# Obligations and Permissions as Mental Entities

**How cognitive agents reason about what is permitted and what counts as a violation**

**Guido Boella**
Dipartimento di Informatica - Università di Torino
Cso Svizzera 185
Torino 10149, Italy
guido@di.unito.it

**Leendert van der Torre**
SEN-3 - CWI Amsterdam
P.O. Box 94079, NL-1090 GB
Amsterdam, The Netherlands
torre@cwi.nl

## Abstract

Cognitive agents must have an explicit representation of their beliefs, desires and goals, and also a 'theory of mind' of the other agents. In this paper we propose a cognitive agent model where norms are based on three dimensions. First, we distinguish between agents whose behavior is governed by norms and autonomous normative systems; we call the latter normative agents. Second, we distinguish some of the usual mental attitudes for all agents. Third, we distinguish between behavior that counts as a violation, and sanctions that are applied. The decisions of agents are based on a qualitative decision theory extended with recursive modelling: an agent explicitly models the normative agent who monitors violations and applies sanctions. Our framework enables agents to reason also on what is permitted and on how permissions work as exceptions to behaviors which are considered violations and thus are punishable.

## 1 Introduction

Agents that contain explicit representations for goals, intentions and beliefs are called *cognitive agents* [Conte *et al.*, 1998]. Norms contribute to fill the gap between individual cognitive agents and multiagent societies - i.e. the gap between cognitive and social concepts. In this paper we present a definition of norm based on the mental attitudes of cognitive agents: their beliefs and goals and, most importantly, their beliefs about the beliefs and goals of other agents. Thus, the concept of norm presupposes cognitive agents and multiagent interaction: "[agents] must have a 'theory of mind', an 'intentional stance' towards the others. In other words *they must be able to reason about the beliefs, the goals, the intentions, and the normative reasoning of the others*" [Castelfranchi *et al.*, 1999].

In this paper we consider not only *imperative norms*, i.e., obligations, but also *permissive norms*. In most works, permissions are considered a dependent normative category, since they would amount simply to the negation of some obligation. In this paper we consider permissions as irreducible to obligations, even if, in some sense, we show that permissions have a meaning only in the context of an obligation.

In [Boella and van der Torre, 2003a], we proposed a model of norms based on the three ideas:

**Agent dimension.** Inspired by [Boella and Lesmo, 2002], we distinguish between the agent A who is the bearer of the obligation and the normative agent N. We attribute mental states to the normative agent (thus taking an *intentional stance* [Dennett, 1987] towards normative systems). Roughly, the content of an obligation is a goal of the normative agent.

**Mental attitudes dimension.** We distinguish between the agents' abilities, their beliefs and their motivations (goals and desires). We assume that agent A adopts the goal of agent N as one of its possible options, where adoption is "having a state of affairs as a goal *because* another agent has the same state as a goal" [Castelfranchi, 1998].

**Obligation dimension.** We distinguish between behavior which counts as a violation, and behavior that is sanctioned. The "counts as" relation is the same proposed by [Searle, 1995] in his *construction of social reality*.

The research questions we address in this paper are:

- Which are the elements of a cognitive agent that are necessary to reason about obligations? And in particular, how can a mentalistic definition of obligation be extended to reason also about permissions?
- Which is the mental state of a normative agent who enacted a permission which behaves like an exception to some obligation it previously enacted?
- Can permissions be defined in the same way as obligations? Are permissions goals of the normative agent in the same way as obligations are?

In Section 2 we introduce our definition of obligation and the elements of a cognitive agent for reasoning about obligations. Then in Section 3 we discuss the role of permissions in a normative system and we propose a definition in terms of non punishability. These definitions are formalized in Section 4 where we present our cognitive agent based on a qualitative decision theory extended with recursive modelling. We discuss the behavior of an agent subject to obligations and permissions with examples in Section 5.

## 2 Reasoning about obligations

The sociologist Goffman argues that taking the other agents' actions into consideration is unavoidable, and call this situation "strategic interaction":

> "When an agent considers which course of action to follow, before he takes a decision, he depicts in his mind the consequences of his action for the other involved agents, their likely reaction, and the influence of this reaction on his own welfare" [Goffman, 1970, p.12].

Moreover, Goffman gives a game-theoretic interpretation of obligations. Norms produce a form of strategic interaction where the agent and the normative system are the players.

In the context of a game including norms, the outcome depends on whether the agent's behavior is considered as a violation and it is sanctioned by the normative agent.

However, decision and game theory have been criticized for their assumptions of ideality. Several alternatives have been proposed that take the limited or bounded rationality of decision makers into account. For example, Simon and others develop theories in artificial intelligence and agent theory replace probabilities and utilities by informational (knowledge, belief) and motivational attitudes (goal, desire), and the decision rule by a process of deliberation. Bratman further extends such theories with intentions for sequential decisions and norms for multiagent decision making. Finally, Gmytrasiewitcz and Durfee replace in game theory the equilibria analysis by recursive modelling:

> "Recursive modelling method views a multi agent situation from the perspective of an agent that is individually trying to decide what physical and/or communicative actions it should take right now." [Gmytrasiewicz and Durfee, 1995]

Moreover, [Gmytrasiewicz and Durfee, 1995] consider the cognitive limitations of agents in realistic settings such as acquiring knowledge and reasoning so that an agent can only build a finite nesting of models about other agents' decisions.

In the cognitive agent model we propose, an agent does not only have a representation of the other agents' mental states, and in particular of the normative agent, but it also uses this representation to reason about which decisions they will take. So our agent will recursively model the decision of the other agents: the same decision process employed to take its own decision is used by the agent to predict the other agents' reactions.

In a different context, [Conte *et al.*, 1998] support a similar view: "autonomous normative agents not only have minds (goals, beliefs, intentions, et cetera) but they have some explicit representation of the other agents' mental concepts".

Since a cognitive agent model is based on a symbolic representation of mental attitudes, instead of representing preferences by means of a quantitative utility function, we base our model on a qualitative decision theory inspired on [Broersen *et al.*, 2002a]. The representation of beliefs and motivational attitudes is homogeneous and based on conditional rules. For what concerns the reasoning on beliefs, our agent is based on conditional rules extended with a priority relation to resolve conflicts among them. In contrast, decisions are taken on the basis of which desires and goals remain unsatisfied and which ones have the priority.

Our definition of obligation is inspired to [Boella and Lesmo, 2002]:

> "An obligation holds when there is an agent A, the *normative* agent, who has a goal that another (or more than one) agent B, the *bearer* agent, satisfy a goal G and who, in case he knows that the agent B has not adopted the goal G, can decide to perform an action Act which (negatively) affects some aspect of the world which (presumably) interests B. Both agents know these facts." [Boella and Lesmo, 2002, p.496]

The definition is extended as follows by explicitly taking into account the conditional desires and goals of the bearer agent A and of the normative agent N. Moreover, in the context of a cognitive agent, we consider only the perspective of an agent who believes to be obliged: a norm is considered from the perspective of an individual agent's cognitive state who attributes certain mental states to other agents.

Agent A believes to be obliged to do $a$ by a norm $n$ issued by agent N if:

- Agent A believes that agent N desires and has as a goal that $a$.
- Agent A believes that agent N desires that there is no violation $\neg V(\neg a)$, but if agent N believes $\neg a$ then it has the goal and the desire that $V(\neg a)$: $\neg a$ counts as a violation.
- Agent A believes that agent N desires not to sanction $\neg s$, but if agent N decides $V(\neg a)$ then it desires and has as a goal that it sanctions agent A with $s$. Agent N only sanctions in case of violations: the normative agent is not a sadist that sanctions anyway.
- Agent A has the desire for $\neg s$: the agent does not like the sanction.

