Obligations as Social Constructs

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

[1] Dipartimento di Informatica - Università di Torino- Italy. E-mail: guido@di.unito.it
[2] SEN-3 - CWI Amsterdam - The Netherlands. E-mail: torre@cwi.nl

Abstract. In this paper we formalize sanction-based obligations in the context of Searle's notion of construction of social reality. In particular, we define obligations using a *counts as* conditional, Anderson's reduction to alethic modal logic and Boella and Lesmo's normative agent. Our analysis presents an alternative criticism to the weakening rule, which has already been criticized in the philosophical literature for its role in the Ross paradox and the Forrester paradox, and the analysis presents a criticism to the generally accepted conjunction rule. Moreover, we show a possible application of these results in a qualitative decision theory. Finally, our analysis also contributes to philosophical discussions such as the distinction between violations and sanctions in Anderson's reduction, and between implicit and explicit normative systems.

1 Introduction

In agent theory, mental and social attitudes used in folk psychology such as knowledge, beliefs, desires, goals, intentions, commitments, norms, obligations, permissions, *et cetera*, are attributed to artificial systems [15]. The conceptual and logical study of these attitudes changes with the change of emphasis from autonomous agent systems to multiagent systems. For example, new challenges have been posed by new forms of multiagent systems such as web based virtual communities realized by the grid and peer to peer paradigms. In these settings it is not possible to design a central control since they are made of heterogeneous agents which cannot be assumed always to stick to the system regulations. The main driving force of single agent systems was Newell and Simon's study of knowledge and goals as knowledge level concepts in bounded or limited reasoning in knowledge based systems [20], and more recently Bratman's study of intentions as, amongst others, stabilizers of behavior in the agent's deliberation and planning process [9]. Likewise, joint intentions, joint commitments, norms and obligations are studied as stabilizers of multiagent systems.

However, from philosophical and sociological studies it is well known that there is more to multiagent concepts than stabilizing behavior. For example, multiagent behavior may spontaneously emerge without being reducible to the behavior of individual agents (known as the micro-macro dichotomy). Moreover, in a society the emersion of normative concepts is possible since they are constructed due to social processes. Searle's notion of construction of social reality explains these processes, e.g., how due to social conventions banknotes may be more than just pieces of paper, and what it means to be married [22]. The core concept of this construction is that in a social reality, certain actions and facts may *count as* something else. Under certain conditions, a priest performing a ritual counts as marrying a couple. Considering normative conceptions inspired by how human societies work and are constructed may have a decisive role also in the coordination of multiagent systems such as virtual communities, especially when artificial agents have to interact with human ones, as they do on the web.

We are interested in formal accounts of obligations that build on Searle's notion of construction of social reality. The obvious candidate for a formalization of norms and obligations is deontic logic, the logic of obligations. In particular, we may use Anderson's reduction of $O(p)$, read as 'p is obligatory', to $\Box(\neg p \rightarrow V)$, read as either 'the absence of p leads to a sanction' or 'the absence of p leads to a bad state' [1]. Anderson's reduction has proven useful in agent theory as part of Meyer's reduction of deontic action logic to dynamic logic [19], in which $F(\alpha)$, to be read as 'action α is forbidden, is reduced to $[\alpha]V$, after the execution of α V holds. However, these studies do not distinguish violations from sanctions, and they do not show how Searle's notion of social construction may fit in.

In this paper we introduce and study a deontic logic, using ideas developed in agent theory to formalize the notion of social construction. We formalize and extend an idea recently proposed by Boella and Lesmo [2]. They attribute mental attitudes to the normative system - which they therefore call the normative *agent*. They relate the external motivation of the agent to the internal motivation of the normative agent. The wishes (or desires, or goals) of the normative agent are the commands (or obligations) of the agent.

