A dyadic operator for the gradation of desirability

Guillaume Piolle

INRIA Grenoble Rhône-Alpes,
Inovallée, 655 avenue de l'Europe
38334 Saint-Ismier Cedex - France
guillaume.piolle@inria.fr

Abstract. We propose a normal modal deontic logic based on a dyadic operator, similar in structure to the temporal "until". By bringing significant expressiveness to the logic, it allows both the definition of a monadic desirability operator similar to the SDL obligation, and the expression of the relative level of desirability of target formulae. The interpretation of this logic on a linear structure of worlds ordered by desirability makes its semantics more intuitive and concrete than the SDL deontic accessibility relation. We also show that the core modality of the logic permits to represent the Chisholm and Forrester paradoxes of deontic logic in a more precise way, which does not lead to inconsistencies.

1 Introduction

Limitations of Standard Deontic Logic (SDL) have constantly been pointed out, almost since its introduction following von Wright's seminal proposal [1]. However, no other unified mathematical formalization of this philosophical logic has alighted. Instead, many specialized logics have been proposed, each aimed at addressing one particular issue. One of these problems is that the usual Kripke semantics for SDL is rather abstract and unintuitive, being based on a binary deontic accessibility relation over possible worlds that would never be directly manipulated in any agent model. One of its consequences is that the obligation modality in SDL, as in many other deontic formalisms, is absolute and binary: all obliged formulae are considered on the same level, as are all possible or forbidden formulae. Many researchers have called for a richer notion of obligation, often based on source-based classification [2], conditional structures [3] or abstract contexts [4], and leading to a gradation of the notion of obligation.

What we propose here is to base this gradation over a deontic interpretation of linear temporal logic [5], which is already a formal structure embedding rich possibilities of organization between formulae, while keeping the formal framework relatively simple. The associated Kripke models will allow us to compare worlds, some being more ideal than others. In particular, we will interest ourselves in giving a deontic meaning to the structure of the dyadic "until" operator, which provides great expressiveness by formally linking formulae to each other. What we get is a deontic logic dealing with formulae that can be more or less

"obliged", unlike with the too simple binary SDL obligation. To acknowledge this wider semantics, we will rather speak of both formulae and worlds in terms of desirability, which covers obligation, inasmuch as an obliged formula is one that occurs in any desirable world, while providing room for gradation. This logic should be seen as the starting point of a new research track, and we examine here how it can be exploited. In addition to its conceptual interest, we will show that it provides very expressive tools to deal with contrary-to-duty norms, thus allowing to deal with the classical deontic paradoxes of Chisholm [6] and Forrester [7] in a way that does not lead to inconsistency.

In section 2, we present our logical framework. In section 3, we illustrate the differences with the SDL formalism, particularly based on an analysis of well-known paradoxes, showing that some of them can be nicely addressed in our logic. We compare our proposal to related works in section 4 and then conclude on possible improvements.

2 Structure of the logic

The logic we propose is a dyadic deontic logic based on a single primitive operator, similar in structure to the "until" operator of the temporal logic [8]. It is interpreted over linear semi-finite Kripke structures. Given the nature of our logic, it seems more straightforward and intuivite to start with a description of the semantic structures of the logic, before detailing the operators and their axiomatics. To begin, it is enough to know that we work with deontic operators designed to point out whether (and possibly to which extent) a given formula is desirable or not.

2.1 Semantic structures

The formulas of our logic are interpreted over rooted Kripke structures consisting in one root world (the current world, labelled w_0) and a countable set of possible worlds, ordered starting from w_1. The set of all possible worlds (possibly including other worlds than the ones described here) is noted $\mathcal{W}$. Fig. 1 represents a semantic structure (Kripke frame) for the logic. We will now detail its various components.

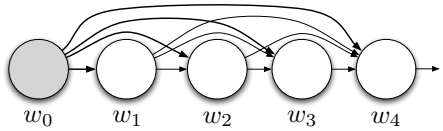

Fig. 1. A semantic structure of the logic, showing a sequence of worlds and the $<$ relation between them. The current world (w_0) and the relation instances it is involved in are highlighted.

w_0 is the current world, representing the actual state of the considered system, via the truth value of various formulae. Some of these formulae, exempt from deontic modalities, are related to facts, whereas some other, making use of the deontic operators we are about to introduce, tell us about what is desirable and what is not. In this current world, this expressed desirability or ideality may or may not be fully respected by the facts. If we choose to reason on desirability formulae in terms of obligations, then obligations expressed in w_0 can be violated in w_0.

The sequence of possible worlds $\{w_i\}_{i\in\mathbb{N}^*}$, on the other hand, refers to the desirable worlds. Each of them represents a possible state of the system that can be considered as acceptable, although some states are preferred to others. Namely, if $i < j (i, j > 0)$, then w_j represents a state more desirable than w_i.

The worlds of the structure are linked by a binary relation $\hookrightarrow \in \mathcal{W}^2$ which is serial (eq. 1) and linear (eq. 2-3), thus making the chain of linked worlds linear and infinite to the right (w_0 being the start of the chain on the left).

$$\forall w \in \mathcal{W}, \exists x' \in \mathcal{W}, w \hookrightarrow w' \tag{1}$$

$$\forall w, w', w'' \in \mathcal{W}, \text{ if } w \hookrightarrow w' \text{ and } w \hookrightarrow w'' \text{ then } w' = w'' \tag{2}$$

$$\forall w, w', w'' \in \mathcal{W}, \text{ if } w' \hookrightarrow w \text{ and } w'' \hookrightarrow w \text{ then } w' = w'' \tag{3}$$

We also introduce $<$ as the transitive closure of $\hookrightarrow$. Its reflexive version, $\leq$, is a total order relation. Therefore, $w_i < w_j (i, j \neq 0)$ means that w_j is strictly more desirable than w_i. Formally, this relationship ("less/more desirable") is defined by the binary relations we have introduced, and indexes are used to show the isomorphism between $(\mathcal{W}, <)$ and $(\mathbb{N}, <)$.