This definition does not presuppose that agents always stick to the obligations they are subject to. There are many reasons why agents should be able to reason about norm violation. Some of them are listed by [Castelfranchi *et al.*, 1999]:

> "[N]obody can avoid that norms - and in particular their instances - might be incoherent. There might be conflicts, and the deliberate agents should be able to manage these conflicts. Norms also cannot predict and successfully frame all possible circumstances. There might be some important event or fact to be handled, where no norm applies or some norm applies with bad results."

[Boella and Lesmo, 2002] give the following non-exhaustive list of situations where an agent may violate an obligation ($\Omega$ stands for an obligation):

1. There is some plan which does not fulfill $\Omega$ but which make the sanction impossible to apply.

2. The bearer may choose a plan which misleads the normative agent so that she selects a sanction which she believes can be applied, whereas, as a matter of fact, it cannot be applied.
3. The bearer of the obligation can bribe (or menace) the normative agent so that she does not apply the sanction.

No agent could be said to be really autonomous if it is not able to decide whether to respect or violate norms. However, as [Castelfranchi, 2000] argued, an agent should fulfill an obligation because it is an obligation, not because there is a sanction associated with it:

> "True norms are aimed in fact at the internal control by the addressee itself as a cognitive deliberative agent, able to understand a norm as such and adopt it. [...] The use of external control and sanction is only a sub-ideal situation and obligation." [Castelfranchi, 2000]

As a response to this criticism, we use the distinction between violations and sanctions – the third dimension of our classification in Section 1 – to distinguish between the agent's interpretation of the obligation, and its personal characteristics or agent type. The agent types are inspired by the use of agent types in the goal generation components of Broersen et al.'s BOID architecture [Broersen *et al.*, 2002a]. *Roughly*, we distinguish between *respectful* agents that are motivated by what counts as and does not count as a violation, and *selfish* agents that are motivated by sanctions only. An obligation without a sanction *should* be fulfilled, as Castelfranchi argues; but if fulfilling the obligations has a cost then it *is* only fulfilled by respectful agents, not by selfish agents, unless some incentives are provided or the agents dislike some social consequences of the violations. A selfish agent fulfills its obligations due to fear of consequences, whereas a respectful agent fulfills its obligations due to the existence of the obligation. Respectful agents not only accept norms but they *internalize norms* [Verhagen, 1999].

Furthermore, Castelfranchi argues that sanctions are only one of the means which motivate agents to respect obligations, besides "pro-active actions, prevention from deviation and reinforcement of correct behavior, and then also 'positive sanctions', social approval" [Castelfranchi, 2000]. In response to this criticism, we also show that variants of the definition can be used for other types of obligations, such as reward-based obligations.

## 3 Permissions

In this paper, we consider not only imperative norms, i.e., obligations, but also permissive norms. In most approaches, permissions are considered a dependent normative category, since they would amount simply to the negation of obligations: $P(q)$ iff $\neg O(\neg q)$. These are called *weak permissions* and it seems that there is no role for *strong permissions* which are the content of an explicit permissive norm.

But some authors have argued that this reduction is not acceptable. As concerns permissions in informal settings - 'practical permissions' - [Castelfranchi, 1997] argues that permissions are not simply the absence of obligations. In fact, permissions presuppose the dependency of the agent who asks the permission from the other agent who can give the permission.

Some legal scholars argue that permissive norms play a role in a normative system. For example, [Bobbio, 1980] says that "the difference between weak and strong permission becomes clear when we think of the function of permissive norms. Permissive norms are subsidiary norms: subsidiary in that their existence presupposes the existence of imperative norms [...] a permissive norms is necessary when we have to repeal a preceding imperative norm or to derogate to it", p. 891-892.[1]

In this work the focus is on formal permissions, or *permissive norms* issued by a normative agent. We show that in a normative system a meaningful permission "presupposes" an obligation it is an exception of; this obligation can be an existing one or an obligation which could be later added by some competent authority. So, the similarity with [Castelfranchi, 1997]'s approach is based on the analogy between being dependent on another agent and being subject to an obligation of the normative system. In an informal setting an agent has to be permitted to achieve its goal if there is an agent who has the power to prevent it to do what he asked; in a formal setting, an agent has to be permitted to do $a$ if there is some legal authority which has put or has the power to put an obligation not to do $a$.

Surprisingly, the definition of permission in our setting is simple. Agent A believes to be permitted to do $a$ under some condition $q$, if, when the normative agent believes that the condition $q$ holds, it has a desire and goal not to consider $a$ as a violation ($\neg V(a)$). So if the condition holds and the permission is used, the normative agent does not want to consider what is permitted as a violation, and consequently, as a behavior that deserves a sanction (or as a behavior that does not deserve a reward).

While an obligation to do $a$ implies that $a$ is a goal of the normative agent, a permission to do $a$ does not imply that $a$ is among its goals: otherwise, the agent could consider to adopt such a goal among its owns as it does in case of obligations. Rather, if a permission to do $a$ is an exception to an obligation not to do $a$, the permission does not cancel the goal of the normative agent that the agent does not do $a$. This relation between permissions and obligations is shown in the following example: it is forbidden to kill (i.e., it is obligatory not to kill), but it is permitted to kill in self defence. Such a permission does not allow to think that agent N's goal that nobody is killed is cancelled in a situation of self defence: in case of self defense, such a goal remains unsatisfied, but the murder is not considered as a violation.

Finally, this model of permission reflects the way in which permissions are enacted in formal laws. For example, the Italian penal code describes self defence exactly as an immunity from being considered a violator who should be sanctioned: "It is not punishable who has committed a crime compelled by the need to defend his own rights."[2]

[1] Authors' translation from Italian.

[2] "Non è punibile chi ha commesso il fatto per esservi stato costretto dalla necessità di difendere un diritto proprio ...", art.52.

Note that given a permission to do $a$, the goal that a behavior does not count as a violation ($\neg V(a)$) conflicts with the goal that the same behavior counts as a violation ($V(a)$) which is implied by the definition of a corresponding obligation not to $a$. How does the normative system decide between the two alternative goals? Legal theories say that some laws are more basic than other; the priority ordering between norms "may be determined in part by considerations arising from the text of the regulations themselves, such as the existence of cross-references from one to another; and it may also be determined in part by factors of a more extrinsic kind, such as the powers and competence of the issuing bodies, dates of promulgation and amendment, and the degree of specificity or generality of the regulations made" [Alchourron and Makinson, 1981, p.125].

In this work we do not consider the issue of how such priority relation between laws is generated. We assume it as given and embedded in the preferences of the normative agent.

Finally, our definition of permission does not imply an existing prohibition to which it is an exception. Permissions are independent from obligations: the fact that it is meaningless to enact a permission which does not play (and it neither can play) the role of an exception is a property of the normative system it belongs to, and not a feature of the single permission. Since [Boella and van der Torre, 2003b] focus on this problem, we do not address this issue here further.

# 4 Formalization

In this section we formalize the obligations and permissions in a qualitative game theory.

The presentation of the decision and game theory in this section consists of two parts. We first define all elements of the agent and its profile of the normative agent, and only thereafter consider the decision problem of the agent.

The basic picture is visualized in Figure 1 and reflects the deliberation of agent A in various stages.

Agent A has to make certain decisions, and it is deliberating about the effects of the fulfilment or the violation of norms. Agent N is the normative system, which may recognize and sanction violations. Agent A recursively models agent N's decision (taken from its point of view) and bases its choice on the effects of agent N's predicted actions. When agent A makes its decision $d_A$, it believes that it is in state $s_A^0$. The expected consequences of this decision (due to belief rules $B_A^1$) are called state $s_A^1$. Then agent N makes a decision $d_N$, typically whether it counts this decision as a violation and whether it sanctions agent A or not. Now, to find out which decision agent N will make, agent A has a *profile* of agent N: it has a representation of the initial state agent N believes to be in and of the following stages. When agent A makes its decision, it believes that agent N believes that it is in state $s_N^0$. This may be the same situation as state $s_A^0$, but it may also be different. Then, agent A believes that its own decision $d_A$ will have the consequence that agent N believes that it is in state $s_N^1$, due to its observations and the expected consequences of these observations. Agent A expects that agent N believes that the expected result of decision $d_N$ is state $s_N^2$. Finally, agent A's expected consequences of $d_N$ from A's point of view are called state $s_A^2$.