The relevance of this paper for agent theory is that it can be applied to several norm-based agent architectures that have recently been developed [2, 10, 12, 16]. The formalization of sanction-based obligations shows which are the motivations for the agents to fulfil the obligations they are subject to. In this way it is not necessary to assume that agents are designed so to fulfil obligations from which they do not gain any advantage. In this paper we also consider a decision-theoretic account of norms and obligations as an application of our results. Moreover, the relevance of our study for deontic logic is an alternative criticism to the weakening rule, which has already been criticized in the philosophical literature for its role in the Ross and Forrester paradoxes, and a criticism, based on legal reasoning, to the generally accepted conjunction rule.

We are motivated in this study by our research on norms for multiagent systems. In other works we propose obligations defined in a qualitative decision theory inspired by the BOID architecture of Broersen *et al.* [10]. In this paper we study logical relations between such obligations. This paper is thus a kind of analysis of an element of the model we present in [3, 8], which extends Boella and Lesmo's definition of sanction-based obligations, and distinguish between what counts as a violation, and which sanctions are applied: the agents take a decision based on the recursive modelling of the behavior of the normative agent according to whether it sanctions them or not. In [5, 4, 6] the same model is used to formalize policies regulating virtual communities on the web.

2 Social constructions

We have to fix some terminology. First, we identify normative systems with normative agents and switch freely between them. This is based on the attribution of mental attitudes to the normative system, as discussed in the introduction [3, 8]. We specifically do not restrict the normative system or normative agent to human agents. Second, during the last two decades, knowledge and goals have been replaced by beliefs, desires and intentions. Since for the normative agent we have to choose between goals and desires we opt for the latter, though in the context of this paper desires and goals can be interchanged (see Section 2.3).

Our method to formalize obligations is modal logic [13]. Assume a modal language that contains at least the modalities $O_{AN}(p)$: in normative system N, agent A is obliged to see to it that p, $D_N(p)$: the normative agent desires p, $V_{NA}(p)$: according to N, p counts as a violation by A, and $S_{NA}(p)$: according to N, A is sanctioned for p. The following two choices determine our deontic logic. First, the definition of O_{AN} in terms of D_N, V_{NA} and S_{NA}. An agent is obliged to see to it that p iff the normative agent desires that p, the normative agent desires that $\neg p$ counts as a violation by A, and the normative agent desires that if $\neg p$ counts as a violation, then A is sanctioned for $\neg p$. Note that an obligation for p implies that the normative agent has a desire for p, but this does not imply that all agents have an obligation for p. For the other agents absence of p does not have to count as a violation. Moreover, the fact that $\neg p$ counts as a violation is not a fact independent from the normative agent's behavior: rather, it is a desire of N, so that it must decide to do something for making $\neg p$ count as violation.

Given this definition, in case of a violation, it is possible to predict N's behavior from his desires and goals: he will decide that $\neg p$ is a violation and he will sanction A. Second, the logical properties of D_N, V_{NA} and S_{NA}. Instead of choosing one particular logic for these three primitive concepts, which would lead to a unique deontic logic for a particular definition, in this paper we take the logical properties of D_N, V_{NA} and S_{NA} as a parameter. That is, we show that O_{AN} has a certain property if D_N, V_{NA} and S_{NA} have certain other properties. In this way our results can be applied to a wide variety of logical systems.

Boella and Lesmo's construction introduces a new problem, which may be called the obligation distribution problem. Given a set of goals or desires of the normative agent, how are they distributed as obligations over the agents? Typical subproblems which may be discussed here are whether a group of agents can be jointly obliged to see to something, without being individually obliged. Similar problems are studied, e.g., by [11]. Another subproblem is whether agents can transfer their obligations to other agents. In this paper we do not study these questions, and we simply define that a desire of the normative agent counts as an obligation of agent A, when the unfulfillment of this desire counts as a violation by agent A.

2.1 Counts as a violation

In Searle's theory, *counts as* is a conditional relativized to an institution or society. Thus, when p and q are descriptions of some state of affairs or action, and N is a description of an institution, then $p \Rightarrow_N q$ may be read as 'p counts as q according to institution N'. A conditional logic along this line has been developed by Jones and Sergot [17]. Jones and Sergot study the *counts as* conditional $p \Rightarrow_i q$ in the context of modal action logic $E_a(p)$ for agent a sees to it that p. The conditional $p \Rightarrow_i q$ is closed under left-OR, right-AND, and TRANS, but not under right-W nor left-S. The latter makes their conditional a defeasible one. Their motivation is that their action operator satisfies the success postulate $E_a(p) \rightarrow p$, and that they do not like to infer $E_y(E_x(A)) \Rightarrow_i B$ from $E_x(A) \Rightarrow_i B$.

