A Gentzen System for Reasoning with Contrary-To-Duty Obligations. A Preliminary Study

Guido Governatori
School of Information Technology and Electrical Engineering
The University of Queensland
Brisbane, QLD 4072, Australia
email: guido@itee.uq.edu.au

Antonino Rotolo
CIRSFID, University of Bologna,
Via Galliera 3, I-40121 Bologna, Italy
email: rotolo@cirfid.unibo.it

Abstract

In this paper we present a Gentzen system for reasoning with contrary-to-duty obligations. The intuition behind the system is that a contrary-to-duty is a special kind of normative exception. The logical machinery to formalize this idea is taken from substructural logics and it is based on the definition of a new non-classical connective capturing the notion of reparational obligation. Then the system is tested against well-known contrary-to-duty paradoxes.

1 Introduction

One of the main themes in the philosophical discussion on deontic logic is about reasoning with contrary-to-duty (CTD) obligations. In this perspective, it is widely acknowledged that the crisis of Standard Deontic Logic (SDL) is strongly related to the formulation of some notorious paradoxes centering around the regulation of the violation of obligations. Puzzles like Chisholm's and Forrester's paradoxes, Reykjavik scenario and Möbius strip example, depict situations where various combinations of reparational and unconditional obligations give rise to logical contradictions or counterintuitive conclusions. As a matter of fact, a great part of the efforts in deontic logic have been driven by solving these problems and a plethora of different strategies have been accordingly proposed. A full analysis of all these contributions is obviously beyond the scope of this paper.

However, we believe that some of those approaches deserve to be considered here in some detail. We refer in particular to the works whose starting-point can be summarized in the following thesis: "no logic of norms without attention to the normative systems in which they occur" [12, p. 32]. Even though this idea at first sight seems to be obvious, it is greatly valuable since it proposes what we could call a "holistic reading" of normative reasoning. Actually, we think this intuition is fundamental for at least two reasons. Firstly and generally speaking, it makes closer areas often too far such as philosophy of norms (and, more specifically, philosophy of law) and deontic logic. It is quite odd, for a legal philosopher, to conceive of norms in isolation. Norms interplay each other. Thus, a normative set can (must) have different meanings and may (should) contribute to diverse conclusions if it is included in distinct normative systems. The "spirit" of such norms changes according to their systematic reading. Secondly, thanks to this approach to normative reasoning, it is possible to give both a simple and appealing account of CTD obligations and consequently solutions to the just mentioned paradoxes of deontic logic.

In a wide sense, significant examples in this direction are some papers by H. Prakken and M. Sergot [15, 16]. Basically, they regard CTD structures as contextual obligations, that is obligations strictly relative to a certain context of application. Accordingly, they are not just conditional obligations which hold without restriction and so factual detachment is not in general permitted. On the other hand, it must be the case that primary obligations related to CTDs are still in force at least outside their specific context of violation. Thus, the authors argue that some cases are to be inconsistent, in particular when a CTD norm states a reparational obligation which is in contradiction with an another primary obligation in the system. In logical terms, this idea has been first implemented by the so-called principle of "downwards inheritance" for checking the unrelatedness between contexts and primary obligations [15]. Later, they developed a peculiar semantical construction to characterize a preference ordering over the worlds which is strictly sensitive of the "number" of reciprocal incompatibilities (potential violations) between norms.

A different approach, inspired by similar intuitions, has been developed in particular by D. Makinson and L. van der Torre. Their main idea, as pointed out by Makinson [12] himself, is to be traced back to a pionieristic work by Stenius [19] and it has been later improved by Alchourrón and Bulygin [2, 1]. This line of investigation is based on the intuition that any obligation can be thought in terms of a *consequence relation* of what is explicitly stated as obligatory in a normative system. Actually, Makinson and van der Torre's approach is a further step in this direction. In particular, if their analysis is meant to capture this original idea, on the other hand it claims to impose some constraints on the manipulation of conditional norms. As expected, some restrictions are required both on strengthening of antecedent and on transitivity, since this is vital in CTD contexts. A related point thus concerns the directionality of normative conditionals. It is commonly acknowledged that contraposition cannot be accepted. For the consequence relation under which a normative system is closed is not classical but is to be modelled by permitting only a directional

iterative detachment of obligations. The conditions of such a detachment are in turn strictly connected with performing a consistency check in the normative system. More precisely, the detachment of an obligation B can only be obtained by using the regulations which are consistent with the condition stated for B. If it is not the case, then B is a not a consequence of the normative system. This task can be done within a labelled deductive system based on the so-called Input-Output logic developed by L. van der Torre [20, 13, 14].

Basically, our system starts from this last conception of normative reasoning. First, it is based on a purely syntactic view of deontic logic so that all the machinery consists of defining a suitable consequence relation for dealing with norms. Second, some intuitive conditions are required to capture adequately the global interplay between the norms included in a given normative system. In particular this is done by introducing a new logical operator of "normative reparation" in order to make explicit the relation between primary obligations and their related CTDs and to combine them in single regulations. This will allow us to give a plausible reading of CTD structures.

In what follows we first argue that, logically, CTD obligations are a special kind of exceptions (Section 2). Then we propose a Gentzen system specially tailored to cope with the above intuition. In particular, as usual with Gentzen systems, we provide general inference rules for the introduction and the elimination of a non-classical connective intended to capture the meaning of CTD structures (Section 3). Before introducing the formal notions of normative system, ideal situation, violation, etc., in Section 5, we present some of the most common instances of the inference rules, and we shortly discuss them in relation to well-know patterns of normative reasoning (Section 4). At this point we have all the formal machinery needed to examine in depth some of the most important CTD paradoxes (Section 6). We conclude the paper with some insights about possible extensions of the system, such as the definition of a normative consequence relation for the notion of permission.