It is important to note, that w_0 is not identified as the least desirable world, it is considered separately from the ordered set. w_1 is the least desirable worlds *among the set of all desirable worlds*. Therefore, $w_0 \hookrightarrow w_1$ only identifies w_1 as the first desirable world of the sequence (provided w_0 is identified as the current world), not that w_1 is in some way more desirable than w_0, for we do not have any information about that. Indeed, w_0 may be little desirable, very desirable, or even not desirable at all (which means that there may or may not be a world $w_{i\neq 0}$ with the same valuations as w_0). In the same way, $w_0 < w$ only means that w is one desirable world in the set. To summarize, we have given a specific meaning to w_0 in the semantics, and therefore the relation instances in which it is involved bear a different meaning as well. Since the current world is identified in the structure, unlike in SDL, the interpretation of formulae will primarily take place in w_0. Interpreting the same formulae in the desirable worlds, although regulated by the same mechanisms, will bear slightly different meanings. Therefore the organization of the desirable worlds can be considered as a kind of "anchored" structure, the formulas in the root world being given a specific importance in applications. This idea is to be compared with the structures of anchored temporal logic, where the truth value of formulae in the initial world bear a distinct meaning. This is formally supported by the fact that w_0 is the only world with no predecessor, and by the strict version of the operators that we will choose, which will not include the current world in their

semantics. It is also worth noting that undesirable worlds are not represented. More precisely, they may be present but they are not reachable by the means of the relations we have introduced, so they will not have any influence on the interpretation of formulae: this logic focuses on the nuances of desirability but tells very little about undesirability.

The structure we propose is based on two strong hypotheses, that desirable worlds can be totally ordered and that there is a least desirable world among them. The first hypothesis (total order) is the most questionable one, since even though it may be easy for a human expert to point out the most desirable world among two, this might not aggregate in a total relation, or could eventually result in a preorder. Yet, we will keep this hypothesis as a working context, assuming that in most cases a total order can be used to approximate an expert's evaluation of desirability. On the other hand, the idea behind the least desirable world notion is that if it is possible to order desirable worlds, it is because all possible worlds (desirable and undesirable) can be ordered and that it is somehow possible to position an acceptability threshold somewhere on the scale. Therefore, the identity of the least desirable world and the beginning of the world sequence may vary, for a same system, according to the chosen level of expectation. It is even possible to include states that are only marginally desirable, or even slightly undesirable, if we can rely on an efficient gradation, allowing us to point out the most and least desirable formulae.

In a classical way, a model $\mathcal{M}$ is a triplet $(\mathcal{W}, \hookrightarrow, h)$ where h is a valuation function, such that a proposition p holds in a world w if and only if $w \in h(p)$.

2.2 The Δ dyadic modality

For the sake of clarity, we will begin by presenting the various operators, their meaning and the way they are interpreted over the semantic structures before detailing their axiomatics.

The sole primary operator of our logic (apart from the operators of propositional logic) is a dyadic deontic modality that we will note Δ. $\varphi\Delta\psi$, when evaluated in the current world w_0, means that there is a desirable world in which ψ is true, and that in all desirable worlds strictly less desirable than w, φ is true. Interpreted in another (desirable) world w, its meaning will be slightly different: it means that there is a world w' more desirable than w in which ψ holds, and that in all worlds strictly more desirable than w and strictly less than w', φ holds. Formally, the interpretation operator $\models$ is defined for this modality as per eq. 4 (we skip the definition of the interpretation operator on propositional structures, which remains very classical).

$$\mathcal{M}, w \models \varphi\Delta\psi \text{ if and only if } \exists w' \in \mathcal{W}, \begin{cases} w < w' \\ \mathcal{M}, w' \models \psi \\ \forall w'' \in \mathcal{W}, \text{ if } w < w'' \text{ and } w'' < w' \text{ then } \mathcal{M}, w' \models \varphi \end{cases} \tag{4}$$

This formal semantics covers both cases of interpretation, on the current worlds and on desirable worlds. One can check that this is the exact structure of the

strict "until" operator of the linear temporal logic. An example model is presented in fig. 2.

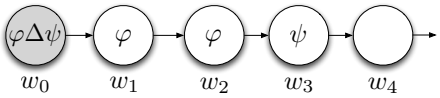

Fig. 2. A model illustrating the semantics of the Δ modality.

2.3 Reaching classical deontic operators

On the basis of the Δ modality, it is possible to build more easily manipulable operators. In each case, the meaning assigned to the operators will refer to an evaluation in the current world. First, we define a monadic operator *Acc* (for "acceptable"), comparable in some way to the SDL permission, as in eq. 5. It refers to a notion of acceptability: it means that there is at least one desirable world in which the formula is true (if the evaluation takes place in another world, it means that the formula occurs in at least one *more* desirable world). Therefore, the formula is considered as "acceptable", since it occurs in one desirable world. *Acc* is similar in structure to the existential operator F in temporal logic (in its strict version excluding the present).

$$Acc\ \varphi \stackrel{def}{=} \top \Delta \varphi \tag{5}$$

The formal semantics of the operator (eq. 6) results from its definition and from the semantics of Δ.

$$\mathcal{M}, w \models Acc\ \varphi \text{ if and only if } \exists w' \in \mathcal{M}, \begin{cases} w < w' \\ \mathcal{M}, w \models \varphi \end{cases} \tag{6}$$

One could note that this concept of acceptability is contextual: a formula is not acceptable is every case, and this presence in a desirable world may be subject to conditions. For instance, φ may be true in desirable worlds only when ψ is true as well. Including this kind of condition in the target formula of the *Acc* operator allows to capture this contextual characteristic. Again, this reasoning is similar to the one occurring with the SDL permission operator.

