An Integrated Framework for Ought-to-be and Ought-to-do Constraints

P. d'Altan, J.-J.Ch. Meyer and R.J. Wieringa

UU-CS-1995-30
September 1995

Utrecht University

Department of Computer Science

Padualaan 14, P.O. Box 80.089,
3508 TB Utrecht, The Netherlands,
Tel. : ... + 31 - 30 - 531454

An Integrated Framework for Ought-to-be and Ought-to-do Constraints

P. d'Altan, J.-J.Ch. Meyer and R.J. Wieringa

Technical Report UU-CS-1995-30
September 1995

Department of Computer Science
Utrecht University
P.O.Box 80.089
3508 TB Utrecht
The Netherlands

ISSN: 0924-3275

Contents

An integrated framework for ought-to-be and ought-to-do constraints[1]

P. d'Altan[2]
J.– J.Ch. Meyer[3]
R.J. Wieringa[4]

1 Introduction

Deontic logic is the logic to reason about ideal and actual behaviour. Besides the traditional role as an underlying logic for law and ethics (for a survey see [MW93]), deontic logic has been proposed as a logic for the specification of legal expert systems [BMT87],[Sta80], authorization mechanisms [ML85], decision support systems [KL88], [Lee88a],[Lee88b], database security rules [GMP89], fault-tolerant software [KM87],[Coe93], and database integrity constraints [WMW89], [WWMD91]. A survey of applications can be found in [WM93b]. In all these areas, we must be able to reason about the difference between ideal and actual behaviour. In many cases, it is important to distinguish *ought-to-do* statements (which may be interpreted as expressing imperatives of the form "an actor ought to perform an action") from *ought-to-be* statements (which express a desired state of affairs without necessarily mentioning actors and actions bearing relations with that state of affairs). There are situations where we would like to relate the two oughts with each other. For example, suppose we want to specify deontic integrity constraints for a bank data base. From the ought-to-be constraint

(1.) The balance of a bank account must be non-negative

we would like to derive the ought-to-do statement

[1]Partial support is acknowledged from the Esprit Basic Research Action 8319 ModelAge.

[2]Current address: Universitá di Milano, Dipartimento di Filosofia, via F.del Perdono 7, 20122 Milano, Italy. Email: daltan@imiucca.csi.unimi.it.

[3]Current address: Department of Mathematics and Computer Science, Utrecht University, P.O. Box 80089, 3508 TB Utrecht, The Netherlands. email: jj@cs.ruu.nl

[4]Department of Mathematics and Computer Science, Vrije Universiteit, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands. Email: roelw@cs.vu.nl.

(2.) If the balance of a bank account is n and $n - m < 0$, then it is forbidden to withdraw m from the account.

In addition, we would like to be able to express

(3.) If the balance of a bank account is n and $n < 0$, then an action *deposit*(m) with $n + m \geq 0$ ought to be performed.

As several philosophers have argued, the distinction between ought-to-be and ought-to-do is not only a matter of syntax. We follow Castañeda [Cas70, page 452] in separating deontic statements into those that involve agents and actions and support imperatives (ought-to-do) and those that involve states of affairs and are agentless and have by themselves nothing to do with imperatives. After introducing operators for ought-to-be and ought-to-do we first follow up the suggestion of Geach [Gea81] and try to reduce the ought-to-be to the ought-to-do, using a formalization of ought-to-do that we gave earlier [Mey88, WMW89, WWMD91]. Perhaps not so surprisingly, we will see that some rather plausible attempts at giving such a reduction do not yield satisfactory results. We then try to circumvent the problems encountered with this reduction by using another reduction, viz. by reducing both ought-to-do and ought-to-be to alethic modalities and then considering the relations between the so reduced formulas as to ought-to-be and ought-to-do. We will work out this possibility using Meyer's reduction of ought-to-do to dynamic logic and using Anderson's reduction of ought-to-be to alethic modal logic. It will turn out that the integrated framework that we obtain in this way, provides a sound — albeit rather minimal — basis for giving specifications involving both ought-to-do's and ought-to-be's as well as reasoning with these.

The structure of the paper is as follows. In section 2, we state and criticize a few plausible assumptions reducing ought-to-be to ought-to-do. These relations will be expressed in the language of first order dynamic logic (PD_eL) with deontic operators, but without any formal semantic characterization of the expressions involved. The aim of this section is that of establishing a few intuitive constraints on the results we will later obtain by logic.

Section 3.1 concerns a different formalization of deontic sentences in terms of what in the literature is known as Anderson's reduction. We briefly review Anderson's reduction of ought-to-be statements to alethic ones and successively Meyer's reduction of ought-to-do statements to dynamic ones (section 3.2). Both reductions reduce deontic modalities to alethic ones.