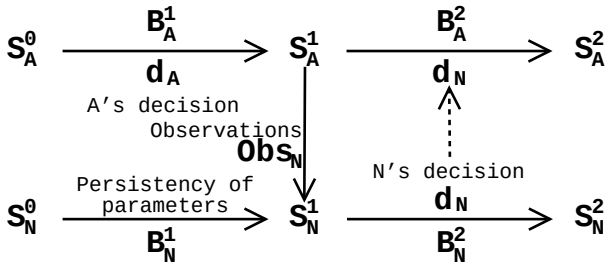


Figure 1: The agent A and the normative system N

In Section 4.1 we introduce our model of the agent's decisions and mental states, and in Section 4.2 the agent characteristics; in Section 4.3 we define obligations and permissions, and in Section 4.4 we define how the agent makes a decision by reasoning about these norms.

## 4.1 Decisions, epistemic states and mental states

In this section we formalize the various elements of Figure 1, and we define when states respect belief rules and violate or fulfill desire and goal rules. However, we do not define the agent's decision problem or the way it deliberates about these elements.

The decisions and mental states of the agents are based on the qualitative decision theory proposed in the BOID architecture by [Broersen *et al.*, 2002a].

We start with the *decisions*. We assume that the base language contains boolean variables and logical connectives. The variables are either *decision variables* of an agent, whose truth value is directly determined by it, or *parameters*, whose truth value can only be determined indirectly. The distinction between decision variables and parameters is a fundamental principle in all decision theories or decision logics. Our terminology is borrowed from Lang et al. [Lang *et al.*, 2002], they are called respectively controllable and uncontrollable propositions by Boutilier [Boutilier, 1994].

**Definition 1 (Decisions)** *Let $A = \{a_1, a_2, \ldots\}$, $N = \{m_1, m_2, \ldots\}$ and $P = \{p_1, p_2, \ldots\}$ be three disjoint sets of propositional variables, i.e. $A \cap N = \emptyset$, $A \cap P = \emptyset$, and $N \cap P = \emptyset$. A literal is a variable or its negation. For a propositional variable $p$ we write $\overline{p} = \neg p$ and $\overline{\neg p} = p$.*

*A decision set is a tuple $\langle d_A, d_N \rangle$ where $d_A$ is a set of literals of $A$ (the decision of agent A) and $d_N$ is a set of literals of $N$ (the decision of agent N). We assume that decisions are complete, in the sense that for example for each decision variable $a$ in $A$, we have either $a \in d_A$ or $\neg a \in d_A$. Moreover, $a \notin d_A$ means that $\neg a$ is in $d_A$. Analogously for $d_N$.*

We distinguish between what we call the agent's *epistemic states* and its *mental states*. By epistemic states we mean beliefs about the world and about the beliefs of the normative agent's beliefs about the world, i.e. about the states $s_{A/N}^{0/1/2}$ in Figure 1. By mental states we mean the sets of belief, desire and goal *rules* of the agents.

We first formalize the epistemic states. Since the value of the propositional variables must be distinguished in the three

stages, we label the parameters according to the stage number they describe.

**Definition 2 (Epistemic states)** *Let $P^0$, $P^1$ and $P^2$ be the sets of propositional variables defined by $P^i = \{p^i \mid p \in P\}$. We write $L_A$, $L_{AP^1}$, ... for the propositional languages built up from $A$, $A \cup P^1$, ... with the usual truth-functional connectives. We assume that the propositional language contains a symbol $\top$ for a tautology.*

*The epistemic state is a tuple $\langle s^0_A, s^1_A, s^2_A, s^0_N, s^1_N, s^2_N \rangle$ where $s^0_A$ and $s^0_N$ are sets of literals of $L_{P^0}$ (the initial state), $s^1_A$ and $s^1_N$ are sets of literals of $L_{AP^1}$ (the states after the decision $d_A$ of agent A), and $s^2_A$ and $s^2_N$ are sets of literals of $L_{NP^2}$ (the states after the decision $d_N$ of agent N). Moreover, let $s_A = s^0_A \cup s^1_A \cup s^2_A$, $s_N = s^0_N \cup s^1_N \cup s^2_N$. The states are assumed to be complete.*

We now formalize the agent's mental state. The mental state contains four sets of rules for each agent. Two sets of *belief rules* are used to calculate the expected consequences of decisions and two sets of *desire* and *goal rules* express the attitudes of the agents towards a given state of affairs, depending on the context. In particular, $B^1_A$ and $B^1_N$ represent the belief rules from which the second stage is produced (according to agent A and agent N respectively) on the basis of the initial one and of the decision of agent A. $B^2_A$ and $B^2_N$ are the belief rules which produce the last stage. $D_A$, $G_A$, $D_N$ and $G_N$ represent the desires and goals of the two agents. Note that they cover all the aspects of the three stages.

**Definition 3 (Mental states)** *Let a rule of one of the propositional languages $L_A$, $L_{AP^1}$, ... be an ordered sequence of literals $l_1, \ldots, l_n, l$ of this language written as $l_1 \wedge \ldots \wedge l_n \rightarrow l$.*

*The mental state $M^N_A$ is a tuple $\langle B^1_A, B^2_A, B^1_N, B^2_N, D_A, G_A, D_N, G_N \rangle$ where $B^1_A$ and $B^1_N$ are sets of rules of $L_{AP^0P^1}$, $B^2_A$ and $B^2_N$ are sets of rules of $L_{ANP^0P^1P^2}$, $D_A$, $G_A$, $D_N$ and $G_N$ are sets of rules of $L_{ANP^0P^1P^2}$. We write $B_A = B^1_A \cup B^2_A$ and $B_N = B^1_N \cup B^2_N$. We assume that $G_N \subseteq D_N$, such that the goals of the normative agent correspond to its desires only.*

The normative agent's beliefs depend on what it can observe. Again we accept a simple formalization of this complex phenomena, based on an explicit enumeration of all propositions which can be observed.

**Definition 4 (Observations)** *The set of observable propositions $OP_N$ is a subset of the description of the second stage (according to agent A's point of view) and of agent A's decision: $P^1 \cup A$. The expected observation of agent N in state $s^1_A$ is $Obs_N = \{p \mid p \in OP_N \text{ and } p \in s^1_A\} \cup \{\neg p \mid p \in OP_N \text{ and } \neg p \in s^1_A\}$: if a proposition describing state $s^1_N$ is observable, then agent N knows its value in $s^1_A$.*

## 4.2 Agent characteristics

How the agents reason about norms, and in particular how they deliberate whether they fulfill or violate obligations, depends not only on their interpretation of the obligations in terms of their beliefs, desires and goals, as defined above, but also on their *agent characteristics*. Given the same set of rules, distinct agents reason and act differently. For what concerns reasoning, different agents can deal with conflicts among belief rules in different ways. For what concerns acting, a respectful agent always tries to fulfill the goals of the normative system, whereas a selfish agent first tries to achieve its own goals. We express these agent characteristics by a priority relation on the rules. As detailed in [Broersen *et al.*, 2002a], this encodes the way in which the agent resolves its conflicts.