The dual operator of *Acc* is the universal modality *Des* (as defined in eq. 7), standing for "fully desirable" and roughly corresponding, in meaning, to a deontic obligation. In this universal version of desirability, we consider a formula fully desirable, in a kind of absolute, context-free way, if it is true in any desirable world (in any *more* desirable world, if the formula is evaluated in a wold other than the current one). This is expressed by eq. 8 obtained from the definition of

Des and the semantics of Δ, and the temporal equivalent is the strict universal modality G.

$$Des\ \varphi \stackrel{def}{=} \neg Acc\ \neg\varphi = \neg(\top \Delta \neg\varphi) \tag{7}$$

$$\mathcal{M}, w \models Des\ \varphi \text{ if and only if } \forall w' \in \mathcal{M},\ \text{if } w < w' \text{ then } \mathcal{M}, w \models \varphi \tag{8}$$

From this, it is possible, if needed, to build a *Udes* operator for undesirability (eq. 9), capturing the idea that a formula is true in no desirable world (eq. 10) and linked in meaning to the SDL interdiction modality.

$$Udes\ \varphi \stackrel{def}{=} Des\ \neg\varphi = \neg(\top \Delta \varphi) \tag{9}$$

$$\mathcal{M}, w \models Udes\ \varphi \text{ if and only if } \forall w' \in \mathcal{M},\ \text{if } w < w' \text{ then } \mathcal{M}, w \models \neg\varphi \tag{10}$$

2.4 Axiomatics

We have seen that the modality Δ is similar in structure to the temporal "until". Consequently, its axiomatics (eq. 11-19) is very close as well, only adapted to the limitation of the chain to the left.

$$Des\ (\varphi \rightarrow \psi) \rightarrow ((\rho \Delta \varphi) \rightarrow (\rho \Delta \psi)) \tag{11}$$

$$Des\ (\varphi \rightarrow \psi) \rightarrow ((\varphi \Delta \rho) \rightarrow (\psi \Delta \rho)) \tag{12}$$

$$((\varphi \Delta \psi) \wedge \neg(\rho \Delta \psi)) \rightarrow (\varphi \Delta (\varphi \wedge \neg\rho)) \tag{13}$$

$$(\varphi \Delta \psi) \rightarrow ((\varphi \wedge (\varphi \Delta \psi)) \Delta \psi) \tag{14}$$

$$(\varphi \Delta (\varphi \wedge (\varphi \Delta \psi))) \rightarrow (\varphi \Delta \psi) \tag{15}$$

$$(\varphi \Delta \psi) \wedge (\rho \Delta \sigma) \rightarrow \begin{pmatrix} ((\varphi \wedge \rho) \Delta (\psi \wedge \sigma)) \\ \vee((\varphi \wedge \rho) \Delta (\psi \wedge \rho)) \\ \vee((\varphi \wedge \rho) \Delta (\varphi \wedge \sigma)) \end{pmatrix} \tag{16}$$

$$Acc\ \top \rightarrow (\bot \Delta \top) \tag{17}$$

$$Acc\ \varphi \rightarrow ((\neg\varphi) \Delta \varphi) \tag{18}$$

$$Acc\ \top \tag{19}$$

Axioms 11 and 12 are the equivalent of the K axiom for monadic modalities, providing a kind of distributivity of the operator over logical implication. Axioms 13 and 14 build the link between the two arguments of the operator, giving its specificity to the construct, and ensure the maintaining of the left hand side formula over the corresponding range. Axiom 15 provides transitivity, while the following axioms regulate more closely the organization of the worlds in the structure: axiom 16 ensures linearity, axiom 17 ensures discreteness, axiom 18 provides a sound ordering and axiom 19 makes the chain of worlds infinite to the right (*i.e.* in the direction of higher desirability).

From this axiomatics and the definition of the *Des* and *Acc* abbreviations, it results that *Des* has the properties of a $KD4.3$ modality. The KD part of it is obviously reasonable for a universal deontic modality.

$$Des\ \varphi \rightarrow Des\ Des\ \varphi \tag{20}$$

$$(Acc\ \varphi \wedge Acc\ \psi) \rightarrow Acc\ (\varphi \wedge Acc\ \psi) \vee Acc\ (\varphi \wedge \psi) \vee Acc\ (Acc\ \varphi \wedge \psi) \tag{21}$$

The 4 axiom (eq. 20), corresponding to the transitivity of $<$, has already been discussed as an interesting property for obligations, for instance by Brian Chellas [9], while deemed undesirable in other cases, depending on the precise sense given to the obligation modality. In our case, it gives an additional meaning to the succession and ordering of the desirable worlds. Seen from the current world w_0, it means that every formula that is currently fully desirable is also fully desirable in all desirable world, which is of little consequence if the desirability of formulae is to be interpreted only in the current world. Seen from any desirable world, it means that any formula fully desirable in this world is also fully desirable in all the more desirable worlds. In other words, it brings a kind of monotony in the concept of full desirability. Axiom .3 (eq. 21) takes meaning only in this context. It tunes the monotony introduced by 4 to the linear structure of the semantics. It means that if two formulae are acceptable in the current world, then either it is acceptable that they occur simultaneously, or it is acceptable that one of them occurs and that the second is still acceptable. The last point is directly linked to the notion of monotony we have introduced.