In a normative system with norms $\{n_1, \ldots, n_k\}$, with p and N as before, n a norm and V as a violation operator, $p \Rightarrow_N V_A(n)$ may be read as 'p counts as a violation by agent A according to norm n of institution N'. However, in deontic logic the formal language usually abstracts away from agents, institutions and explicit norms, because either they are irrelevant for the logical relations between obligations, or they seem to block such an analysis. In Section 2 and 3 we also abstract away from the explicit norms, such that 'p counts as a violation' may be represented as $p \Rightarrow_N V_{NA}$, which we abbreviate by $V_{NA}(p)$. In Section 4 we discuss explicit normative systems. For an extensive discussion for and against explicit norms in deontic logic see the discussion on the so-called diagnostic framework for deontic reasoning (diOde) in [24].

There is no consensus on the logical properties of the *counts as* conditional, maybe because the conditional can be used in many different kinds of applications. We therefore do not build our analysis on the conditional. The approach we follow in this paper is to study a default interpretation of $V_{NA}(p)$, together with various other alternatives. That is, a particular interpretation of it will be used by default, in absence of information to the contrary.

For our default interpretation, we say that the following property called strengthening (S) holds, whereas the property called weakening (W) does not hold. For example, if speeding counts as a violation, then speeding in a red car counts as a violation too. However, if driving under 18 counts as a violation, then driving by itself does not count as a violation. Note that the property called conjunction (AND) follows from S. We write → for the material implication.

S $\vdash V_{NA}(p) \rightarrow V_{NA}(p \wedge q)$
not-W $\nvdash V_{NA}(p) \rightarrow V_{NA}(p \vee q)$
AND $\vdash V_{NA}(p) \wedge V_{NA}(q) \rightarrow V_{NA}(p \wedge q)$

If both S and W hold, then we have $V_{NA}(p) \rightarrow V_{NA}(q)$, i.e., when some formula counts as a violation, then all formulas count as a violation. In other words, in such a case the logic only distinguishes between no violation and violation. In such a case, we say that the 'counts as a violation' operator V_{NA} trivializes. This trivial operator V_{NA} corresponds to the notion of violability studied by Anderson [1], because it does not distinguish between distinct violations (see e.g., [24]). Note that this kind of trivialization should be distinguished from the trivialization represented by $p \leftrightarrow V_{NA}(p)$. In the latter kind of trivialization, the modal operator has become superfluous. In our kind of trivialization, we go from a fine-grained to binary distinction.

Moreover, by default we assume that the following property called disjunction (OR) holds. For example, assume that 'driving 120 km/hour' counts as a violation and that 'driving drunk' counts as a violation. By default we conclude that 'driving drunk or 120 km/hour' counts as a violation, because we know that some norm has been violated.

OR $\vdash V_{NA}(p) \wedge V_{NA}(q) \rightarrow V_{NA}(p \vee q)$

Clearly, for our default interpretation we cannot use standard normal modal operators, because they satisfy W. This suggests the use of a minimal modal logic, as used in several recent agent logics [18]. However, when $\Box$ is a normal modal operator, then $\Box'$ defined by $\Box'(p) =_{\text{def}} \Box(\neg p)$, satisfies S instead of W. This definition in terms of a normal modal operator is the default choice for V_{NA}. We say that $\Box'$ is the negation or negative of $\Box$. For example, prohibitions are the negative of obligations. Note that permission $P(p) =_{\text{def}} \neg O(\neg p)$ is also sometimes called the negation (or dual) of an obligation. Thus, what we call the negation should be distinguished from other uses in the literature.

2.2 Being sanctioned

Sanctioning is an action of the normative agent. The normative agent sanctioning A for $\neg p$ with s due to norm n may be represented by $S_{NA}(\neg p, s, n)$. A logical property we discuss later in this paper is that the normative agent can sanction only if the agent's behavior counts as a violation of this norm.