**Definition 5 (Agent characteristics)** *The characteristics of the agent A are a tuple $\langle \geq^B_A, \geq_A, \geq^B_N, \geq_N \rangle$ where $\geq^B_A$ is a transitive and reflexive relation on the powerset of $B_A$, $\geq_A$ is a transitive and reflexive relation on the powerset of $D_A \cup G_A \cup D_N \cup G_N$, which contains at least the subset relation, $\geq^B_N$ is a transitive and reflexive relation on the powerset of $B_N$, containing at least the subset relation, and $\geq_N$ is a transitive and reflexive relation on the powerset of $D_N \cup G_N$, which containing at least the subset relation.*

The decision sets and epistemic states are related to each other by the agent's mental state. There are two different kinds of relations. First, the belief rules express whether the states are the expected consequences of the decisions. Second, the desire and goal rules are used to evaluate the consequences of decisions. In other words, the relations are whether the decisions and epistemic states respect the belief rules of the mental state, and whether the decisions and epistemic state violate or fulfill the desires and goals.

We first formalize respecting mental states. For rational agents, the epistemic state is a consequence of applying belief rules to the previous state, together with persistence of the previous state. In such a case, we say that the epistemic states respect the mental state. In this paper we use the following definition to express how belief rules are used to compute the subsequent state starting from a state and a decision or a set of observations. This is clearly a very simple definition, because for example it does not formalize the iterative application of belief rules. However, it suffices for our purposes, and extensions are straightforward.

The definition of respecting a mental state accounts also for the priority relation on beliefs defined in the agent characteristics: the $\max$ function does not only consider only maximal sets of rules but also only the ones which are maximal in the given ordering. If the order is total, then the $\max$ function generates a single state.

**Definition 6 (Respecting mental states and beliefs)** *For $s$ a state, $f$ a set of literals in $L_{ANP^1}$, $R$ a set of rules, and $\geq$ a transitive and reflexive relation on the powerset of $R$ containing at least the superset relation, let $\max(s, f, R, \geq)$ be defined as the set of states obtained by:*

1. *$S$ is the set of states after applying a consistent subset of $R$ to the union of the state $s$ with $f$:*

$$S = \{\{l_1, \ldots, l_n\} \cup f \mid l_{i,1} \wedge \ldots \wedge l_{i,m_i} \rightarrow l_i \in R \text{ and } l_{i,j} \in s \cup f \text{ for } j = 1 \ldots m_i \text{ for } i = 1 \ldots n \text{ and } \{l_1, \ldots, l_n\} \cup f \text{ consistent } \}$$

2. *$S'$ is the set of maximal elements of $S$, i.e.*

$$S' = \{s \in S \mid \nexists s' \in S \text{ such that } s \subset s'\}$$

3. *$S''$ is the set of maximal (with respect to the $\geq$ ordering) elements of $S'$, i.e.*

$$\{s \in S' \mid \nexists s' \in S' s' \geq s \text{ and } s \not\geq s'\}$$

4. *$\max(s, f, R, \geq)$ is the set of states that contain an element of $S''$ together with some elements from $s$, then* $\max(s, f, R, \geq) = \{s' \cup s'' \mid s' \in S''$ *and* $s'' = \{l^i \in s \mid l^i \in P^i$ *and* $\overline{l^{i+1}} \notin s'\}\}$

*A state description $\langle s^0_A, s^1_A, s^2_A, s^0_N, s^1_N, s^2_N\rangle$ respects the decision set $\langle d_A, d_N\rangle$, the expected observation of agent N $Obs_N$ together with the mental state description $\langle B^1_A, B^2_A, B^1_N, B^2_N, D_A, G_A, D_N, G_N\rangle$ and the relations $\geq^B_A$ and $\geq^B_N$ if $s^1_A \in \max(s^0_A, d_A, B^1_A, \geq^B_A)$, $s^2_A \in \max(s^0_A \cup s^1_A, d_N, B^2_A, \geq^B_A)$, $s^1_N \in \max(s^0_N, Obs_N, B^1_N, \geq^B_N)$ and $s^2_N \in \max(s^0_N \cup s^1_N, d_N, B^2_N, \geq^B_N)$.*

### 4.3 Obligations and permissions

To define obligations and permissions in the qualitative game theory, we use the mental states and the distinction between decision variables and parameters, the latter to distinguish various notions of norm. Moreover, to define norms we introduce in this section a formalization of the notion "counts as a violation". Our approach is inspired by the so-called Anderson's reduction of deontic logic [Anderson, 1958] to alethic modal logic, which may be written as $O(p) = \Box(\neg p \rightarrow V)$. This modal formula says that if $p$ is obliged, then it is necessarily the case that the negation of $p$ implies the violation constant $V$. There has been much discussion in the philosophical logic literature on the interpretation of this constant $V$. In Anderson's original proposal, he interpreted it as a sanction. However, it was soon observed that many violations are not sanctioned. He therefore later in [Anderson, 1967] interpreted it as 'something bad has happened'.

For each literal built from a propositional variable $a$, we introduce a new decision variable $V(a) \in N$. The variable $V(a)$ is intended to mean that agent N determines that $a$ counts as a violation. All obligations which have the same content, when they are violated, count as the same violation.

The following definition presents the general notion of obligation.

**Definition 7 (Obligations)** *Let $NS$ be a normative system or set of norms $\{n_1, \ldots, n_m\}$ and let the decision variables of agent N contain a set of so-called violation variables $V = \{V(a) | a \in P^1 \cup P^2 \cup A\}$.*

*Agent A believes that it is obliged to decide to do $a$, a literal built from a parameter in $P^1 \cup P^2$ or a decision variable in $A$, $O_{AN}(a)$, iff agent A believes that there is a norm $n \in NS$ such that:*

1. *$\top \rightarrow a \in D_N \cap G_N$: agent A believes that agent N desires and has as a goal that $a$ and wants agent A to adopt $a$ as a goal.*
2. *$\neg a \rightarrow V(\neg a) \in D_N \cap G_N$: agent A believes that if agent N believes $\neg a$ then it has the goal and the desire to recognize it as a violation $V(\neg a)$.*
3. *$\top \rightarrow \neg V(\neg a) \in D_N$: agent A believes that agent N desires that there are no violations.*

*When the literal $a$ is built from a decision variable, then we call the obligation an ought-to-do obligation, and when it is built from a parameter then we call it an ought-to-be obligation.*

We now consider conditional obligations with sanctions.

**Definition 8 (Conditional obligations with sanction (parameter))** *Agent A believes that under condition $q$, $q \in L_{AP^0P^1}$, it is obliged to decide to do $a$ (a literal built from a propositional variable in $P^1 \cup P^2 \cup A$) with sanction $s$ (a parameter in $P^2$ to be achieved by agent N) $O_{AN}(a, s|q)$ iff for some $n \in NS$*

1. *$q \rightarrow a \in D_N \cap G_N$: agent A believes that in context $q$ agent N desires and has as a goal that $a$ and wants agent A to adopt $a$ as a goal.*
2. *$q \wedge \neg a \rightarrow V(\neg a) \in D_N \cap G_N$: agent A believes that if agent N believes that $q$ and $\neg a$ then it has the goal and the desire to $V(\neg a)$: it recognizes it as a violation.*
3. *$\top \rightarrow \neg V(\neg a) \in D_N$: agent A believes that agent N desires that there are no violations.*
4. *$V(\neg a) \rightarrow s \in D_N \cap G_N$: agent A believes that if agent N decides $V(\neg a)$ then it desires and has as a goal that it sanctions agent A.*
5. *$\top \rightarrow \neg s \in D_N$: agent A believes that agent N desires not to sanction $\neg s$. This desire of the normative system expresses that it only sanctions in case of violation.*
6. *$\top \rightarrow \neg s \in D_A$: agent A has the desire for $\neg s$, which expresses that it does not like to be sanctioned.*
7. *for some $m \in N$ we have $m \rightarrow s \in B_N$: agent A believes that agent N has a way to apply the sanction.*

Analogously, if $V(\neg a)$ is a parameter, then also the prosecution process can be subject to failures. We assume that similar extensions are provided for $V(\neg a)$ as a parameter in $P^2$, though we do not give the details here.

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. Again we can consider variations with decision variables and parameters for the obligation, the violation constant, and the sanction. We only give the details of one of the options.