2.5 Building more expressive deontic operators

So far, these operators do not bring much more than their SDL counterparts, except in terms of an axiomatics which is specific to our vision of the system. But thanks to the expressiveness of Δ, it is possible to build other deontic operators, either unreachable with SDL modalities or bearing a meaning that could not be supported by SDL's semantics. First, we introduce *Ult* as a unary operator for "ultimate desirability" (eq. 22). A formula is ultimately desirable when it is true in a connected subset of most desirable worlds, that is all worlds more desirable than a given reference (eq. 23). An example model is given in fig. 3.

$$Ult\ \psi \stackrel{def}{=} \top \Delta(\psi \wedge Des\ \psi) \tag{22}$$

$$\mathcal{M}, w \models Ult\ \psi \text{ if and only if}$$
$$\exists w' \in \mathcal{W}, \begin{cases} w < w' \\ \forall w'' \in \mathcal{W}, \text{ if } w' \leq w'' \text{ then } \mathcal{M}, w \models \psi \end{cases} \tag{23}$$

Obviously, the set of fully desirable formulae is included in the set of all ulti-

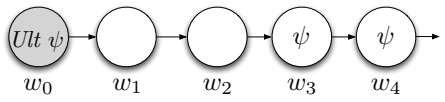

Fig. 3. A model illustrating the semantics of the *Ult* operator.

mately desirable formulae. This operator allows us to identify the formulae that are likely to guide us towards the most desirable worlds. The abstract notion of these "most desirable" worlds put us in the need of means to compare those

ultimately desirable formulae. This can be done by considering that among ultimately desirable formulae, a formula φ is more widely desirable than a formula ψ if and only if the corresponding set of worlds begins sooner in the scale of desirability, that is if one can derive $\mathit{Ult}\ (\varphi \wedge \mathit{Ult}\ \psi) \wedge \neg \mathit{Ult}\ (\psi \wedge \mathit{Ult}\ \varphi)$.

However, if an ultimately desirable formula is something good to achieve, it may not be easily reachable There could be alternative ways leading to worlds that are less desirable, but acceptable nonetheless. We will note $\varphi \lhd \psi$ when ψ is an ultimately desirable formula, and φ is such an alternative to it. This operator is defined by eq. 24 and its deduced semantics detailed in eq. 25.

$$\varphi \lhd \psi \stackrel{def}{=} \varphi \Delta (\psi \wedge \mathit{Des}\ \psi) \tag{24}$$

$$\mathcal{M}, w \models \varphi \lhd \psi \text{ if and only if}$$
$$\exists w' \in \mathcal{W}, \begin{cases} w < w' \\ \forall w'' \in \mathcal{W},\ \text{if } w < w'' \text{ and } w'' < w' \text{ then } \mathcal{M}, w \models \varphi \\ \forall w'' \in \mathcal{W},\ \text{if } w' \leq w'' \text{ then } \mathcal{M}, w \models \psi \end{cases} \tag{25}$$

Basically, $\varphi \lhd \psi$ means that φ is true in the first desirable worlds, and that it

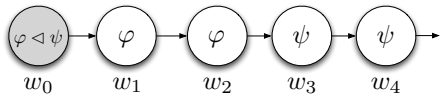

Fig. 4. A model illustrating the semantics of the $\lhd$ operator.

is true until ψ becomes true and remains so in all most desirable worlds. This operator cannot be obtain from the SDL KD obligation, in the same way that the temporal "until" cannot be expressed by monadic temporal modalities.

3 Comparison with SDL

We understand our logic as a possible alternative to Standard Deontic Logic [1], at least in some cases. It is therefore necessary to shed light on a few points of comparison between the two formalisms.

On a very formal aspect, for instance, our logic and SDL share the same complexity class. Indeed, since our logic relies on the same axiomatics as the until-based propositional linear temporal logic, it inherits its PSPACE-completeness [10], and it is known that SDL, as a KD system, is also PSPACE-complete [11].

3.1 Expressiveness

From the point of view of expressiveness, we have seen that SDL-like operators (full desirability comparing to obligation and acceptability to permission) can be used. Although they have additional axioms in our version, we have seen

that these are justified in the context of our linear semantics. The meaning of the *Ult* operator is obviously unreachable from SDL, since it relies on the specific structure of graded world desirability. It illustrates the new opportunities brought by this logic in terms of comparison between several target formulae. We have begun to explore this pathway by suggesting the notion of "more widely desirable formula".