Whether an action of the normative agent is a sanction or just any other action, i.e., whether it counts as a sanction, is also a social construction. For example, whether giving a fine counts as a sanction for late delivery, may depend on a convention in the society. We may thus write $s \Rightarrow_N S_{NA}(\neg p, n)$: according to institution N, s counts as a sanction for $\neg p$, agent A and norm n. However, it is important to notice that $S_{NA}(\neg p)$ in the definition of obligation should not be read as '$\neg p$ counts as a sanction'. The normative agent does not desire that s counts as a sanction, but that $\neg p$ is sanctioned with s. This is subtly different.

If we abstract away from norm n and sanction s, then we write $S_{NA}(\neg p)$ for $\neg p$ is being sanctioned. As far as we know, this operator has not been discussed in the literature. Again, it seems reasonable to accept, as a starting point, S, AND, and OR, and reject W. This is therefore our default choice.

S $\vdash S_{NA}(p) \rightarrow S_{NA}(p \wedge q)$
not-W $\not\vdash S_{NA}(p) \rightarrow S_{NA}(p \vee q)$
AND $\vdash S_{NA}(p) \wedge S_{NA}(q) \rightarrow S_{NA}(p \wedge q)$
OR $\vdash S_{NA}(p) \wedge S_{NA}(q) \rightarrow S_{NA}(p \vee q)$

Alternatively, we may abstract away from the reason for the sanction, and write $S_{NA}(s)$ for A is sanctioned with s. The latter can also be simplified to a single proposition s.

2.3 Desires

There has been some discussion on the distinction between desires and goals. If we consider a deliberation cycle, then desires are usually considered to be more primitive, because goals have to be *adopted* [14] or *generated* [10]. Goals can be based on desires, but also on other sources. For example, a social agent may adopt as a goal the desires of another agent, or an obligation. In knowledge based systems [20], goals are related to utility aspiration level and to limited (bounded) rationality. Moreover, here goals have desirability aspect as well as intentionality aspect, whereas in BDI circles it has been argued that this desirability aspect should be separated.

An important distinction for our present purposes is whether we may have $D_N(p)$ and $D_N(\neg p)$ at the same time. If such conflicts are considered to be inconsistent, then the desires can be formalized by a normal modal operator of type KD. System KD is the smallest set that contains the propositional formulas, the axioms $K : D_N(p \rightarrow q) \rightarrow (D_N(p) \rightarrow D_N(q))$ and $D : \neg(D_N(p) \wedge D_N(\neg p))$, and is closed under modus ponens and necessitation. This is the formalization used in e.g., [21], and our default choice.

If desires are allowed to conflict, and $D_N(p) \wedge D_N(\neg p)$ has to be represented in a consistent way, then desires may be represented by a so-called minimal modal operator [13, 18], in which the conjunction rule AND is not valid.

3 Obligations

3.1 Basic definition

We start with the definition of obligations in terms of desires, counts as a violation, and being sanctioned. The basic definition contains three clauses. (1) says that an obligation of A is a desire of N. (2) says that if $\neg p$ is the case, then N desires that it counts as a violation. (3) says that if $\neg p$ counts as a violation, then N desires that it is sanctioned. Permissions are defined as usual.

Definition 1 (Obligation). *Consider a modal logic with modal operators D_N (for desire or goal), V_{NA} (for counts as a violation) and S_{NA} (for being sanctioned). Obligation and permission are defined by:*

$$\begin{array}{lll} O_{AN}(p) =_{def} & D_N(p) \wedge & (1) \\ & \neg p \rightarrow D_N(V_{NA}(\neg p)) \wedge & (2) \\ & V_{NA}(\neg p) \rightarrow D_N(S_{NA}(\neg p)) & (3) \\ P_{AN}(p) =_{def} & \neg O_{AN}(\neg p) & \end{array}$$

We now consider various properties for the three modal operators of the normative agent. We first consider the case in which the three modal operators are defined as either modal operators of type KD or negatives of them.