2 The Main Intuition of Our Approach

What is a contrary-to-duty obligation? The common reading suggests that is nothing but a reparational obligation of a violated norm; accordingly, it is in force only when a violation occurs. In this paper we would like to argue that a CTD obligation can be conceived of in a slightly different way, namely as special kind of normative exception.

What does it mean that a CTD obligation is an exception?

Norms are by definition violable: a norm which cannot be violated is meaningless or, at least, seems to be useless. If a norm says that it is obligatory to kill or not to kill, then the norm says nothing. It is not a reason to act. In other words, norms do not concern simply what should be the case in any ideal situation but they should be open to their violation. This intuition is widely accepted but we feel it has not been fully investigated from a logical point of view. If we look at realistic normative domains (e.g., law) we realize that the

obligation not to kill is usually rendered as a norm stating an appropriate sanction which ought to follow in case of violation. Actually, it is not by chance that H. Kelsen [11] talks about legal obligations, that he calls primary norms, as norms stipulating sanctions. A similar approach can be found in the analysis of deontic logicians like A.R. Anderson [3] who define ought-assertions as obligations to do something or to repair their violations by means of sanctions. We are aware that this is one of the most discussed issues in contemporary philosophy of norms (for a recent overview, see, e.g., [22]) since it concerns hard problems such as the very nature of conditional obligations. However, besides the plethora of different opinions on this matter, a point seems intuitively to be clear. If a norm is categorical, then it does not admit violations. In logical terms this means that is to be *impossible* to derive from it a secondary obligation. Otherwise, it is not categorical at all. In the case of CTD structures we are not dealing with this kind of norms but with different normative domains. Of course, a norm-giver who makes norms as obligations conditioned to sanctions is trying first to state what is obligatory. On the other hand, he/she is to be ready to reply to violations. The notion of CTD norms as exceptions is clear if we reason from the point of view of the addressee of a norm. In this case, norms like these can be interpreted in terms of alternative reasons to act: do x or you will be sanctioned. Actually the addressee has two logical options. Even the second can be acceptable; however, since it is a sanction, it has to be considered as a normative exception to the primary obligation. Something similar holds also in the perspective of the norm-giver. In fact, he/she has to impose a fair and proportional sanction for the violation of a given obligation; in this way, any action which is a violation of a primary obligation must be understood as an "exceptional action" with respect to what is obligatory.

In this perspective, a CTD obligation (1) is a special kind of logical exception of the normative content of a primary obligation, and (2) is not a usual conflicting obligation which overrides such a primary obligation. As we shall see, an immediate consequence of this thesis is that a primary obligation and its CTDs can and must give raise to a unique norm, expressing the true meaning of the normative content they define in a given normative system.

Given this general background, let us see in detail why and how CTD norms can be logically thought of as special exceptions of primary obligations. According to the usual logical account, a norm with exception can be represented as

$$\begin{array}{rcl} A & \Rightarrow & B \\ A, E_1 & \Rightarrow & \neg B \\ & \vdots & \\ A, E_n & \Rightarrow & \neg B \end{array}$$

Let us now consider $\Rightarrow$ as a sub-structural consequence relation $\vdash$ without the structural rules of contraction, duplication, and exchange. The main reason for this choice is that we want to investigate the very nature of normative consequence without any commitment to the classical interpretation. In this perspective the comma does not correspond to the classical conjunction on the

left side of $\vdash$ and the classical disjunction on the right side. Thus the meaning of the expression

$$A_1, \ldots, A_n \vdash B_1, \ldots, B_m$$

is: the sequence $A_1, \ldots, A_n$ comports that B_1 is the case; but if B_1 is not satisfied, then B_2 should be the case, and so on. In a normative context, this means that the content of the obligation determined by the conditions $A_1, \ldots, A_n$ is B_1; however the violation B_1 can be repaired by B_2 and so on.

Now, let us consider the standard rules for negation, that is:

$$\frac{A, B \vdash C}{A \vdash \neg B, C} \qquad \frac{A \vdash B, C}{A, \neg B \vdash C}$$

If $\neg$ is an involutive operator (i.e., $\neg\neg A \equiv A$), the effect of these rules is to move a formula on the other side of $\vdash$, changing the polarity. Accordingly, given the norm

$$A \Rightarrow B \tag{1}$$

and its exception

$$A, \neg B \Rightarrow C \tag{2}$$

we can obtain

$$A \Rightarrow B, C \tag{3}$$

applying the rule for the negation on (2).

The norm in (1) says that B should be the case when the condition A obtains. According to (2) C should be the case given A and $\neg B$; thus (2) is, at the same time, a CTD obligation and an exception of (1). In a classical reading of (3), "," would correspond to the classical disjunction. However, this is not the case in the present interpretation, where, *intuitively*, the expression on the right side of $\vdash$ in (3) can be thought of as

$$B \vee (\neg B \rightarrow C) \tag{4}$$

Hence the norm in (3) subsumes the norms in (1) and (2). In other words it states that, given A, B ought to be the case; otherwise, under the same condition A, C is obligatory. Therefore, not surprisingly, (3) is a CTD obligation of (1).

What we have just said gives us the possibility to deal with CTD reasoning within a purely syntactic framework. The next section provides a formalization of CTDs in terms of a Gentzen system for the non-classical connective $\otimes$, corresponding to the "," on the right side of a normative consequence relation $\vdash$ characterizing obligations. Given the intended interpretation of $A \otimes B$ as "B is the reparation of A", the connective $\otimes$ permits to combine primary and CTD obligations into unique regulations. It has been argued that violations are different from exceptions [15, 21]. We think this analysis is correct insofar as it maintains that a norm is still in force even when is violated, whereas a default like 'birds fly' is cancelled, e.g., by the fact that Tweety does not fly. As a matter of fact, it is quite odd to say that an obligation is *cancelled* by its violation. On the other hand, our idea that a CTD is a special kind of

exception does not mean that the primary obligation has to be cancelled or even overriden by its CTDs. As we shall see, we do not introduce any kind of machinery to account for the overriding of a nprimary norm by its CTDs. We simply argue that a normative system containing primary and CTD obligations actually gives its addressees the possibility to comply with either primary or, as exceptions, secondary (tertiary, etc.) obligations. Obviously, compliance with primary norms or their CTDs are not put at the same level, but refer to different degrees of ideality. In this perspective, it should be noted in advance that the introduction of $\otimes$ can be done iteratively depending on the number of levels of ideality determined by the chains of CTDs contained in the normative system. This is in a way the sytactic counterpart of the thesis, quite common in the DL community, that CTDs are semantically rendered in a preference (ordering) semantics, where the order among sets of worlds expresses different levels of ideality and violability.