So far, the most interesting contribution of this logic in terms of expressiveness seems to be the notion of alternative formula ($\lhd$ operator). Indeed, not only does it bring to light a clear relationship between two formulae (one being possibly more accessible and the other more desirable), but it allows reasoning agents to build elaborate strategies on this basis. Actually we think that $\lhd$ would be a nice operator for contrary-to-duty (CTD) obligations. $\varphi \lhd \psi$ can read "ψ is what you should aim for, but in case it is not possible, then to reach an acceptable situation φ should hold". In this formula, ψ is only ultimately desirable ($Ult\ \psi$), and not fully desirable ($\neg Des\ \psi$). However, it is the case that $Des\ (\varphi \vee \psi)$. This translates fairly well the meaning of a contrary-to-duty, imperfectly stated in SDL as $\{Ob\ \psi, \neg\psi \rightarrow Ob\ \varphi\}$. It is part of the meaning of the problem that ψ is obliged, but that there are ways of resolution in the case this obligation is violated. It is to be noted, however, that φ is not necessarily an alternative *goal* for an agent, since it can be a logical consequence of not ensuring ψ. It should be seen as a necessary characteristic of any acceptable world (provided the primary target formula is deemed unreachable). On the other hand, $\neg\psi \wedge \neg\varphi$ is something that should be avoided at all cost. The use of the dyadic operator $\lhd$ captures this hierarchy between the formulae, and the order among worlds allows to characterize the degree of severity of an eventual failure to comply with all norms (a world with $\neg\psi \wedge \varphi$ is more desirable than a world with $\neg\psi \wedge \neg\varphi$), while in SDL Kripke semantics, both kinds of failures (even the one complying with the CTD obligation) are captured by the binary notion of deontic inaccessibility. Therefore, we claim that our logic has a CTD tool more adapted than the standard SDL translations of the concept.

3.2 Structural paradoxes of deontic logic

We propose now to explore how the well-known paradoxes of Standard Deontic Logic apply in the logic we have designed.

Both formalisms being normal modal logics based on propositional logic, the paradoxes linked to its intrinsic elements still hold. This is the case of Ross's paradox [12], linked to the interpretation of disjunction ($Des\ \varphi \rightarrow Des\ (\varphi \vee \psi)$ is still valid in our logic), of the paradox of derived obligation [13], based on the nature of logical implication ($\vdash Des\ \neg\varphi \rightarrow Des\ (\varphi \rightarrow \psi)$), of the paradox of the good Samaritan [14], direct consequence of the necessitation rule common to all normal modal logics (if $\vdash (\varphi \rightarrow \psi)$ then $\vdash Des\ \varphi \rightarrow Des\ \psi$). Similarly, the free choice paradox [12] ($\nvdash Acc\ (\varphi \vee \psi) \rightarrow (Acc\ \varphi \vee Acc\ \psi)$) also holds. Basically, *Des* has the same rules and axioms as SDL's obligation, therefore the structural paradoxes necessarily apply.

Sartre and Plato dilemmas [15], which can be described as SDL's inability to express conflicting obligations without leading to the logical inconsistency of the system, can be considered from a new perspective in the light of the notion of ultimately desirable formula. Sartre dilemma is a conflict between two strict obligations of high moral importance, therefore it seems reasonable, if one's expectations in terms of desirability are high enough, to consider it as a problem in our logic as well, SDL's $Ob\ \varphi \wedge Ob\neg\varphi$ corresponding to $Des\ \varphi \wedge Des\ \neg\varphi$, both leading to inconsistency. Plato dilemma, on the other hand, is a conflict between two obligations of different importance, one having the priority over the other. One example instance of the dilemma (taken from McNamara [16]) would be :

- I am obligated to meet you for a light lunch meeting at the restaurant ($Ob\ \varphi$);
- I'm obligated to rush my choking child to the hospital ($Ob\ \psi$, with $\{\varphi, \psi\} \vdash \bot$).

In SDL, the notion of precedence or priority among norms is impossible to express and we end up with the same kind of inconsistency as for Sartre dilemma. In our new formalism, however, we can manipulate the gradation of desirability. A global vision of the problem makes it obvious that in the most desirable worlds, φ is false and ψ is true, if ψ is the only way to save the child. Therefore, ψ should be (at least) ultimately desirable ($Ult\ \psi$), and φ is not fully desirable ($\neg Des\ \varphi$). The acceptability of φ (and the full desirability of ψ) then depends of the limit that should be set on the desirability of the worlds. Indeed, if a world where the child is dead but the lunch meeting has been attended is considered marginally desirable (and not totally unthinkable), then we have $Acc\ \varphi$, $Ult\ \psi$, $\neg Des\ \varphi$ and $\neg Des\ \psi$. On the other hand, if we consider that a world when the child is dead is necessarily not desirable, then we have $Udes\ \varphi$ and $Des\ \psi$. However, this analysis of the situation based on our logic is made a posteriori, after an implicit evaluation of the priorities. It is not an expression of how the norms are individually expressed independently from each other. Therefore, it is not decently possible to conclude that our logic is strong enough to resolve conflicts by itself.

3.3 SDL paradoxes based on contrary-to-duties

The paradoxes of deontic logic based on CTD obligations seem to be the most difficult ones, making the formal description of possible situations logically inconsistent. The first of these paradoxes is the Chisholm paradox [6], which we describe here in parallel with its SDL translation (taken from McNamara [16]):

It ought to be that Jones go to the assistance of his neighbours.	$Ob\ \varphi$
It ought to be that if Jones does go then he tells them he is coming.	$Ob(\varphi \rightarrow \psi)$
If Jones doesn't go, then he ought not tell them he is coming.	$\neg\varphi \rightarrow Ob\neg\psi$
Jones doesn't go.	$\neg\varphi$