Proposition 1. *Let the modal operator D_N be a normal modal operator of type KD, and let V_{NA} and S_{NA} be negated operators of type KD in the sense that*

$V_{NA}\neg$ and $S_{NA}\neg$ are normal modal operators of type KD. The logic does not satisfy weakening (W), strengthening (S), conjunction (AND), or disjunction (OR). It only satisfies the following formula called Deontic (D):

not-S	$\not\vdash O_{AN}(p) \rightarrow O_{AN}(p \wedge q)$
not-W	$\not\vdash O_{AN}(p) \rightarrow O_{AN}(p \vee q)$
not-AND	$\not\vdash O_{AN}(p) \wedge O_{AN}(q) \rightarrow O_{AN}(p \wedge q)$
not-OR	$\not\vdash O_{AN}(p) \wedge O_{AN}(q) \rightarrow O_{AN}(p \vee q)$
D	$\vdash O_{AN}(p) \rightarrow P_{AN}(p)$

Proof. *AND does not hold due to (2), and W and OR do not hold due to (3).*

The following proposition studies in more detail the conditions under which the properties are satisfied.

Proposition 2. *O_{AN} does not satisfy S.*

O_{AN} satisfies W if D_N satisfies W, V_{NA} trivializes in the sense that it satisfies W as well as S, and S_{NA} satisfies W.

O_{AN} satisfies AND if D_N satisfies AND and W, V_{NA} trivializes, and S_{NA} satisfies OR.

O_{AN} satisfies OR if D_N satisfies OR and AND, V_{NA} satisfies W and AND, and S_{NA} satisfies AND.

O_{AN} satisfies D if D_N satisfies D.

Corollary 1. *O_{AN} satisfies W, AND and OR if D_N is a normal modal operator, V_{NA} trivializes and S_{NA} is the negative of a normal modal operator.*

3.2 Interpretation of results

The corollary explains why Anderson's reduction, as well as most deontic logics developed along this line, only consider a single violation constant. Such a simple notion of violability leads to a logic with many desirable properties. Let us consider the results in more detail.

In so-called Standard Deontic Logic (SDL), a normal modal system of type KD, the obligations satisfy weakening and conjunction, but lack strengthening. The result that our O_{AN} lacks weakening is thus in conflict with this logic, but it is in line with a long standing tradition in deontic logic that rejects it, see [23] for a survey and discussion. The reason is that this proof rule leads to counterintuitive results in the so-called Ross paradox ('you ought to mail the letter' implies that 'you ought to mail the letter or burn it') and the Forrester paradox ('you should not kill', but 'if you do so then you should do it gently'). However, here the reason is completely different. W does not hold due to (3), which means that the reason is not the violability but the association of sanctions with violations.

The result that O_{AN} lacks conjunction is surprising, because most deontic logics satisfy this rule. The motivation of deontic logics not satisfying this rule is that they want to represent conflicts in a consistent way. Moreover, our result is in particular surprising since the rule is already blocked by clause (2), i.e., it is blocked due to the violability clause. The reason is the condition of (2). For an example, consider the two obligations 'driving 120 km/hour' counts as a violation and 'driving drunk' counts as a violation. In the logic, if we have that 'either someone drives 120 km/hour or he drives drunk', then this does *not* count as a violation. This phenomena can also be observed in reality. For example, in many legal courts someone cannot be sentenced if it is not clear which norm has been violated. There is only a violation if the norm which is violated can be identified. In such circumstances, if someone has committed a violation, but we do not know which one, then we cannot sanction him.

3.3 Two variants that disturb AND

There are several issues in this formalization of obligation in Definition 1. For example, the three conditions informally given in the introduction can be represented in another way, and additional conditions can be added. However, from the perspective of our logical analysis, all changes we have considered only lead to minor variations of the two propositions, and they do not interfere with the analysis. The following two definitions imply a small change to Proposition 2.

First, the formalization of 'the absence of p'. In clause (2), the absence of a is represented by $\neg a$. Consequently, if nothing is known then it does not count as a violation. An alternative way to formalize it is to use $not(a)$, where *not* is the negation by failure as used in logic programming.