3 A Gentzen System for CTD Obligations

First of all, let us define our formal language L. It consists of a countable set of atomic formulas. Well-formed-formulas are then defined using the unary connective $\neg$ (negation) and the binary connective $\otimes$ which is intended to formalize CTD statements.

Definition 1 *Let $\vdash_O$ be a binary consequence relation defined over $\mathcal{P}(WFF) \times WFF$. Thus the expression $\Gamma \vdash_O A$ is a sequent where Γ (the antecedent) is a finite (possibly empty) set of formulas and A is a formula.*

As usual in Gentzen systems the meaning of operators and connectives is given by the rules for their introduction and elimination (cf., e.g., [17]). More precisely, this is true in the presence of the structural rules of exchange, duplication and contraction. Otherwise, the introduction and elimination rules have to be supplemented by rules for the "structural" meaning of the operators involved [8, 18].

According to Definition 1 the usual rules of contraction, duplication and exchange hold trivially for the formulas in the antecedent. However, they do not make any sense for the consequent so that we need properties describing the structural behaviour of $\neg$ and $\otimes$.

The only property we assume for $\neg$ is that it is an involutive operator, i.e., $\neg\neg A \equiv A$ for any formula A; while the basic logical properties for $\otimes$ are the following:

1. $A \otimes (B \otimes C) \equiv (A \otimes B) \otimes C$
2. $\bigotimes_{i=1}^{n} A_i \equiv (\bigotimes_{i=1}^{k-1} A_i) \otimes (\bigotimes_{i=k+1}^{n} A_i)$ where $A_j = A_k$ and $j < k$

Condition 1 is just associativity of $\otimes$, while condition 2 corresponds to duplication and contraction. In fact, according to the intuitive reading of this connective given in the previous section, the expression on the right side of $\vdash$

can be considered as an ordered set. However, the full meaning of the operator $\otimes$ is given by a rule for its introduction ($\otimes$I) and the corresponding rule for its elimination ($\otimes$E). Thus, let us see its logical characterization wrt the normative consequence relation $\vdash_O$.

$$\frac{\Gamma \vdash_O A \otimes (\bigotimes_{i=1}^{n} B_i) \otimes C \qquad \Delta, \neg B_1, \ldots, \neg B_n \vdash_O D}{\Gamma, \Delta \vdash_O A \otimes (\bigotimes_{i=1}^{n} B_i) \otimes D} \quad (\otimes\text{I})$$

$$\frac{\Gamma \vdash_O A \otimes B \otimes C \qquad \Delta \vdash_O A \otimes \neg B \otimes D}{\Gamma, \Delta \vdash_O A \otimes C} \quad (\otimes\text{E})$$

To complete the formal description of the system we have to give the conditions for $\top$ –an always true formula– and for $\bot$, a generic formula for a contradiction (or normative conflict).

$\top$ is defined in terms of $\otimes$; more precicesly

$$A \otimes \neg A \equiv \top.$$

This formula states that the reparation of A is $\neg A$; but a reparation occurs when the thing it repairs fails, so $\neg A$ should be the case when $\neg A$ is the case. Thus $A \otimes \neg A$ is fulfilled when we have either A or $\neg A$; in each state of affairs we have either A or $\neg A$, so any state of affairs satisfies $A \otimes \neg A$. It is immediate to see that

$$A \otimes \top \equiv \top \otimes A \equiv \top;$$

accordingly it is reasonable to stipulate that

$$A \otimes \bot \equiv \bot \otimes A \equiv A.$$

The following rule is devised for making explicit conflicting norms (contradictory norms) within the system:

$$\frac{\Gamma \vdash_O A \qquad \Delta \vdash_O \neg A}{\Gamma, \Delta \vdash_O \bot} \quad (\text{I}\bot)$$

where

1. there is no sequent $\Gamma' \vdash_O X$ such that either $\neg A \in \Gamma'$ or $X = A \otimes B$; and
2. there is no conditional norm $\Delta' \vdash_O X$ such that either $A \in \Delta'$ or $X = \neg A \otimes B$; and
3. for any formula B, $\{B, \neg B\} \not\subseteq \Gamma \cup \Delta$.

The meaning of these three conditions is that given two conditional norms (sequents), we have a conflict if the normative content of the two norms is opposite, such that none of them can be repaired, and the the states of affairs they require are consistent.

The last aspect of the system we want to deal with is the relation of subsumption between two sequents.

Definition 2 *Let* $n_1 = \Gamma \vdash_O \bigotimes_{i=1}^{m} A_i \otimes B$ *and* $n_2 = \Gamma' \vdash_O C$ *be two sequents. Then* n_1 subsumes n_2 *iff*

1. $\Gamma = \Gamma'$ *and* $\bigotimes_{i=1}^{m} A_i = C$*; or*
2. $\Gamma \cup \{\neg A_1, \ldots, \neg A_m\} = \Gamma'$ *and* $B = (C \otimes D)$.

The idea behind this definition is that the normative content of the norm n_2 is fully included in the norm n_1. Thus n_2 does not add anything new to the system and it can be safely discarded.

