we used the reasoning process involving the prediction of the behavior of other agents for modeling cooperation among agents ([5]) and for modeling dialog ([6]); this form of reasoning is becoming a widespread methodology in the multi-agent field, as works like [22] witness.

In [7], the details and limitations of the planning process underlying this framework are discussed, while the phenomenon of deceptions for avoiding the fulfillment of obligation is the topic of the ongoing work.

8. REFERENCES

- [1] A. Anderson. The logic of norms. *Logic et analyse*, 2, 1958.
- [2] L. Ardissono, G. Boella, and L. Lesmo. Plan based agent architecture for interpreting natural language dialogue. *International Journal of Human-Computer Studies*, (52):583–636, 2000.
- [3] I. Asimov. *The naked sun*. Garden City, New York, 1957.
- [4] G. Boella. *Cooperation among economically rational agents*. PhD thesis, Università di Torino, Italy, 2000.
- [5] G. Boella, R. Damiano, and L. Lesmo. Cooperation and group utility. In N. Jennings and Y. Lespérance, editors, *Intelligent Agents VI — Proceedings of the Sixth International Workshop on Agent Theories, Architectures, and Languages (ATAL-99, Orlando FL)*. Springer-Verlag, Berlin, 1999.
- [6] G. Boella, R. Damiano, L. Lesmo, and L. Ardissono. Conversational cooperation: the leading role of intentions. In *Amstelogue'99 Workshop on Dialogue*, Amsterdam, 1999.
- [7] G. Boella and L. Lesmo. A rational treatment of obligations in intelligent agents. Submitted for conference review.
- [8] M. Boman. Norms as constraints on real-time autonomous agent action. In *Proc. of MAAMAW'97*, pages 36–44, Berlin, 1997. Springer Verlag.
- [9] W. Briggs and D. Cook. Flexible social laws. In *Proc. of 14th IJCAI'95*, pages 688–693, Montreal, 1995.
- [10] S. Carberry. *Plan Recognition in Natural Language Dialogue*. MIT Press, 1990.
- [11] J. Carmo and O. Pacheco. Deontic and action logics for collective agency and roles. In R. Demolombe and R. Hilpinen, editors, *Proc. Fifth International Workshop on Deontic Logic in Computer Science (DEON'00)*, pages 93–124, ONERA-DGA, 2000.
- [12] C. Castelfranchi. Commitment: from intentions to groups and organizations. In *Proc. of ICMAS-96*, Cambridge (MA), 1996. AAAI/MIT Press.
- [13] C. Castelfranchi. Modeling social action for AI agents. *Artificial Intelligence*, 103:157–182, 1998.
- [14] C. Castelfranchi, F. Dignum, C. M. Jonker, and J. Treur. Deliberate normative agents: Principles and architecture. In N. Jennings and Y. Lespérance, editors, *Intelligent Agents VI — Proceedings of the Sixth International Workshop on Agent Theories, Architectures, and Languages (ATAL-99)*, Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin, 2000.
- [15] C. Castelfranchi and R. Falcone. Principles of trust for mas: Cognitive anatomy, social importance, and quantification. In *Proc. of ICMAS'98*. 1998.
- [16] P. Cohen and H. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
- [17] R. Conte and C. Castelfranchi. *Cognitive and Social Action*. UCL Press, 1995.
- [18] R. Conte, C. Castelfranchi, and F. Dignum. Autonomous norm-acceptance. In J. P. Mueller, M. Singh, and A. Rao, editors, *Intelligent Agents V — Proc. of 5th Int. Workshop on Agent Theories, Architectures, and Languages (ATAL-98)*. Springer Verlag, Berlin, 1998.
- [19] F. Dignum. Autonomous agents and social norms. In *ICMAS'96 Workshop on Norms, Obligations and Conventions*, 1996.
- [20] F. Dignum, J.-J. Meyer, R. Wieringa, and R. Kuiper. A modal approach to intentions, commitments and obligations: intention plus commitment yields obligation. In *Proc. of DEON'96*, 1996.
- [21] R. E. Fikes. A commitment-based framework for describing informal cooperative work. *Cognitive Science*, 6:331–347, 1982.
- [22] P. J. Gmytrasiewicz and E. H. Durfee. Formalization of recursive modeling. In *Proc. of first ICMAS-95*, 1995.
- [23] P. J. Gmytrasiewicz and E. H. Durfee. Rational interaction in multiagent environments: Communication. In *Submitted for publication*, available at <http://www-cse.uta.edu/~piotr/www/piotr.html>, 1997.
- [24] B. Grosz and S. Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

- [25] P. Haddawy and S. Hanks. Utility models for goal-directed, decision-theoretic planners. *Computational Intelligence*, 14:392–429, 1998.
- [26] H. Isozaki and H. Katsuno. Observability-based nested belief computation for multiagent systems and its formalization. In N. Jennings and Y. Lespérance, editors, *Intelligent Agents VI — Proceedings of the Sixth International Workshop on Agent Theories, Architectures, and Languages (ATAL-99)*, Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin, 2000.
- [27] C. Krogh. *Normative Structures in Natural and Artificial Systems*. Complex, TANO, Oslo, 1997.
- [28] A. Ndiaye and A. Jameson. Predictive role taking in dialog: global anticipation feedback based on transmutability. In *Proc. 5th Int. Conf. on User Modeling*, pages 137–144, Kailua-Kona, Hawaii, 1996.
- [29] D. Pautler and A. Quilici. A computational model of social perlocutions. In *Proc. 36th Conf. of ACL*, pages 1020–1026, Montreal, 1998.
- [30] Y. Shoham. Agent-oriented programming. *Artificial Intelligence*, 60(1):51–92, 1993.
- [31] D. Traum and J. Allen. Discourse obligations in dialogue processing. In *Proc. 32nd Annual Meeting of ACL*, pages 1–8, Las Cruces, New Mexico, 1994.
- [32] L. van der Torre and Y. Tan. Contrary-to-duty reasoning with preference-based dyadic obligations. *Annals of Mathematics and Artificial Intelligence*, 2000.
- [33] G. von Wright. Deontic logic. *Mind*, 60:1–15, 1950.
- [34] G. von Wright. A new system of deontic logic. In R. Hilpinen, editor, *Deontic Logic*, pages 105–120. D. Reidel, Dordrecht-Holland, 1971.
- [35] P. Xuan and V. R. Lesser. Incorporating uncertainty in agent commitments. In N. Jennings and Y. Lespérance, editors, *Intelligent Agents VI — Proceedings of the Sixth International Workshop on Agent Theories, Architectures, and Languages (ATAL-99)*, Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin, 2000.