**Definition 9 (Obligations with reward (parameter))** *Agent A believes that it is obliged to decide to do $a$ (a literal built from a propositional variable in $P^1 \cup P^2 \cup A$) with reward $r$ (a parameter in $P^2$ to be achieved by agent N) $O_{AN}(a, r)$ iff for some $n \in NS$:*

1. *$\top \rightarrow a \in D_N \cap G_N$: agent A believes that agent N desires and has as a goal that $a$ and wants agent A to adopt $a$ as a goal.*
2. *$\neg a \rightarrow V(\neg a) \in D_N \cap G_N$: agent A believes that if agent N believes $\neg a$ then it has the goal and the desire to recognize it as a violation $v$.*

3. $\top \rightarrow \neg V(\neg a) \in D_N$: *agent A believes that agent N desires that there are no violations.*
4. $\neg V(\neg a) \rightarrow r \in D_N \cap G_N$: *agent A believes that if agent N decides $\neg V(\neg a)$ then it desires and has as a goal that it rewards agent A.*
5. $\top \rightarrow \neg r \in D_N$: *agent A believes that agent N desires not to reward $\neg r$. This desire of the normative system expresses that it only rewards in case of absence of violation.*
6. $\top \rightarrow r \in D_A$: *agent A has the desire for $r$, which expresses that it likes to be rewarded.*
7. *for some $m \in N$ such that $m \rightarrow r \in B_N$: agent A believes that agent N knows a way to apply the reward.*

We give the following definition of permission:

**Definition 10 (Conditional permissions (parameter))** *Agent A believes that it is permitted by the normative system N to decide to do $a$ (a literal built from a propositional variable in $P^1 \cup P^2 \cup A$) in context $q, q \in L_{AP^0P^1}$, $P_{AN}(a|q)$ iff for some $n \in NS$*

1. $q \wedge a \rightarrow \neg V(a) \in D_N \cap G_N$: *agent A believes that if agent N believes that $q$ and $a$ hold then it has the goal and the desire that $a$ does not count as a violation $V(a)$.*

In our definition, a given permission is not directly related to any obligation it can be the exception of. Thus, how can a permission be an exception to an obligation? The relation between permissions and obligations is given by the fact that the content of the permission is not considered as a violation: $\neg V(a)$. This propositional variable is exactly the negation of the violation of an obligation which prescribes that agent A should not do $a$: $O_{AN}(\neg a)$. Hence, a permission in our definition can be an exception to all those obligations which consider a given behavior as a violation and not just to a particular one. This explains how our definition of permission does not imply the existence of a corresponding obligation it is an exception to. We require that the relation between permissions and obligations is only indirect because, as discussed in [Bulygin, 1986], in case of hierarchies of authorities it is possible that a permission is issued by an higher level authority to prevent lower level ones to eventually forbid such behavior; in [Boella and van der Torre, 2003b] we provide a model of such subtle situation.

Finally, the conflict between the rule defining a permission and the corresponding rule in an obligation is resolved according to the normative agent characteristics, as any other conflict among rules. Since we do not consider here the problem of the legal sources of norms, we just consider how a permission and an obligation are related in a given normative agent characteristics, without addressing the problem of how it is generated according the different legal principles.

### 4.4 Agent types

The agents value, and thus induce an ordering on, the epistemic states by considering which desires and goals have been fulfilled and which have not. In the general case, agent A may consider its own desires and goals as well as the desires and goals of agent N, whereas agent 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 of the normative agent. Note that $U_A^{D_N}$ and $U_A^{G_N}$ refer to the state from agent A's point of view: agent A cares that it believes that the desires and goals of agent N are satisfied, not that agent N believes that they are satisfied while this is not the case.

**Definition 11 (Unfulfilled mental states)** *Let $U(R,s)$ be the unfulfilled rules of state $s$, $U(R,s) = \{l_1 \wedge \ldots \wedge l_n \rightarrow l \in R \mid \{l_1, \ldots, l_n\} \subseteq s \text{ and } l \notin s\}$. The* unfulfilled mental state description *of agent A is the tuple $U_A^N = \langle U_A^{D_A}, U_A^{G_A}, U_A^{D_N}, U_A^{G_N} \rangle$ where $U_A^{D_A} = U(D_A, s_A)$, $U_A^{G_A} = U(G_A, s_A)$, $U_A^{D_N} = U(D_N, s_A)$, $U_A^{G_N} = U(G_N, s_A)$. The* unfulfilled mental state description *of agent N is $U_N = \langle U_N^{D_N}, U_N^{G_N} \rangle$, where $U_N^{D_N} = U(D_N, s_N)$ and $U_N^{G_N} = U(G_N, s_N)$.*

For what concerns the priorities on desire and goal rules, our model of agent characteristics is very general, and it is possible to define some categories; we classify agents according to the way they solve the conflicts among the rules belonging to different components: desires, goals and desires and goals of the normative system that can be adopted. Agent types are defined in a similar way as they have been introduced in the BOID architecture [Broersen *et al.*, 2002a].

Here, for space reasons, we introduce only a selfish agent type, which bases its decisions only on its unsatisfied goals and desires.

**Definition 12 (Agent types)** *if $s_A \leq s'_A$ iff $U'^{G_A}_A \geq_A U_A^{G_A}$ and if $U'^{G_A}_A \geq_A U_A^{G_A}$ and $U_A^{G_A} \geq_A U'^{G_A}_A$ then $U'^{D_A}_A \geq_A U_A^{D_A}$*

We now consider the agent's *decision problem*. In the decision problems we consider, the agent only knows the initial state, considers the consequences of its possible decisions, and chooses the decision which results in the best states.

**Definition 13 (Optimal decisions)** *Given initial states $s_A^0$ and $s_N^0$, a mental state, observations $OP_N$ and agent characteristics, the decision set $\Delta = \langle d_A, d_N \rangle$ is N optimal if for all state descriptions $S$ which respect $\Delta$, for all other decision sets $\Delta = \langle d_A, d'_N \rangle'$, for all state descriptions $S'$ respecting $\Delta'$, $s_N \leq s'_N$. $\Delta$ is A optimal if it is N optimal and for all state descriptions $S$ which respect $\Delta$, for all other N optimal decision sets $\Delta'$, for all state descriptions $S'$ respecting $\Delta'$, $s_A \leq s'_A$.*

## 5 Examples

In this section we present some examples which show which roles permissions play in agent interaction, which is the mental state of the agent who authorizes another one and of the agent who is authorized. Moreover, we show some situations where an obligation is not respected but the normative agent is not aware of that or it cannot do anything for sanctioning such violations; also permissions are exploited in such situations.[3]

[3] Since the states and decision sets are complete, when parameters and decision variables are negated, we do not include them in the examples below if there is no risk of ambiguity.

The first example represents an agent which sticks to the obligation to do $a$ even if it has $\neg a$ as a goal or if there were no sanction $s$ since it desires not to violate the norm:

**Example 1** $O_{AN}(a, s)$
$s^0_A = \emptyset, B_A = \emptyset, \geq^B_A = \emptyset, a \in A$,
$G_A = \emptyset, D_A = \{\top \rightarrow \neg V(\neg a), \top \rightarrow \neg a\}, T_A = stable$,
$\geq_A = \{\top \rightarrow \neg V(\neg a)\} > \{\top \rightarrow \neg a\}$,
$s^0_N = \emptyset, OP_N = A \cup P^1, B_N = \emptyset, \geq^B_N = \emptyset$,
$G_N = \{\top \rightarrow a, \neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s\}, V(\neg a) \in V, s \in N, n \in NS$,
$D_N = \{\top \rightarrow a, \neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s, \top \rightarrow \neg V(\neg a), \top \rightarrow \neg s\}, T_N = stable$,
$\geq_N \supseteq \{\top \rightarrow a, \neg a \rightarrow V(\neg a)\} > \{\top \rightarrow a, \top \rightarrow \neg V(\neg a), \top \rightarrow \neg s\}$
*Optimal decision set:* $\langle d_A = \{a\}, d_N = \emptyset\rangle$
*Expected state description:*
$s^1_A = \{a\}, s^1_N = \{a\}, s^2_N = \emptyset, s^2_A = \emptyset$
*Unfulfilled mental states:*
$U^{D_A}_A = \{\top \rightarrow \neg a\}, U^{G_A}_A = U^{D_N}_N = U^{G_N}_N = \emptyset$

Agent A's desire not to be a violator is fulfilled: the antecedent $\top$ of the unconditional rule $\top \rightarrow \neg V(\neg a)$ is true, but the consequent is consistent with state $s^2_A = \emptyset$ since agent N decides that the situation does not count as a violation and thus it does not sanction ($\neg s$); recall that $V(\neg a), s \in N$, so they are implicitly variables of the last stage. In contrast, the unconditional (and hence applicable) desire $\top \rightarrow \neg a$ is in conflict with state $s^1_A = \{a\}$ and it remains unsatisfied: $a \in A$, so it is a decision variable describing second stage.