4 Commentary and Examples

The inference rules introduced in the previous section allow us to characterize formally the notion of CTD obligation with respect to $\vdash_O$. They are presented there in the most general version. In order to make clearer their intuitive meaning, in this section we will give to the reader some simplified variants which correspond to intuitive situations in which CTDs may occur.

Let us consider a norm like

$$\Gamma \vdash_O A.$$

Given an obligation like this, if we have that

$$\Delta, \neg A \vdash_O C,$$

then the latter must be a good candidate as reparational obligation of the former. This idea is formalized is as follows:

$$\frac{\Gamma \vdash_O A \qquad \Delta, \neg A \vdash_O C}{\Gamma, \Delta \vdash_O A \otimes C}$$

According to this view, if there exists a conditional obligation whose antecedent is the negation of the propositional content of a different norm, then the latter is a reparational obligation of the former. In this way, the CTD obligation can be forced to be an *explicit reparational obligation* with respect to the violation of its primary counterpart. Accordingly, it seems to be reasonable to discard both premises when they are subsumed by the conclusion. Their reciprocal interplay makes them two related norms so that they cannot be viewed anymore as independent obligations. Notice that if Γ and Δ are empty, then we are dealing with the basic case in which the primary obligation has the format of an apparently categorical obligation.

As we have alluded to above, the rule $\otimes$I can also generate chains of CTDs in order to deal iteratively with violations of reparational obligations. The following case is just an example of this process.

$$\frac{\Gamma \vdash_O A \otimes B \qquad \neg A, \neg B \vdash_O C}{\Gamma \vdash_O A \otimes B \otimes C}$$

Chains of CTDs can be manipulated in different ways. An interesting case is when other reparations are added inside a sequence of CTDs built via the $\otimes$

operator. This is possible since *any* conditional norm can be combined with a different obligation insofar as the former regulates the violation of the latter. Given an obligation we may thus infer more than a single explicit new $\otimes$-norm conditioned to its violation: in fact, a norm-giver can stipulate different reparations for a particular violation. The presence of such new regulations in the normative system is equivalent to saying that it is obligatory to fulfil the conjunction of several CTD obligations if the same violation occurs. More precisely, even if a primary obligation can be discarded after some applications of $\otimes$I, another explicit CTD regulation can be drawn with respect to the first obligation of the chain of reparational obligations we have already in the system:

$$\frac{\Gamma \vdash_O A \otimes B \otimes C \qquad \Delta \neg A \vdash_O D}{\Gamma, \Delta \vdash_O A \otimes D}$$

What about disjunctions of CTDs? It is quite common in our every-day experienceto tackle situations where different obligations can repair alternatively to the violation of the primary obligation. Suppose John eats a piece of cake even though his mother commanded him not to touch it, since it is for some guests invited for dinner. When she realizes that John has eaten the cake she could say to John: buy another cake or apologize for your bad action! As a matter of fact, both secondary obligations can repair alternatively to the violation of the primary obligation. Situations like this are far from being unusual also in legal contexts. It is not hard to find examples, at least in most western countries, where the legislator states different sanctions for certain kind of crimes as alternatives to the prison. Actually, the system we provided seems to be unable to capture these cases for the trivial reason that it is based on a language which does not include the boolean connectives. However, something very close to a disjunctive obligation can be represented when the normative system permits to get the symmetry of two obligations wrt $\otimes$:

$$\frac{\Gamma \vdash_O A \quad \neg A \vdash_O B}{\Gamma \vdash_O A \otimes B} \qquad \frac{\Gamma \vdash_O B \quad \neg B \vdash_O A}{\Gamma \vdash_O B \otimes A}$$

In this case, A and B repair each other, so that it can be said that it is obligatory to do A or B if Γ holds.

Let us see now the rule $\otimes$E. To understand its intuitive meaning, it is useful to look at a couple of derived rules. First of all, consider its trivial instance when $n = 2$:

$$\frac{\Gamma \vdash_O A \otimes B \qquad \Delta, \neg A \vdash_O \neg B}{\Gamma, \Delta \vdash_O A} \tag{5}$$

Informally, if the normative system contains both a reparational obligation of A and a norm stating the negation of such a reparation as a CTD obligation of the violation of A, then each of the two secondary obligations makes meaningless the other as a true reparation of A. Notice that these norms do not generate a contradiction: both premises are consistent with the original primary obligation A. This fact should not be strange: a "contradiction" between two secondary obligations conditioned to the violation of the same primary obligation A is

nothing but a (perhaps bizarre) way for restating A as obligatory. The presence of B and its negation as CTDs of A is in a way irrelevant for A. For similar reasons, we can derive a rule like the following:

$$\frac{\Gamma \vdash_O A \otimes B \qquad \Delta \vdash_O A \otimes \neg B}{\Gamma, \Delta \vdash_O A} \tag{6}$$

In general we have to distinguish between genuine normative conflicts from apparent ones. By normative conflict we mean any situation ruled by opposite norms and which results in an impossible state of affairs; or, in other words, a situation in which the normative content of all relevant norms cannot be fulfilled, ending inevitably in a violation that cannot be repaired.

The simplest case of conflict of norms obtains when only two categorical obligations are given, that is, when we have both $\vdash_O A$ and $\vdash_O \neg A$. It is immediate to see that we can apply I$\bot$, thus deriving $\vdash_O \bot$.

Let us consider the following patterns of apparent conflicts. In the case of

$$A \vdash_O B \qquad \neg A \vdash_O \neg B$$

the conflict is apparent because the conditions of application of the two norms are mutually exclusive; thus a situations where both norms are applicable does not exist.

On the other side, given

$$\vdash_O A \qquad \vdash_O \neg A \qquad \neg A \vdash_O B$$

we have two conflicting categorical obligations. However, a closer analysis shows that actually one of them is not categorical insofar as it admits a CTD. Thus the situation where the CTD obligation is in force is still normatively acceptable, even if the corresponding primary obligation is violated. But in this case the other categorical obligation is fulfilled.