This set of formulae leads to an inconsistency ($Ob(\psi \wedge \neg\psi)$). Of course, this situation could be modelled in a similar way in our logic by simply replacing Ob by Des, and then it would lead to the same conflict and the paradox would hold. However, it is possible to interpret the text of the paradox in terms of graded desirability. In this case, the first and third obligations obviously don't have the same strength, as the third one is conditioned by the violation of the first one. Therefore, it seems that a world where Jones does not go to his neighbours and he does not tell them that he comes (one violation) is still somewhat acceptable, at least more than a world where he does not go in spite of his word, which sounds undesirable. For this reason it seems reasonable to merge these two norms in a single one, more complex than what SDL can achieve, and making use of our $\lhd$ operator to model the CTD:

$$(\neg\psi) \lhd \varphi \tag{26}$$

$$Des\ (\varphi \rightarrow \psi) \tag{27}$$

$$\neg\varphi \tag{28}$$

Eq. 26 says that it is ultimately desirable that Jones goes to his neighbours, and also that not telling them that he is going is an acceptable but less desirable alternative to it. In other words, the second part of the sentence means that if Jones does not go, then not telling is the only way for the situation to be acceptable, which expresses the intention of the third obligation of the initial SDL formalization. Eq. 27 is the expression of the absolute obligation (uninterpreted in terms of gradation) of the second obligation, and eq. 28 is unchanged. It is easy to show that this formalization is consistent, a model of it exposing $\neg\psi$ in a first sequence of worlds and then φ in all more desirable worlds (the formulae 26-28 being true at w_0). It is true, however, that the construction of eq. 26 somehow blurs the independence between the initial first and third obligation, but it does so in a manner that keeps their meaning to all the initial sentences while respecting their organization, to the difference of SDL-based attempts involving formulae like $Ob(\neg\varphi \rightarrow \neg\psi)$ (which is almost void in meaning since it can be deduced from $Ob\ \varphi$). One can also note that the current formalization of the Chisholm scenario does not lead to the "pragmatic oddity" (Jones being required both to help and not to tell) that can be found in some other proposals [17] and was pointed out by Prakken and Sergot [18]. This is achieved by not deriving $Des\ \neg\psi$ from $\neg\varphi$, but instead presenting $\neg\psi$ as a fallback alternative for a failed φ via the $\lhd$ operator. This way, $\neg\varphi \wedge \neg\psi$ holds in no desirable world (which would anyway be contradictory with eq. 27).

The Forrester paradox, or paradox of the gentle murderer [7], is another puzzle based on CTD formulae. It is presented as follows in SDL (quoted from McNamara [16]):

It is obligatory that John Doe does not kill his mother.	$Ob\ \neg\varphi$
If Doe does kill his mother, then it is obligatory that Doe kills her gently.	$\varphi \rightarrow Ob\ \psi$
Doe does kill his mother.	φ

This must be read along with the theorem saying that if Doe kills his mother gently, then he kills his mother ($\vdash \psi \rightarrow \varphi$). This time the set of sentences is not inconsistent, but it counter-intuitively results in John Doe being obliged to kill his mother. Here again we propose an interpretation of the first and second sentences based on our $\lhd$ operator:

$$\psi \lhd \neg\varphi \tag{29}$$

$$\varphi \tag{30}$$

Here eq. 29 says that Doe not killing his mother is ultimately desirable, and that an (extreme) alternative to it would be to kill her gently. In other words, there still are a few marginally desirable worlds in which John Doe kills her mother, but then he does it gently. Inclusion of these worlds at the beginning of the chain of all desirable worlds is necessary if one wants to make a distinction between the violation of the first obligation only, and the violation of both obligations. This formalization, while capturing the essence of the situation, does not lead to φ being fully desirable or even ultimately desirable, it remains only acceptable.

To conclude, it would certainly be exaggerated to say that this proposal solves CTD-related paradoxes, since they still hold in their SDL-equivalent formalization and since the $\lhd$-based formalization takes them out from their limited formal context, but it certainly provides another point of view on how to deal with CTD-related issues.

4 Related works

This works mostly relies on a new semantic interpretation of modal deontic logic. Other people have explored such ways. This is notably the case of Sven Ove Hansson [19], who based a non-Kripke semantics on preference relations among worlds (an idea very close to ours), leading to preferences among actions, from which obligations are constructed. By refusing to base his system on the necessitation rule, Hansson avoids a number of deontic paradoxes, but being outside of Sahlqvist's theorem conditions [20], the correspondence between the axioms and properties of the obligation operator and the characteristics of the preference relation lead to other counter-intuitive situations.

Addressing the Plato and Sartre dilemma is the core activity of the community working on normative conflict resolution. We have admitted the limit of our current proposal in the domain. Efficient ways to deal with conflicting obligations seem to rely on the directed obligations introduced by Ryu [2], consisting in linking each norm to the agent having enacted it. Identifying the source of the obligation allows to break contradiction relationships, allowing conflicts to be spotted and then arbitrated. This is often done by the means of norm fusion mechanisms [21, 22], allowing certain norms to be deactivated on the basis of preference relations, thus providing maximal consistent sets of norms.

The idea of using dyadic modalities to express deontic notions is not new either. However, existing dyadic operators (like the system developed by Bengt

Hansson [3] and David Lewis [23]) focus on the proper formalization of conditional or contextual obligations, rather than on a gradation of alternative formulae. In more recent works, though, the need for the gradation of desirability or ideality is clearly exposed and exploited, for instance by Prakken and Sergot [4]. These notions are closely related to contrary-to-duty obligations, often addressed by the means of conditional obligations or preferences, like in the case of Cholvy and Garion [24]. In their work, based on Boutilier's CO^* formalism, the preference relation is constructed among propositions (and formulae), not among possible worlds. Besides, their model is enriched with the notion of fixed, controllable and "influenceable" formulae, which interacts with the preference relation to express subtler notions of agency.