Second, introduction of a particular perspective, such as the perspective of an external observer, of agent A or of the normative agent. For example, if everything is considered from the perspective of agent A, then we may write:

Definition 2 (Subjective obligations). *Consider a modal logic as before, with additionally a normal modal operator B_A for 'agent A believes ... '. Agent A believes to be obliged to see to it that p iff:*

$$\begin{array}{rll} BO_{AN}(p) =_{def} & B_A(D_N(p)) \wedge & \textit{(1)} \\ & B_A(\neg p) \rightarrow B_A(D_N(V_{NA}(\neg p))) \wedge & \textit{(2)} \\ & B_A(V_{NA}(\neg p)) \rightarrow B_A(D_N(S_{NA}(\neg p))) & \textit{(3)} \end{array}$$

Clearly, for obligations based on *not* operator and for subjective obligations, Proposition 1 still holds. Proposition 2 also holds, with the minor adaptation that AND no longer holds under these conditions (nor under any other reasonable conditions).

Moreover, for various variations of Definition 2, for example the one in which (2) would read $B_A(\neg p \rightarrow D_N(V_{NA}(\neg p)))$, Proposition 2 still valid, but for other variations, such as the one in which (2) reads $B_A(B_N(\neg p) \rightarrow D_N(V_{NA}(\neg p)))$, the adapted proposition holds. Summarizing, our analysis can directly be applied to such subjective obligations.

3.4 Four equivalent variations

In this section we discuss four more variations to the central definition, which do not influence our result.

First, the formalization of ‘if … then … ’ structures. In clause (2) and (3), they are represented by a material implication (within the desire modality), whereas it is well known that this is problematic. However, other conditional logics proposed in the literature are weaker than the material implication, such that the logic of O_{AN} can only become weaker.

Second, additional clauses that represent realism and other borderline cases. For example, we may add a clause that $O_{AN}(p)$ implies that p is consistent, or that $O_{AN}(p)$ implies that $\neg p$ is consistent. Such borderline cases do not influence the two propositions in any significant way.

Third, additional clauses that distinguish goals from desires (i.e., by introducing besides desires also goals), require that the normative agent does not desire violations (or desires that there are no violations), assume that the normative agent has at least one way to apply the sanction, etc. Again, for any reasonable additional clauses we have considered, such additional clauses only make the logic of O_{AN} weaker.

Fourth, in the following definition sanctions are made explicit. That is, we may say not only that $\neg p$ is sanctioned but also which sanction is applied. This leads to the introduction of an additional clause which says that the normative agent does not desire to apply the sanction anyway, i.e., even without a violation. Such rare cases are known, of course, but they are excluded in our model. The formalization of this new clause seems not completely satisfactory. We would have like to add the unconditional $D_N(\neg s)$. However, this unconditional clause is incompatible with our interpretation of D_N as a normality, because (3) and (4) together would imply $\neg V_{NA}(\neg p)$. In other words, (4) can only be formalized by $D_N(\neg s)$ if we adopt for D_N a non-normal modality, or a non-monotonic logic.

Definition 3 (Modal logic with explicit sanctions). *Consider a modal logic with modal operators D_N (for desire or goal) and V_{NA} (for counts as a violation). Obligation with explicit sanction is defined by:*

$$\begin{array}{lll} O_{AN}(p,s) =_{def} & D_N(p) \wedge & (1) \\ & \neg p \rightarrow D_N(V_{NA}(\neg p)) \wedge & (2) \\ & V_{NA}(\neg p) \rightarrow D_N(s) \wedge & (3) \\ & \neg V_{NA}(\neg p) \rightarrow D_N(\neg s) & (4) \end{array}$$

For fixed s, Proposition 1 and 2 both still hold, when in the latter the conditions on S_{NA} are dropped.

4 Decision theory

4.1 Normative systems

This section illustrates an area where our theory can be applied. The logical analysis has shown that there are many ways to formalize obligations in a modal logic of desires, counts as and being sanctioned. However, in the logical analysis of such obligations, the following pattern emerges. If V_{NA} does not trivialize, then the logic does not satisfy several proof rules which are often accepted in deontic logic. Now consider the following definition of a normative system.