This pattern also shows that the system at hand is nonmonotonic: the presence of $\neg A \vdash_O B$ prevents the application of I$\bot$. Hence $\vdash_O A$ and $\vdash_O \neg A$ no longer derive $\vdash_O \bot$.

The above discussion points out that the only conflicts we have to worry about are the so called genuine conflicts. Those conflicts indicate that part of the normative system they are part of is not rational. In idealized situations they should not occur. Unfortunately this is seldom the case in real-life. Thus methods to restore rationality should be devised. Indeed, many and many of them have been put forth, and it is beyond the scope of the paper to investigate such a topic. However, unlikely more traditional treatments of CTDs, the present approach considers CTDs and normative conflicts just as two orthogonal aspects of normative reasoning. Accordingly the interested reader can try to plug-in it her preferred way to deal with (genuine) conflicts.

5 Consequence Relations and Normative Systems

Now we need to introduce a formal definition of normative system. We distinguish between normative codes and normative systems. The former can just be considered as the set of explicitly promulgated norms, while its related normative system is obtained from the normative code by adding principles to derive other norms. Formally:

Definition 3 *Let D be a set of deontic notions (e.g., obligation, permission, etc).*

- *A* Normative Code *is a set $\{S_i\}_{i\in D}$ where each S_i is a finite set of norms.*
- *An* Implicit Normative Systems *is a set $\{(S_i, \vdash_i)\}_{i\in D}$, where each S_i is a finite set of sequents, and each $\vdash_i$ is a normative consequence relation for i.*
- *An* Explicit Normative Systems *is a set $\{\uparrow (S_i, \vdash_i)\}_{i\in D}$, where each $\uparrow (S_i, \vdash_i)$ is the least fixed point (if it exists) of the closure under $\vdash_i$ and subsumption of S_i.*

This is a very general definition of normative system. One of the main advantages of explicite normative systems resides in the fact that complete meaning and content of a norm is entirely encoded in the formulation of the norm itself and not scattered around the normative systems. In fact, in this paper, we consider the normative system obtained from the deontic notion of obligation and the corresponding normative consequence relation we have investigated in Section 3. Some insights about the integration of obligation and permission will be given in Section 7.

We are now ready to give conditions under which we are able to determine whether a state of affairs is compatible with a normative system or it represents a violation of some norms. To this end, we shall consider a state of affairs as a set of literals; moreover we will restrict ourselves to the case where all the formulas made explicit in the norms (sequents) of a normative system are literals as well. Notice that this choice does not allow us the use of expressions like $\neg(A \otimes B)$ on the right side of $\vdash_O$ nor occurrences of $\otimes$ in the antecedents of the sequents.

We are aware that this is a debatable limitation. However, the intuitive meaning of $\neg(A \otimes B)$ is unclear, or at least it seems to admit several possible interpretations. What does it mean that B is not a reparation of A?[1] Until we have a precise answer to this question we prefer not to commit ourselves to any particular interpretation; therefore we do not give the logical meaning of negation. Indeed, the introduction and elimination rules for $\neg$ have not been given. Moreover, it is not easy to give an intuitive account of formulas such as $A \otimes B$ if they occur on the left side of the consequence relation. A possible

[1] See Section 7 for some intuitions about the relation between $\neg$ and $\otimes$ wrt an explicit consequence relation of permission.

interpretation could be that such occurrences mean something like: "It is a fact that B is a reparational obligation of A". However, we prefer here to refrain from presenting solutions to these problems. Of course, these are matter of further investigations, but we feel that they can be resolved in a satisfactory way as soon as a suitable machinery for reasoning *about* norms is introduced in the system.

First of all we define when a state of affairs is either ideal, sub-ideal, or non-ideal with respect to a norm. Then we extend this notions both to explicit and to implicit normative systems.

Definition 4

- *A state of affairs s is* ideal *wrt to a sequent (norm)* $\Gamma \vdash_O A_1 \otimes \cdots \otimes A_n$ *iff if* $\Gamma \subseteq s$*, then* $A_1 \in s$*.*
- *A state of affairs s is* sub-ideal *wrt to a sequent (norm)* $\Gamma \vdash_O A_1 \otimes \cdots \otimes A_n$ *iff if* $\Gamma \subseteq s$ *and* $\exists A_i$*,* $1 < i \leq n$ *such that* $\forall A_j$*,* $j < i$ $\{\neg A_1, \ldots, \neg A_j\} \subseteq s$*, then* $A_i \in s$*.*
- *A state of affairs s is* non-ideal *wrt to a sequent (norm)* $\Gamma \vdash_O A_1 \otimes \cdots \otimes A_n$ *iff it is neither ideal nor sub-ideal.*

According to Definition 4, a situation is ideal wrt to a norm if the norm is not violated; sub-ideal when the primary obligation is violated but the norm admits a reparation, which is satisfied; non-ideal when the primary obligation and all its reparations are violated. This definition can be easily extended to the case of explicit normative systems:

Definition 5

- *A state of affairs s is* ideal *wrt to an explicit normative system iff there is no norm in the system for which s is either sub-ideal or non-ideal.*
- *A state of affairs s is* sub-ideal *wrt to an explicit normative system iff there is a norm for which s is sub-ideal, and there is no norm in the system for which s is non-ideal.*
- *A state of affairs s is* non-ideal *wrt to an explicit normative system iff there is a norm for which s is non-ideal.*

Definition 5 follows immediately from the intuitive interpretation of ideality and of the related notions we have provided in Definition 4. On the other side, the relation between an explicit normative system and the implicit one from which is obtained seems to be a more delicate matter. A careful analysis of the conditions for constructing an explicit normative system allow us to state the following general criterion:

Definition 6 *A state of affairs s is ideal (sub-ideal, non-ideal) wrt an implicit normative system N* if *s is ideal (sub-ideal, non-ideal) wrt the explicit normative system obtained from N.*