To the best of our knowledge though, it is the first time that a graded deontic interpretation is proposed to the traditional structures of the "until" operator of linear temporal logic (which is very different from adopting a mainly temporal definition of obligations, as proposed in some logical frameworks [25]). It seems to us that this point of view, as a nice side effect, appears to be an interesting way of considering the CTD issue. There remains, however, to compare more closely our logical framework to the many other approaches to contrary-to-duties. This class of problems has attracted much attention since the 60's and we have not been able yet to make formal comparisons between our logic and all existing approaches. The CTD-related paradoxes are probably the main reason why normal modal logic is now often considered an unsuited basis for deontic logic. Many research tracks investigate deontic modelling using weaker forms of logics or alternative inference systems. For instance, Governatori and Rotolo decide, on the basis of their analysis of specific CTD issues, to avoid relying on classical modal logic for deontic concepts [26]. They represent contrary-to-duty norms in a dedicated Gentzen system, with a specific CTD operator $\otimes$ associated to inference rules capturing the essence of the preference between various formulae without giving rise to the usual SDL paradoxes. We also feel that particular attention should be given to defeasible deontic logic [27, 28], which sees CTD norms as exceptions. Although this last philosophical position is subject to debate and often criticized [24], it seems that it is rather close, yet not identical, to the perspective proposed here. Defeasibility has indeed been integrated in some CTD formalisms, for instance by Governatori in a RuleML version of his previous formalization of contrary-to-duties. [29].

5 Conclusion

We have proposed, in this paper, a deontic interpretation of an anchored version of the "until" linear temporal logic, leading to a kind of dyadic deontic logic which allows reasoning on the gradation of desirability. This formal basis has allowed to devise SDL-like operators for desirability and acceptability, as well as specific modalities able to capture more subtle notions. The key interest of this logic seems to be the proposal of a new approach on contrary-to-duty norms,

providing a method to bypass some of the CTD-related inconsistencies found in Standard Deontic Logic.

5.1 Limitations and perspectives

The presented work is the beginning of a research track. As such, it has identified limitations that should be addressed in future developments and it opens perspectives for improvements.

It seems to us that the main limitation of this logical framework is the hypothesis of linearity and total order that we have accepted for the semantics of gradually desirable worlds. These two properties are linked, in the sense that the building of an tree structure (*i.e.* branching to the right) instead of the linear one would allow to alleviate both. This is only one example of how it is possible to improve the semantics, but having several possible paths of desirability (and being able to quantify on them) should allow us to manipulate worlds that are not comparable in terms of direct desirability, because they would belong to different deontic alternatives. This perspective might also allow us to drop the strict monotony of desirability (eq. 20). This needs serious evaluation, but it might be possible to construct a deontic interpretation upon a formalism similar to Computation Tree Logic, which already provides the tree structure and the modalities quantifying over it.

This leads us to the introduction of time in our logic of graded desirability. Although the formal inspiration of this logic is LTL, temporal concepts have not been discussed at all so far, and yet they are omnipresent in terms of analogies. More generally, time is considered a vital notion in deontic framework, because obligations are almost always deeply linked to deadlines or delays [30]. It should be examined how the arrow of time can be mixed, in a clean way, with the gradation of desirability in a possible world semantics compatible with our Δ operator. Another, related point worth investigation is the "anchored" nature of our semantics, where w_0 has a specific meanings among all worlds, by representing the current world where evaluation take place. Currently, if the situation (the facts, not the norms) in the current world evolves, then the whole model should be switched for another one, where the chain of desirable worlds is identical but where w_0 alone has been updated. This calls for a new kind of structure providing both the chain of desirable worlds and a graph of possible current worlds, linked by another accessibility relation. This new relation could be dynamic (based on actions) or temporal in nature, thus providing the expressiveness needed to address the absence of time in the current version of the framework.

We would also like to explore the many possible operators that could arise from the Δ operator. Namely, many other comparison operators (among which some could be the base of order relations) could be devised based on the successive apparition of formulae in the chain of desirable worlds. Also, the interest of a weak version of our CTD operator ($(\varphi \vee \psi) \lhd \psi$ instead of $\varphi \lhd \psi$) should be examined.

It should be made clear that these proposed pathways are only suggestions, calling for exploration and evaluation.

5.2 Applications

We have only discussed the syntactic and semantic aspects of the logic so far, but we are convinced that the concepts borne by the proposed operators could be useful in many application areas. For instance, it has already been suggested that CTD could be a nice tool for designing security policies, because they would allow to distinguish between different grades of obligations and link sanctions to them [31]. We think that our formalism would allow to provide more expressiveness to the notion of violation and sanction, perhaps allowing to introduce and manipulate the sibling concept of reparation, by formally linking an ultimately desirable formula with one or several graded alternatives. Tools of this kind can help provide more expressive security policies, more closely representing, for instance, the complex workflows and regulations on data usage in organizations.

However, one could question the discrete nature of the preference structure. Indeed, the possible alternatives of a human agent, and a posteriori of a human organization, often represent a continuum. Therefore, in order to deal with realistic philosophical scenarios inspired by human experience and common life situation, dense structures would certainly be more suitable. The formal and computational impact of a similar dyadic desirability operator over a dense structure remains to be examined, but the existing works on until-based logics or other dyadic temporal constructions over dense time can provide a basis for such studies [32]. The discrete structure could however be fairly usable as presented here for computing applications, which usually present countable or even finite collections of states. This is why we think this logical framework is more adapted to dealing with computer-directed security policies and policy-compliant (or moderately compliant) software components, than to modelling social behaviour.

Acknowledgments. We would like to thank Romuald Thion for the initial interrogation having led to this work and for his useful remarks. We also thank the DEON reviewers for their constructive and very valuable insights. This research has been funded by the ANR 07 SESUR FLUOR project.