For what concerns agent N's attitudes, its unconditional desire and goal that agent A adopts the content of the obligation $\top \rightarrow a$ is satisfied in $s^1_N$. Analogously are the desires not to prosecute and sanction indiscriminately: $\top \rightarrow \neg V(\neg a)$ and $\top \rightarrow \neg s$ (recall that states are complete so $\neg V(\neg a)$ and $\neg s$ are true in $s^2_N = \emptyset$). The remaining conditional attitudes $\neg a \rightarrow V(\neg a)$, etc. are not applicable and hence they are not unsatisfied.

Whatever other decision agent N would have taken, it could not satisfy more goals or desires, so $d_N = \emptyset$ is a minimal and optimal decision. E.g. $d_N = \{s\}$ would leave $\top \rightarrow \neg s$ unsatisfied: $\{\top \rightarrow \neg s\} \geq \emptyset$ (in fact, $\geq_N$ contains at least the subset relation) and then $U^{D_N}_N = \emptyset \geq U^{D_N}_N{}'' = \{\top \rightarrow \neg s\}$ for a stable agent.

Had agent A's decision been $d'_A = \emptyset$, agent N would have chosen $d'_N = \{V(\neg a), s\}$. The unfulfilled desires and goals in state $s'_A = s'_N = \{V(\neg a), s\}$ would have been: $U^{D_A}_A{}' = \{\top \rightarrow \neg V(\neg a)\}, U^{G_A}_A{}' = \emptyset, U^{D_N}_N{}' = \{\top \rightarrow a, \top \rightarrow \neg V(\neg a), \top \rightarrow \neg s\}, U^{G_N}_N{}' = \{\top \rightarrow a\}$

How does agent A takes a decision between $d_A$ and $d'_A$? Since it is a stable agent, it gives its preference to the first decision, since its choice is based on the unsatisfied goals.

We modify the previous example to show that given a conditional permission not to do $a$ which overrides the obligation to $a$ in a context $q$, agent A needs not do $a$ anymore.

**Example 2** $O_{AN}(a, s)$ *and* $P_{AN}(\neg a \mid q^1)$
$s^0_A = \{q^0\}, B_A = \emptyset, \geq^B_A = \emptyset, a \in A$,
$G_A = \emptyset, D_A = \{\top \rightarrow \neg V(\neg a), \top \rightarrow \neg a\}, T_A = stable$,
$\geq_A = \{\top \rightarrow \neg V(\neg a)\} > \{\top \rightarrow \neg a\}$,
$s^0_N = \{q^0\}, OP_N = A \cup P^1, B_N = \emptyset, \geq^B_N = \emptyset$,
$G_N = \{\top \rightarrow a, \neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s, q^1 \wedge \neg a \rightarrow \neg V(\neg a)\}, V(\neg a) \in V, s \in N, n \in NS$
$D_N = \{\top \rightarrow a, \neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s, \top \rightarrow \neg V(\neg a), \top \rightarrow \neg s, q^1 \wedge \neg a \rightarrow \neg V(\neg a)\}, T_N = stable$,
$\geq_N \supseteq \{\top \rightarrow a, q^1 \wedge \neg a \rightarrow V(\neg a)\} > \{\top \rightarrow a, \neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s\} > \{\top \rightarrow a, \neg a \rightarrow V(\neg a)\} > \{\top \rightarrow \neg V(\neg a), \top \rightarrow \neg s\}$
*Optimal decision set:* $\langle d_A = \emptyset, d_N = \emptyset\rangle$
*Expected state description:*
$s^1_A = \{q^1\}, s^1_N = \{q^1\}, s^2_N = \{q^2\}, s^2_A = \{q^2\}$
*Unfulfilled mental states:*
$U^{D_A}_A = \emptyset, U^{G_A}_A = \emptyset, U^{D_N}_N = U^{G_N}_N = \{\top \rightarrow a, \neg a \rightarrow V(\neg a)\}$

Note that with respect to the previous example, agent A does not leave any mental state unsatisfied; instead, agent N is left with its goal that $a$ unsatisfied. However, it does not decide that $\neg a$ is a violation, since it is more important to satisfy the goal $q^1 \wedge \neg a \rightarrow \neg V(\neg a)$.

In the next example, the desire of the *stable* agent A (that $p^1$ is true) conflicts in an indirect way with the new option $a$. The belief rules and the ordering $\geq^B_A$ on them represent the incompatibility of the effects of the two decision variables $a$ (with side effect $\neg p^1$) and $b$ (which achieves $p^1$). Agent A has no advantage in choosing decision $d_A = \{a, b\}$ since $\max(s^0_A, d_A, B^1_A) = \{\{p^1\}, \{\neg p^1\}\}$, but $\max(s^0_A, d_A, B^1_A, \geq^B_A) = \{\{\neg p^1\}\}$; in fact, the decision specification is coherent since the ordering $B^1_A$ is total: it expresses the priority of the rule $a \rightarrow \neg p^1$ over $b \rightarrow \neg p^1$, i.e. in the context $a$ the effects of the second rule is overwritten in a non-monotonic way by the first rule.

The decision between $a$ and $b$ is taken by comparing the results of recursive modelling: if $a$ then $\neg p^1$ and $\neg s$ and if $b$ then $p^1$ and $s$. Since agent A prefers not being sanctioned $s$ with respect to leaving $p^1$ unfulfilled, it chooses $a$.

**Example 3** $O_{AN}(a, s)$
$s^0_A = \{\neg p^0\} = \emptyset, B_A = \{a \rightarrow \neg p^1, b \rightarrow p^1\}, a \in A, p^1 \in P^1$,
$\geq^B_A \supseteq \{a \rightarrow \neg p^1\} > \{b \rightarrow p^1\}$,
$G_A = \emptyset, D_A = \{\top \rightarrow p^1, \top \rightarrow \neg s\}, T_A = stable$,
$\geq_A \supseteq \{\top \rightarrow \neg s\} > \{\top \rightarrow p^1\}$,
$s^0_N = \{\neg p^0\}, OP_N = A \cup P^1, B_N = \emptyset, \geq^B_N = \emptyset$,
$G_N = \{\top \rightarrow a, \neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s\}, V(\neg a) \in V, s \in N, n \in NS$
$D_N = \{\top \rightarrow a, \neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s, \top \rightarrow \neg V(\neg a), \top \rightarrow \neg s\}$,
$\geq_N \supseteq \{\neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s\} > \{\top \rightarrow \neg V(\neg a), \top \rightarrow \neg a\}, T_N = stable$
*Optimal decision set:* $\langle d_A = \{a\}, d_N = \emptyset\rangle$
*Expected state description:*
$s^1_A = \{a, \neg p^1\}, s^1_N = \{a, \neg p^1\}, s^2_N = \emptyset, s^2_A = \emptyset$
*Unfulfilled mental states:*
$U^{D_A}_A = \{\top \rightarrow p^1\}, U^{G_A}_A = U^{D_N}_N = U^{G_N}_N = \emptyset$

We return on the conflict among motivations: agent A has the goals $\neg a$ and $\neg q^1$. However it is obliged to do $a$ unless $q^1$ is true. It can achieve the goal $\neg q^1$ by means of action $b$, but

if it does, it cancels the condition of the permission not to do $a$.