It is worth noting that Definition 6 shows the relevance of the distinction explicit and implicit normative systems. This holds in particular for the case of sub-ideal situations. Suppose you have an implicit normative system consisting of the norms

$$\vdash_O A \qquad \neg A \vdash_O B$$

The corresponding explicit normative system is

$$\vdash_O A \otimes B$$

While the state of affairs $s = \{\neg A, B\}$ is sub-ideal wrt to the latter, it would be non-ideal for the former. In the first case, even if $\neg A \vdash_O B$ expresses in fact an implicit reparational obligation of $\vdash_O A$, this is not made explicit. So, there exists a situation which apparently accomplishes a norm and violates the other without satisfying any reparation. This conclusion cannot be accepted because it is in contrast with our intuition according to which the presence of two norms like $\vdash_O A$ and $\neg A \vdash_O B$ must lead to a unique regulation. For this reason, we can evaluate a situation as sub-ideal wrt an implicit normative system only if is sub-ideal wrt its explicit version.

Finally, let us define the notion of *ought*. It is intended to formalize what any explicit normative system requires as obligatory, if a state of affairs is given.

Definition 7 *Given a state of affairs s and an explicit normative system N, $Ought(s)$ is a set of sets of literals $O(s) - s$ such that for each $O(s)$:*

- *$s \subseteq O(s)$; and*
- *$O(s)$ is one of the smallest sets of literals such that $O(s)$ is at least sub-ideal wrt N; and*
- *$O(s)$ does not contain a literal and its negation.*

This definition is meant to capture the best possible alternatives to a given situation. It also provides a semantics for $\vdash_O$ and $\otimes$. Let $\Gamma \vdash_O A$ be a sequent, and let s be the smallest state of affairs satisfying Γ. Then s satisfies A iff $Ought(s)$ contains a set $O(s)$ which is at least sub-ideal with respect to A. The above construction does not distiguish the degree of ideality between states of affairs. It only says whether complex obligations are fulfilled or violated by some states of affairs. For example given the empty state of affairs and the norm

$$\vdash_O A \otimes B,$$

both $\{A\}$ and $\{\neg A, B\}$ are in $Ought(\emptyset)$. Therefore we have to identify the most ideal situations: in the case at hand $\{A\}$ beacuse it is ideal, while $\{\neg A, B\}$ is sub-ideal.[2] Notice that in general is not possible to determine the most ideal

[2] A possible solution for this problem is to supplement the definition of satisfiability by adding a degree of violation similar to the degree of diasappointment proposd by Brewka, Benferhat and Le Berre [5] for their logic of ordered disjunction. However a careful analysis of this topic is left as a matter of future work.

situation. Let us consider the following normative system

$$\vdash_O A \otimes (B \otimes C) \quad \vdash_O A \otimes (C \otimes B)$$

As we have seen in Section 4, given $s = \{\neg A\}$, both B and C are reparations of A, as well as the reparation of each other. Thus the two states of affairs $s_1 = \{\neg A, B\}$ and $s_2 = \{\neg A, C\}$ are both in $Ought(s)$. It is immediate to see that s_1 is sub-ideal wrt the first norm, while for the second norm every extension containing C or $\neg C$ will be sub-ideal wrt it. Similarly for s_2.

Besides what we said in Section 4 about the consequence relation $\vdash_O$, it is worth noting that also the notion of *ought* exhibits a nonmonotonic behavior. In fact, if we consider $\vdash_O$ as a connective, *ought* can be viewed in terms of a consequence relation where $Ought(s)$ follows from a normative system N and a set of states of affairs. If so, not only would different normative systems imply trivially diverse $Ought(s)$, but, given the same N, different states of affairs (and different violations) could give as well distinct "oughts". This confirms van der Torre and Tan's [21] thesis that violability has to be read as special kind of defeasibility.

In very general terms, our formulation follows the intuition of Jones and Pörn [9, 10] insofar as it permits to represent the *real (actual) obligations* expressed by the system. However, our approach is based on purely syntactical notion of ideality and is strictly related to the role of the operator $\otimes$. In this way, it does not suffer of some drawbacks of Jones and Pörn's analysis such as the necessity of introducing hierarchies of sub-sub-ideal, sub-sub-sub-ideal worlds and so forth.

6 Dealing with CTD Paradoxes

Now, let us see how to deal in our system with some of the most infamous paradoxes of CTD reasoning. In particular, we want to give a formal account of Chisholm's [6] and Forrester's [7] paradoxes, Belzer's [4] "Reykjavik scenario" and Makinson's [12] "Möbius strip example". Since these puzzles are well-known in the deontic logic community we shall not recall any of their intuitive examples but we will confine our analysis to their logical representation in our formalism.

Chisholm's Paradox The basic scenario depicted in Chisholm's paradox corresponds to the following implicit normative system:

$$\{\vdash_O h,\ h \vdash_O i,\ \neg h \vdash_O \neg i\}$$

plus the situation $s = \{\neg h\}$. First of all, note that the system does not determine in itself any normative contradiction. This can be checked by making explicit the normative system. In this perspective, a normative system consisting of the above norms can allow only for the following inference:

$$\frac{\vdash_O h \qquad \neg h \vdash_O \neg i}{\vdash_O h \otimes \neg i} \tag{7}$$

Thus, the explicit system is nothing but

$$\{h \vdash_O i,\ \ \vdash_O h \otimes \neg i\}$$

It is easy to see that s is ideal wrt the first norm. On the other hand, while s is not ideal wrt $\vdash_O h \otimes \neg i$, we do not know if it is sub-ideal wrt such a norm. Then, we have to consider the two states of affairs $s_1 = \{\neg h, i\}$ and $s_2 = \{\neg h, \neg i\}$. It is immediate to check that s_1 is non-ideal in the system whereas s_2 is sub-ideal. If so, given s, we can conclude that the normative system says that $\neg i$ ought to be the case (see Definition 7).