References

1. von Wright, G.H.: Deontic logic. Mind **60** (1951) 1–15
2. Ryu, Y.U.: Relativized deontic modalities for contractual obligations in formal business communication. In: Proceedings of the 30th Hawaii International Conference on System Sciences (HICCS'97), Washington, DC, USA, IEEE Computer Society (1997) 485
3. Hansson, B.: An analysis of some deontic logics. Nôus **3** (1969) 373–398
4. Prakken, H., Sergot, M.J.: Dyadic deontic logics and contrary-to-duty obligations. Synthese Library **263** (1997) 223–262

5. Pnueli, A.: The temporal logic of programs. In: Proceedings of the 18th IEEE Symposium on the Foundations of Computer Science (FOCS-77), Providence, Rhode Island, USA, IEEE Computer Society Press (1977) 46–57
6. Chisholm, R.M.: Contrary-to-duties imperatives and deontic logic. Analysis **24** (1963) 33–36
7. Forrester, J.W.: Gentle murderer, or the adverbial samaritan. Journal of Philosophy **81** (1984) 193–196
8. Kamp, J.A.W.: Tense Logic and the Theory of Linear Order. PhD thesis, University of California at Los Angeles (1968)
9. Chellas, B.F.: Modal Logic, an Introduction. Cambridge University Press (1980)
10. Reynolds, M.: The complexity of the temporal logic with until over general linear time. Journal of Computer and System Sciences **66**(2) (March 2003) 393–426
11. Ladner, R.: The computational complexity of provability in systems of modal propositional logic. SIAM Journal of Computing **6** (1977) 467–480
12. Ross, A.: Imperatives and logic. Theoria (1941) 53–71
13. Prior, A.N.: The paradoxes of derived obligation. Mind **63** (1954) 64–65
14. Prior, A.N.: Escapism. In: Essays in Moral Philosophy. University of Washington Press, Seattle, USA (1958) 135–146
15. Lemmon, E.J.: Moral dilemmas. Philosophical Review **71** (1962) 139–158
16. McNamara, P.: Deontic logic. In Zalta, E.N., ed.: The Stanford Encyclopedia of Philosophy. Stanford University (2006)
17. Jones, A.J.I., Pörn, I.: Ideality, sub-ideality and deontic logic. Synthese **65** (1985) 275–290
18. Prakken, H., Sergot, M.J.: Contrary-to-duty obligations. Studia Logica **57**(1) (1996) 91–115
19. Hansson, S.O.: Semantics for more plausible deontic logics. In: DEON'02, London, UK (May 2002)
20. Sahlqvist, H.: Correspondence and completeness in the first- and second-order semantics for modal logic. In: Proceedings of the Third Scandinavian Logic Symposium, North-Holland, Amsterdam (1975)
21. Cholvy, L., Cuppens, F.: Analyzing consistency of security policies. In: Proceedings of the 18th IEEE Symposium on Research in Security and Privacy, Oakland, CA, USA (May 1997)
22. Cholvy, L., Cuppens, F.: Reasoning about norms provided by conflicting regulations. In Prakken, H., McNamara, P., eds.: Norms, Logics and Information Systems: New Studies in Deontic Logic and Computer Science, Amsterdam, the Netherlands, IOS Press (1998) 247–264
23. Lewis, D.: Semantic analyses for dyadic deontic logic. Logical theory and semantic analysis (1974) 1–14
24. Cholvy, L., Garion, C.: Utilisation d'une logique de préférences conditionnelles pour raisonner avec des normes contrary-to-duties. In: Journées Nationales sur les Modèles de Raisonnement, Arras, France (May 2001)
25. Dignum, F., Broersen, J., Dignum, V., Meyer, J.J.: Meeting the deadline: Why, when and how. In Hinchey, M.G., Rash, J.L., Truszkowski, W., Rouff, C., eds.: Third International Workshop on Formal Approaches to Agent-Based Systems (FAABS'04). Number 3228 in LNCS, Greenbelt, MD, USA, Springer Verlag (april 2004) 30–40
26. Governatori, G., Rotolo, A.: A gentzen system for reasoning with contrary-to-duty obligations. a preliminary study. In Jones, A.J.I., Horty, J., eds.: Sixth international workshop on deontic logic in computer science (DEON'02), London, UK, Imperial College (5 2002) 97–116

27. Nute, D.: Apparent obligation. In: Defeasible Deontic Logic: Essays in Nonmonotonic Normative Reasoning. Kluwer Academic Publishers, Dordrecht, The Netherlands (1997) 287–316
28. van der Torre, L.W.N.: Violated obligations in a defeasible deontic logic. In Cohn, A., ed.: Proceedings of the 11th European Conference on Artificial Intelligence, John Wiley and Sons (1994) 371–375
29. Governatori, G.: Representing business contracts in ruleml. International Journal of Cooperative Information Systems **14**(2-3) (2005) 181–216
30. Demolombe, R.: Formalisation de l'obligation de faire avec délais. In: Troisièmes journées francophones des modèles formels de l'interaction (MFI'05), Caen, France (may 2005)
31. Brunel, J., Cuppens, F., Cuppens-Boulahia, N., Sans, T., Bodeveix, J.P.: Security policy compliance with violation management. In: ACM workshop on Formal methods in security engineering (FMSE'07), Fairfax, Virginia, USA, ACM Press (2007) 31–40
32. Dignum, F., Kuiper, R.: Specifying deadlines with dense time using deontic and temporal logic. International Journal of Electronic Commerce **3**(2) (1999) 67–86