**Example 4** *$O_{AN}(a,s)$ and $P_{AN}(\neg a \mid q^1)$*
$s^0_A = \{q^0\}, B_A = \{b \rightarrow \neg q^1\}, \geq^B_A = \emptyset, a, b \in A, q^1 \in P^1$,
$G_A = \{\top \rightarrow \neg a, \top \rightarrow \neg q^1\}, D_A = \{\top \rightarrow \neg V(\neg a), \top \rightarrow \neg s\}, T_A = stable$,
$s^0_N = \{q^0\}, OP_N = A \cup P^1, B_N = \emptyset, \geq^B_N = \emptyset$,
$G_N = \{\top \rightarrow a, \neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s, q^1 \wedge \neg a \rightarrow \neg V(\neg a)\}, V(\neg a) \in V, s \in N, n \in NS$,
$D_N = \{\top \rightarrow a, \neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s, \top \rightarrow \neg V(\neg a), \top \rightarrow \neg s, q^1 \wedge \neg a \rightarrow \neg V(\neg a)\}, T_N = stable$,
$\geq_N \supseteq \{\top \rightarrow a, q^1 \wedge \neg a \rightarrow \neg V(\neg a)\} > \{\top \rightarrow a, \neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s\} > \{\top \rightarrow a, \neg a \rightarrow V(\neg a)\} > \{\top \rightarrow a, \top \rightarrow \neg V(\neg a), \top \rightarrow \neg s\}$
*Optimal decision set:* $\langle d_A = \{b\}, d_N = \{V(\neg a), s\}\rangle$
*Expected state description:*
$s^1_A = s^1_N = \{b, \neg q^1\}, s^2_N = s^2_A = \{V(\neg a), s, \neg q^2\}$,
*Unfulfilled mental states:*
$U^{D_A}_A = \{\top \rightarrow V(\neg a), \top \rightarrow \neg s\}, U^{G_A}_A = \emptyset, U^{D_N}_N = \{\top \rightarrow a, \top \rightarrow \neg V(\neg a), \top \rightarrow \neg s\}, U^{G_N}_N = \{\top \rightarrow a\}$

Agent A decides to violate the obligation since it is a stable agent: it considers its goals more important that its desires. Agent N decides to sanction agent A since in a state $s^1_N = \{b, \neg q^1\}$ the condition $q^1$ of the permission is not satisfied.

In the next example, we model the sanction $s$ as a parameter which is made true by a decision variable $m \in N$: in this way we account for the possibility of a failure in the execution of the punishment. We examine how agent A exploits the recursive modelling to influence the behavior of the normative agent. Besides the usual goals and desires described by the obligation to do $a$, we assume that, in a situation where $p^1$ is true, agent N has the goal to make the decision variable $r$ true. So, in state $s^1_N = \{\neg a, p^1\}$, it would like to choose decision $\{V(\neg a), m, b\}$: but the two decision variables are incompatible. Since the conditional desire $p^1 \rightarrow r$ is preferred to $V(\neg a) \rightarrow s$, agent N recognizes the violation ($V(\neg a)$) but it does not sanction agent A. This reasoning is reconstructed by agent A: it can skip the sanction by choosing the decision variable $b$ which has the effect $p^1$.

**Example 5** *$O_{AN}(a,s)$*
$s^0_A = \emptyset, B_A = \{b \rightarrow p^1\}, \geq^B_A = \emptyset, a \in A, b \in A, p^1 \in P^1$,
$G_A = \emptyset, D_A = \{\top \rightarrow \neg a, \top \rightarrow \neg s, \top \rightarrow \neg b\}$,
$\geq_A \supseteq \{\top \rightarrow \neg s\} > \{\top \rightarrow \neg a\} > \{\top \rightarrow \neg b\}, T_A = stable$,
$s^0_N = \emptyset, OP_N = A \cup P^1, B_N = \{m \rightarrow s, r \rightarrow \neg s\}$,
$\geq^B_N \supseteq \{r \rightarrow \neg s\} > \{m \rightarrow s\}, V(\neg a) \in V, m \in N, r \in N, s \in P^2, n \in NS$,
$G_N = \{\top \rightarrow a, \neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s, p^1 \rightarrow r\}$,
$D_N = \{\top \rightarrow a, \neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s, \top \rightarrow \neg V(\neg a), \top \rightarrow \neg s\}$,
$\geq_N \supseteq \{p^1 \rightarrow r\} > \{V(\neg a) \rightarrow s\} > \{\top \rightarrow \neg V(\neg a)\}, T_N = stable$
*Optimal decision set:* $\langle d_A = \{b\}, d_N = \{V(\neg a), r\}\rangle$
*Expected state description:*
$s^1_A = s^1_N = \{b, p^1\}, s^2_A = s^2_N = \{r, V(\neg a), p^2\}$
*Unfulfilled mental states:*
$U^{D_A}_A = \{\top \rightarrow \neg b\}, U^{G_A}_A = \emptyset$,
$U^{D_N}_N = \{\top \rightarrow a, \top \rightarrow \neg V(\neg a)\}, U^{G_N}_N = \{\top \rightarrow a, V(\neg a) \rightarrow s\}$

Up to this point we assumed that agent N can observe everything. But what if agent N is not immediately acquainted with the entire state? The limitations on agent N's observations can be exploited by a non-respectful agent A to make agent N believe that it did not commit a violation since it falsely believes that a permission overrides the obligation:

**Example 6** *$O_{AN}(a,s)$ and $P_{AN}(\neg a \mid q^1)$*
$s^0_A = \emptyset, B_A = \emptyset, \geq^B_A = \emptyset, a, b \in A, q^1 \in P^1, T_A = stable$,
$G_A = \emptyset, D_A = \{\top \rightarrow \neg s, \top \rightarrow \neg a\}, \geq_A = \{\top \rightarrow \neg s\} > \{\top \rightarrow \neg a\}$,
$s^0_N = \emptyset, OP_N = A \cup (P^1 - \{q^1\}), B_N = \{b \rightarrow q^1\}, \geq^B_N = \emptyset$,
$G_N = \{\top \rightarrow a, \neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s, b \wedge \neg a \rightarrow \neg V(\neg a)\}, V(\neg a) \in V, s \in N, n \in NS$
$D_N = \{\top \rightarrow a, \neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s, \top \rightarrow \neg V(\neg a), \top \rightarrow \neg s, b \wedge \neg a \rightarrow \neg V(\neg a)\}, T_N = stable$,
$\geq_N \supseteq \{\top \rightarrow a, b \wedge \neg a \rightarrow V(\neg a)\} > \{\top \rightarrow a, \neg a \rightarrow V(\neg a), V(\neg a) \rightarrow s\} > \{\top \rightarrow a, \neg a \rightarrow V(\neg a)\} > \{\top \rightarrow a, \top \rightarrow \neg V(\neg a), \top \rightarrow \neg s\}$
*Optimal decision set:* $\langle d_A = \{b\}, d_N = \emptyset\rangle$
*Expected state description:*
$s^1_A = \{b, \neg q^1\}, s^1_N = \{b, q^1\}, s^2_N = \{q^2\}, s^2_A = \emptyset$
*Unfulfilled mental states:*
$U^{D_A}_A = U^{G_A}_A = \emptyset, U^{D_N}_A = U^{G_N}_A = U^{D_N}_N = U^{G_N}_N = \{\top \rightarrow a, \neg a \rightarrow V(\neg a)\}$

In our model we are also able to model the famous Beccaria's argument against death penalty: if a crime committed by a criminal, say kidnapping ($a$), is punished with a sentence to death ($s$) then every other sanction, such as being jailed $r$, does not prevent the criminal to commit other misbehavior, say killing his hostage ($b$). This is modelled by means of conditional desires: agent A desires not to be sanctioned only if it is alive ($\neg s$).