The Gentle Murderer Let us see now the logical structure of the implicit system of norms which corresponds to Forrester's scenario:

$$\{\vdash_O \neg k,\ k \vdash_O g\}$$

Even in this case, we have a single application of $\otimes$I:

$$\frac{\vdash_O \neg k \qquad k \vdash_O g}{\vdash_O \neg k \otimes g}$$

so that the explicit normative system is trivially as follows:

$$\{\vdash_O \neg k \otimes g\}$$

As is well-known, the paradox is based to the following assumptions: (1) k is given as a fact; (2) g implies k. In SDL such premises permit to apply the inference rule RM thus obtaining a normative contradiction with the obligation $\neg k$. Since our formalism is not able to treat formulas with boolean operators, it seems impossible to represent the implication of k from g. Actually, we think this is not a real problem. It is enough to replace $k \vdash_O g$ with $k \vdash_O k$ in the implicit system. Thus, the explicit system will consist of the norm $\vdash_O \neg k \otimes k$. If so, the situation $s = \{k\}$ is *trivially* sub-ideal wrt the system (remember that $\neg k \otimes k$ is equivalent to $\top$). On the other hand, turning back to the original formulation of the paradox, if s is given, the system consisting of $\vdash_O \neg k \otimes g$ expresses consistently that g ought to be the case. In fact, the situation $s' = \{k, g\}$ is sub-ideal wrt the system.[3]

Reykjavik Scenario Consider now this version of Reykjavik Scenario:

$$\{\vdash_O \neg r,\ \ \vdash_O \neg g,\ r \vdash_O g,\ g \vdash_O r\}$$

Similarly to the previous examples, we can draw the following inferences:

$$\frac{\vdash_O \neg g \quad g \vdash_O r}{\vdash_O \neg g \otimes r} \qquad \frac{\vdash_O \neg r \quad r \vdash_O g}{\vdash_O \neg r \otimes g}$$

[3] And it is compatible with implication of k from g.

Accordingly, the explicit normative system is:

$$\{\vdash_O \neg g \otimes r,\ \ \vdash_O \neg r \otimes g\}$$

Given the situation $s = \{r\}$, the solution of the paradox consists of concluding that g ought to be case without deriving its negation. Actually, in our approach this is easily obtained since the situation $s' = \{r, g\}$ is sub-ideal wrt the explicit normative system.[4]

Möbius Strip Finally, let us look at Makinson's "Möbius Strip" example. Its logical structure is represented as follows:

$$\{c \vdash_O \neg b,\ a \vdash_O c,\ b \vdash_O a\}$$

This implicit normative system can be made explicit by drawing the following inference:

$$\frac{c \vdash_O \neg b \qquad b \vdash_O a}{\vdash_O \neg b \otimes a} \tag{8}$$

Similarly to the previous examples, the explicit normative system is as follows:

$$\{\vdash_O \neg b \otimes a,\ a \vdash_O c\}$$

Given the state of affairs $s = \{b\}$, it is expected to conclude that both a and c should be the case. Actually, this is what we get from the normative system since the situation $s' = \{b, a, c\}$ is sub-ideal wrt it. In fact, if b holds, this means that the primary obligation $\neg b$ is violated. Accordingly, the reparational obligation a ought to be the case. As a consequence, since a is to be given, the obligation c should follow as well.

7 Further Extensions and Final Remarks

Our analysis of the above paradoxes has shown that the sets of norms that characterize each of them are trivially consistent. However, even if such paradoxes correspond to relatively simple cases, our formalism is able to capture, at least potentially, more complicated normative structures. For this reason, we think the notion of normative consistency seems to deserve additional and more general remarks. A normative system is consistent if it does not allow for any application of rule ($\perp$). Roughly speaking, if A or $\neg A$ cannot be repaired by the system[5], then

$$\Gamma \vdash_O A \qquad\qquad \Delta \vdash_O \neg A$$

[4] Makinson [12] pointed out that the conclusion of g must be based on a prioritisation among promulgations. In a way, this remark applies also to our approach insofar as the norms of the explicit normative systems outweigh their counterparts in the implicit normative system. Remember that in our view a CTD is considered as an exception of a primary obligation.

[5] See Section 3 for the formulation of all the conditions of rule ($\perp$), and the end of Section 4 for a discussion.

should correspond to a normative contradiction. However, while this is quite clear when A is an atom, it is more difficult to understand intuitively the reasons why an inconsistency must occur when A is an arbitrary formula. As previously said, a close inspection of the inference rules shows that we can have negations of chains of reparational obligations. Thus, the question to be solved concerns the meaning of expressions like $\neg(A \otimes B)$. The problem is not so easy as the reader could expect. For this reason, in Section 5 we preferred to state the conditions to evaluate a situation with respect a normative system without considering this kind of formulas. Of course, in a certain perspective the question could be viewed as trivial: from a logical point of view, $\neg(A \otimes B)$ is nothing but the negation of a non-atomic formula. On the other hand, we think that the lack of a rule which defines the meaning of $\otimes$ with respect to the negation is in a way unsatisfactory.

Since $\otimes$ is not a boolean connective, it is impossible to establish a suitable definition of it in terms of any combination of formulas built by using $\otimes$ and $\neg$.[6] One of the possible lines of investigation comports to devise an additional consequence relation corresponding to the deontic notion of permission. In particular, such a consequence relation could be characterized at least by the following basic rules:

$$\frac{\Gamma \vdash_O A}{\Gamma \vdash_P A} \tag{9}$$

$$\frac{\Gamma \vdash_O A \otimes B}{\Gamma \vdash_P A} \qquad \frac{\Gamma \vdash_O A \otimes B}{\Gamma, \neg A \vdash_P B} \tag{10}$$

The first rule (9) is the version in our formalism of the notorious Ought-Can principle. The rules in 10 extend 9 to expressions containing the operator $\otimes$. It is easy to understand that, if a norm says that 'A is obligatory, otherwise B', this must imply that both A and B are permitted.