Definition 4. *Let L be a propositional language. A normative system is a tuple $\langle N, V, S\rangle$ in which $N = \{n_1, \dots, n_k\}$ is a set of norms, V is a function that associates with every norm a formula of L called its violation, and S is a function that associates with every norm a propositional formula called its sanction.*

In this setting, we may say that the normative system implies the obligation $O(p, s)$ if there is a norm whose violation condition is $\neg p$. However, it is not very clear what the logical relations between these norms are, and what other methods we have to analyze the properties of such a system. If the norms are closed under for example weakening, then if this system would contain a norm with violation condition p would be equivalent to a normative system which contains the same norm, and moreover a norm with violation condition $p \wedge q$. Moreover, if the system is closed under conjunction and the system contains a norm with violation condition r, then the system is equivalent to a normative system which in addition contains a norm with violation condition $p \vee r$. But what does this equivalence mean? Moreover, such an account does not take the sanctions into account.

We propose the following idea. Given a set of obligations. If for every decision making context, adding a new obligation to this set of obligations does not influence the decision making of the agent, then this new obligation is already implied (or *accepted*) by the set of obligations.

4.2 Decisions

In this section we introduce decisions in the logic.

Definition 5 (Decision). *Let the atomic variables be partitioned into three sets A, the decision variables of agent A, N, the decision variables of the normative agent, and P, the parameters. A state of the world w is a propositional sentence. A decision d of agent A (N) in state w is a propositional formula built from A (N) only, such that $w \wedge d$ is consistent.*

We make several strong assumptions. A full qualitative decision theory has to incorporate a way to encode consequences of decisions. If we assume complete knowledge, i.e., the state of the world implies a truth value for each parameter, then we do not have to consider such effects, because effectively we only reason with ought-to-do obligations. An obligation $O_{AN}(p)$ is an ought-to-do obligation if p contains variables of A only, and an ought-to-be obligation otherwise.

Definition 6. *A state of the world contains complete knowledge if it implies either each variable of P, or its negation.*

With this new machinery, we can formalize the condition that the normative agent has a way to apply the sanction. We may formalize a new variant of our definition of obligation, with an additional clause is thus that there is a decision of N such that this decision implies sanction s. In our case, this means that s is a decision variable of N.

4.3 Decision rule

To evaluate its decisions, an agent may either consider the violations or the sanctions. This represents different agent types: an obedient or respectful agent considers its violations, whereas a selfish agent may only consider the sanctions.

Definition 7 (Decision evaluation). *Let w be a state of the world and d a (partial) decision. The set of violated norms is $Viol(w,d) = \{n \in N \mid w \wedge d \vdash V(n)\}$ and the set of sanctions is $Sanc(w,d) = \{S(n) \mid n \in Viol(w,d)\}$.*

The evaluations are used in the agent's decision rule, assuming that all sanctions have the same cost:

Definition 8 (Decision rule). *Given state of the world w. An obedient agent selects a decision d that minimizes (with respect to set inclusion) $Viol(w,d)$. A selfish agent minimizes (with respect to set inclusion) the logical closure of $Sanc(w,d)$.*

4.4 Acceptance

We analyze the normative system using the notion of acceptance.

Definition 9. *Given an agent type. A normative system accepts an obligation $O(p,s)$ if for any state of the world, adding to the normative system the norm n with violation $V(n) = \neg p$ and sanction $S(n) = s$, does not change the optimal decisions.*

We can consider the logical properties of the acceptance condition by abstracting away from the normative systems. The following proposition implies that the set of accepted obligations is not closed under weakening, strengthening, or conjunction. The results are in line with our logical analysis.

Proposition 3. *There is a normative system that accepts $O(a)$ but not $O(a \vee b)$ or $O(a \wedge b)$, and there is a normative system that accepts $O(a)$ and $O(b)$ but not $O(a \wedge b)$ or $O(a \vee b)$.*

5 Summary

In this paper we obtain the following results.