**Example 7** *$O_{AN}(\neg a, s)$ and $O_{AN}(\neg b, r)$*
$s^0_A = \emptyset, B_A = \emptyset, \geq^B_A = \emptyset, a, b \in A$,
$G_A = \emptyset, D_A = \{\top \rightarrow a, \top \rightarrow \neg s, \top \rightarrow b, \neg s \rightarrow \neg r\}$,
$\geq_A \supseteq \{\top \rightarrow \neg s, \neg s \rightarrow r\} > \{\top \rightarrow a, \top \rightarrow b\} > \{\top \rightarrow a\} > \{\top \rightarrow b\} > \{\top \rightarrow \neg s\} > \{\neg s \rightarrow r\}, T_A = stable$,
$s^0_N = \emptyset, OP_N = A \cup P^1, B_N = \emptyset, \geq^B_N = \emptyset$,
$G_N = \{\top \rightarrow \neg a, a \rightarrow V(a), V(a) \rightarrow s, \top \rightarrow \neg b, b \rightarrow V(b), V(b) \rightarrow r\}, V(a), V(b) \in V, s, r \in N, n, n' \in NS$,
$D_N = \{\top \rightarrow \neg a, a \rightarrow V(a), V(a) \rightarrow s, \top \rightarrow \neg V(a), \top \rightarrow \neg b, b \rightarrow V(b), V(b) \rightarrow r, \top \rightarrow \neg V(b), \top \rightarrow \neg r\}, T_N = stable$,
$\geq_N \supseteq \{a \rightarrow V(a), V(a) \rightarrow s, b \rightarrow V(b), V(b) \rightarrow r\} > \{\top \rightarrow \neg V(a), \top \rightarrow \neg s, \top \rightarrow \neg V(b), \top \rightarrow \neg r\}$
*Optimal decision set:*
$\langle d_A = \{a, b\}, d_N = \{V(a), V(b), s, r\}\rangle$
*Expected state description:*
$s^1_A = \{a, b\}, s^1_N = \{a, b\}, s^2_N = s^2_A = \{V(a), V(b), s, r\}$
*Unfulfilled mental states:*
$U^{D_A}_A = \{\top \rightarrow \neg s\}, U^{G_A}_A = \emptyset, U^{D_N}_N = U^{G_N}_N = \{\top \rightarrow \neg a, \top \rightarrow \neg b, \top \rightarrow \neg V(a), \top \rightarrow \neg V(b), \top \rightarrow \neg s, \top \rightarrow \neg r\}$

In principle, agent A prefers not to commit its crimes to the situation where both its desire not to be killed and not to be jailed are unsatisfied. However, it will kidnap and kill its hostage, since it is not possible that these desires remain unsatisfied at the same time: being jailed ($r$) is relevant only in case it has not been killed ($s$). So the best decision is to kidnap and kill the hostage.

## 6 Conclusions

In this paper we propose a cognitive model for deliberate normative agents who reason about what is permitted or forbidden to them. The definition of obligation is based on three dimensions. The first dimension is the set of agents involved, where we distinguish the agent whose behavior is norm governed, and the normative agent who issues norms and monitors and sanctions violations. The second dimension is the mental attitudes attributed to each agent, where we distinguish beliefs, desires and goals each represented by conditional rules. The third dimension are the elements of the norms and obligations, where we distinguish between behavior that counts as a violation, and sanctions. Obligations and permissions are defined in terms of abilities, beliefs, desires and goals of the normative agent. Permissions are modelled as behavior which does not count as a violation, and they constitute exceptions to obligations which consider the same behavior as a violation which is sanctioned.

We use a qualitative decision theory based on conditional rules extended with recursive modelling to model the interaction between the agent and the normative agent. The qualitative decision theory is inspired by [Broersen *et al.*, 2002a].

In [Dastani and van der Torre, 2002a] a similar framework has been used to model different types of agents according to the way they resolve conflicts. In [Broersen *et al.*, 2002b], it is discussed the relation between beliefs and desires in rational agents who have only realistic motivations.

Issues for further research are the limitation of what can count as a violation for the normative system according to the goals it has been socially delegated to by the society, the creation of norms ([Boella and van der Torre, 2003c]), the hierarchical relations among norms ([Boella and van der Torre, 2003b]) and the separation of all three elements of Montesquieu's *trias politica*: the legislative power (here the normative agent) should be kept distinct from the judicial power who decides whether a certain behavior counts as a violation and the executive one which applies the sanctions to violators.

## References


[Alchourron and Makinson, 1981] C. E. Alchourron and D. Makinson. Hierarchies of regulations and their logic. In R. Hilpinen, editor, *New studies in deontic logic*, pages 125–148. D. Reidel, Dordrecht, 1981.

[Anderson, 1958] A. Anderson. A reduction of deontic logic to alethic modal logic. *Mind*, 67:100–103, 1958.

[Anderson, 1967] A. Anderson. Some nasty problems in the formalization of ethics. *Noûs*, 1:345–360, 1967.

[Bobbio, 1980] N. Bobbio. Norma. In *Enciclopedia Einaudi*, volume 9, pages 876–907, Torino, 1980. Einaudi.

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

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

[Boella and van der Torre, 2003b] G. Boella and L. van der Torre. Permissions and obligations in hierarchical normative systems. In *Procs. of ICAIL 03*, Edimburgh, 2003.

[Boella and van der Torre, 2003c] G. Boella and L. van der Torre. Rational norm creation: A Recursive Model of Normative Systems and Agents In *Procs. of ICAIL 03*, Edimburgh, 2003.

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

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

[Broersen *et al.*, 2002b] J. Broersen, M. Dastani, and L. van der Torre. Realistic desires. *Journal of applied non-classical logics*, 12(2):287–308, 2002.

[Bulygin, 1986] E. Bulygin. Permissive norms and normative systems. In A. Martino and F. Socci Natali, editors, *Automated Analysis of Legal Texts*, pages 211–218. Publishing Company, Amsterdam, 1986.

[Castelfranchi *et al.*, 1999] C. Castelfranchi, F. Dignum, C. Jonker, and J. Treur. Deliberate normative agents: Principles and architecture. In *Intelligent Agents VI*, Springer Verlag, Berlin, 1999.

[Castelfranchi, 1997] C. Castelfranchi. Practical permission: Dependence, power, and social commitment. In *Proceedings of 2nd workshop on Practical Reasoning and Rationality*, London, 1997.

[Castelfranchi, 1998] C. Castelfranchi. Modeling social action for AI agents. *Artificial Intelligence*, 103:157–182, 1998.

[Castelfranchi, 2000] C. Castelfranchi. Engineering social order. In *Proceedings of ESAW00*, Berlin, 2000.

[Conte *et al.*, 1998] R. Conte, C. Castelfranchi, and F. Dignum. Autonomous norm-acceptance. In *Intelligent Agents V* . Springer Verlag, Berlin, 1998.

[Dastani and van der Torre, 2002a] M. Dastani and L. van der Torre. A classification of cognitive agents. In *Procs. of Cogsci02*, Fairfax (VA), 2002.

[Dennett, 1987] D. Dennett. *The intentional stance*. Bradford Books/MIT Press, Cambridge (MA), 1987.

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

[Goffman, 1970] E. Goffman. *Strategic Interaction*. Basil Blackwell, Oxford, 1970.

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

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

[Verhagen, 1999] Harko Verhagen. On the learning of norms. In *Proceedings of MAAMAW99*, 1999.