Moreover, thanks to the introduction of $\vdash_P$ it is possible to give a convincing account of formulas like $\neg(A \otimes B)$. Since the negation applies to $\otimes$, this means that B is not a reparational obligation of A and so its negation is permitted. In other words, we can have the following rule:

$$\frac{\Gamma \vdash_O \neg(A \otimes B)}{\Gamma \neg A \vdash_P \neg B} \tag{$\neg\otimes$}$$

Even though this seems to be a good intuition, some problems are far from being solved. Suppose to have a norm like this

$$\Gamma \vdash_O \neg(A \otimes B) \otimes C$$

This sequent is admitted in our formalism. However, its meaning is not clear and the application of a rule like ($\neg\otimes$) does not make sense in this case. If

[6] In [5] Brewka and colleagues argued that negation transforms their nested ordered disjunctions into standard disjunctions. In this way, the truth of $\neg(A \otimes B)$ makes A and B false. Unfortunately, we do not think this solution is adequate to account for our indended meaning of the $\otimes$ operator.

$\neg(A \otimes B)$ means that B is not a reparational obligation of A, then the question is: What does C stand for?

Problems like this, as well as the role of $\otimes$ on the left side of $\vdash$ or, more generally, the statement of some formal properties enjoyed by the our Gentzen system are a matter of future work. Here we can just advance some lines of further research:

- Identifying the conditions of existence of an explicit normative system, given its corresponding implicit counterpart; and the algorithm to compute the least fixed point of $\vdash_O$.
- Defining an appropriate semantics for the system in order to prove soundness and completeness. As a starting point, we expect it would be a possible world semantics based on selection functions.
- Extending the deductive power of the system by means of non-monotonic patterns such as cumulativity, restricted transitivity, etc.

To sum up we have presented a formal system for reasoning with CTD structures in an easy and natural way. We hope that this can be extended to other forms of normative reasoning.

Acknowledgments

We would like to thank Paola Benassi and Silvia Vida for their fruitful discussions on the topic and Alberto Artosi for his valuable comments and suggestions on an early draft of the paper. Thanks are also due to the anonymous referees for their remarks.

This paper has been written while the second author was visiting the Cooperative Information Systems Research Centre, Queensland University of Technology.

References

[1] C. E. Alchourrón. Philosophical foundations of deontic logic and the logic of defeasible conditionals. In J.-J. Meyer and R.J. Wieringa, editors, *Deontic Logic in Computer Science*, pages 43–84. Wiley and Sons, New York, 1993.

[2] C. E. Alchourrón and E. Bulygin. The expressive conception of norms. In R. Hilpinen, editor, *New Studies in Deontic Logic*, pages 95–124. Reidel, Dordrecht, 1981.

[3] A. R. Anderson. The formal analysis of normative systems. In N. Rescher, editor, *The Logic of Decision and Action*, pages 151–213. Pittsburg University Press, Pittsburg, 1967.

[4] M. Belzer. Legal reasoning in 3-D. In *Proceedings of the First International Conference in Artificial Intelligence and Law*, pages 155–163, Boston, 1987. ACM Press.

[5] G. Brewka, S. Benferhat, and D. Le Berre. Qualitative choice logic. In *Proc. Principles of Knowledge Representation and Reasoning. KR'02*. Morgan Kaufmman, 2002.

[6] R. M. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis*, 24:33–36, 1963.

[7] J. W. Forrester. Gentle murder or the adverbial samaritan. *Journal of Philosophy*, 81:193–197, 1984.

[8] D. M. Gabbay. *Labelled Deductive Systems. Volume 1.* Oxford University Press, Oxford, 1996.

[9] A.J.I. Jones and I. Pörn. Ideality, sub-ideality and deontic logic. *Synthese*, 65:275–290, 1985.

[10] A.J.I. Jones and I. Pörn. "Ought" and "Must". *Synthese*, 66:89–93, 1986.

[11] H. Kelsen. *General Theory of Norms*. Clarendon, Oxford, 1991.

[12] D. Makinson. On a fundamental problem of deontic logic. In P. McNamara and H. Prakken, editors, *Norms, Logics, and Information Systems*, pages 29–53. IOS Press, Amsterdam, 1999.

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

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

[15] H. Prakken and M. Sergot. Contrary-to-duty obligations. *Studia Logica*, 57:91–115, 1996.

[16] H. Prakken and M. Sergot. Dyadic deontic logic and contrary-to-duty obligations. In *Defeasible Deontic Logic*, pages 223–262. Kluwer, Dordrecht, 1997.

[17] D. Prawitz. Ideas and results in proof theory. In J.E. Fenstad, editor, *Proceedings of the Second Scandinavian Logic Symposium*, pages 235–307, Amsterdam, 1971. North Holland.

[18] G. Restall. *An Introduction to Substructural Logics*. Routledge, London, 2000.

[19] Erik Stenius. Principles of a logic of normative systems. *Acta Philosophica Fennica*, 16:247–260, 1963.

[20] L.W.N. van der Torre. The logic of reusable propositional output with fulfilment constraint. In David Basin, Marcello D'Agostino, Dov Gabbay, Sean Matthews, and Luca Viganó, editors, *Labelled Deduction*, volume 17 of *Applied Logic Series*, pages 245–266. Kluwer, Dordrecht, 2000.

[21] L.W.N. van der Torre and Y.-H. Tan. The many faces of defeasibility in defeasible deontic logic. In D. Nute, editor, *Defeasible Deontic Logic*, pages 79–119. Kluwer, Dordrecht, 1997.

[22] Silvia Vida. *Norma e condizione. Uno studio dell'implicazione normativa.* Giuffrè, Milano, 2001.