- We propose a logical framework to study social constructions.
- We define obligations in terms of this social construction, and study its properties.
- We define acceptance relations for normative systems.
- We contribute to the philosophical discussions on the distinction between violations and sanctions in Anderson's reduction, and between implicit and explicit normative systems.

Further relations between deontic logic and the theory of normative systems is subject of ongoing research, e.g., in [7] we consider the notion of strong permission. In [8] we consider the problem of norm creation.

References

1. A. Anderson. A reduction of deontic logic to alethic modal logic. *Mind*, 67:100–103, 1958.
2. G. Boella and L. Lesmo. A game theoretic approach to norms. *Cognitive Science Quarterly*, 2(3-4):492–512, 2002.
3. G. Boella and L. van der Torre. Attributing mental attitudes to normative systems. In *Procs. of AAMAS'03*, Melbourne, 2003. ACM Press.
4. G. Boella and L. van der Torre. Decentralized control obligations and permissions in virtual communities of agents. In *Procs. of ISMIS'03*, 2003. Springer Verlag.
5. G. Boella and L. van der Torre. Local policies for the control of virtual communities. In *Procs. of IEEE/WIC WI'03*, 2003.
6. G. Boella and L. van der Torre. Norm governed multiagent systems: The delegation of control to autonomous agents. In *Procs. of IEEE/WIC IAT'03*, 2003.
7. G. Boella and L. van der Torre. Permissions and obligations in hierarchical normative systems. In *Procs. of ICAIL'03*, Edinburgh, 2003. ACM Press.
8. G. Boella and L. van der Torre. Rational norm creation: Attributing Mental Attitudes to Normative Systems, Part 2. In *Procs. of ICAIL'03*, Edinburgh, 2003. ACM Press.
9. M.Bratman. *Intentions, plans, and practical reason.* Harvard University Press, Harvard (MA), 1987.
10. 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.
11. J. Carmo and O. Pacheco. Deontic and action logics for collective agency and roles. In *Proc. Fifth International Workshop on Deontic Logic in Computer Science (DEON'00)*, pages 93–124, 2000.
12. C.Castelfranchi, F. Dignum, C. M. Jonker, and J. Treur. Deliberate normative agents: Principles and architecture. In *Intelligent Agents VI - Procs. of ATAL'99*, 2000. Springer Verlag.
13. Chellas. *Modal logic: an introduction.* Cambridge University Press, Cambridge (UK), 1980.
14. R. Conte, C. Castelfranchi, and F. Dignum. Autonomous norm-acceptance. In *Intelligent Agents V - Procs of ATAL'98*, pages 319–333. 1999. Springer Verlag.
15. D. Dennett. *The intentional stance.* Bradford Books, Cambridge (MA), 1987.
16. F. Dignum, D. Morley, E. A. Sonenberg, and L. Cavedon. Towards socially sophist icated BDI agents. In *Procs. of ICMAS'00*, pages 111–118, Boston, 2000.
17. A. Jones and M. Sergot. A formal characterisation of institutionalised power. *Journal of IGPL*, 3:427–443, 1996.
18. S. Kraus, K. Sycara, and A. Evenchik. Reaching agreements through argumentation; a logical model and implementation. *Artificial Intelligence*, 104:1–69, 1998.
19. J. J. Ch. Meyer. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame J. of Formal Logic*, 29(1):109–136, 1988.
20. A. Newell and H. Simon. *Human Problem Solving.* Prentice-Hall, 1972.
21. A. Rao and M. Georgeff. Modeling rational agents within a BDI architecture. In *Procs. of KR'91*, pages 473–484. 1991. Morgan Kaufmann.
22. J. Searle. *The Construction of Social Reality.* The Free Press, New York, 1995.
23. L. van der Torre and Y. Tan. Contrary-to-duty reasoning with preference-based dyadic obligations. *Annals of Mathematics and AI*, 27:49–78, 1999.
24. L. van der Torre and Y. Tan. Diagnosis and decision making in normative reasoning. *Artificial Intelligence and Law*, 7(1):51–67, 1999.