In section 4, we consider again the attempted reductions of ought-to-be to ought-to-do of section 2 and analyze the failure of these attempts in the light of the formalization in the combined Anderson-Meyer reduction. This will increase our understanding of the sometimes complex relations between these two modalities. Finally, in section 5 we will show how by making use of both reductions it is possible to solve the expressivity problems sketched in the first part of this introduction. Section 6 concludes the paper witb a number of open problems.

2 Four attempted reductions of ought-to-be to ought-to-do

In the first section of the paper we have reviewed a few positions concerning the relations between ought-to-be and ought-to-do. They span from Geach's conviction that ought-to-be does not exist independently from ought-to-do to the position of the phenomenological school, according to which object of our acts of will are states of affairs and not actions. Between these two extremes we find philosophers who maintain that we can only understand ought-to-do once we have explained what ought-to-be is and philosophers who think that ought-to-do and ought-to-be are indeed separated and no connection exists between them [Gar86].

What seems clear is that in no way these thinkers have conceived ought-to-be as completely reducible to ought-to-do or vice versa; they all maintain either the existence of relations (implicative ones?) or the lack of any such relation. In what follows we are going to consider to what extent it is possible within the formalism adopted to express valid relations between the two concepts.

First of all, we will try to check whether the two concepts may be somehow equalled by reducing ought-to-be to ought-to-do. As explained before, the reason why we choose this direction of reduction is purely pragmatical: we already have a logic for ought-to-do [Mey88], [WMW89]. We distinguish between the two concepts by using traditional operators for ought-to-be (i.e., O, P, F) and the same operators but with a hat for ought-to-do (i.e., $\hat{O}, \hat{P}, \hat{F}$). In particular, we will propose a few definitions expressing important semantic relations between definiens and definitum. Viewed syntactically, these should be considered as *equivalential definitions* in the sense of [Hum93].

In order to correctly understand our analysis of the semantic relations

here presented, note that we deal with formulas of classical logic, where the implication connective lacks any feature of relevance. That is, when we state that φ implies ψ, all we are saying is that φ is a sufficient condition for ψ.

2.1 First attempt: there is an obligatory action that leads to φ

Definition 1 *A state is obligatory iff it is the result of an obligatory action*

$$O\varphi \stackrel{\text{def}}{=} \text{there is an } \alpha \text{ such that } [\alpha]\varphi \wedge \hat{O}\alpha$$

The formula $[\alpha]\varphi$ is to be read as "after any possible way of performing action α, φ will hold". The definition says that state φ is obligatory if and only if there is an obligatory action α that always leads to φ. We refer to the two directions of the equivalence as $1^{\rightarrow}$ and $1^{\leftarrow}$. There are arguments against the intuitive validity of both directions.

Counterexample 1.1 Implication $1^{\rightarrow}$ asserts that a state-of-affairs cannot be considered obligatory independently from the existence of acts for bringing it about. This is simply not true, since we have often prescriptions that do not prescribe what actions should be performed in order to fulfil the prescriptions themselves. In addition, it would allow us to derive factual consequences (the existence of an action to achieve a result) from deontic antecedents (the obligatoriness of the result). This is not only philosophically suspect, it is also empirically falsified. Consider a situation in which a robot cannot perform any action (except for idling perhaps) due to power failure. Then it cannot establish any obligatory / desirable state-of-affairs.

This counterexample exploits the fact that there may be obligatory states for which there is no action to reach them. This may or may not be the case in social systems, but it can certainly be the case in artificial systems like the robot world of the example or the bank account world of the introduction.

Note that part of $1^{\rightarrow}$ can be viewed as a *rationality constraint* on a law-giver. We do not want a law-giver to issue a law that makes a state obligatory, without there being at least one action that leads to that state. It may be that abstractly, there are desirable states for which there is no action that leads to them. For example a world without war may be desirable, but we see no humanly possible action that would lead to such a world. However, the norm that a world without war is desirable is not issued by a law-giver, so we find the practical unreachabilty of this ideal state of affairs acceptable.

However, anyone who explicitly issues a law declaring a state of affairs ideal, should make sure that there is at least one action that leads to this state of affairs.

Counterexample 1.2 $1^{\leftarrow}$ says that if there is an obligatory action that leads to φ, then φ is obligatory. As a counterexample to this, consider the obligation to jog (α) because this is good for your health. Jogging makes you very tired (φ), so that the right-hand side of definition 1 is satisfied. We consider this a good thing in the states reachable by jogging (it means that you did good practice to maintain your health), but nevertheless it is not in all states of the world a good thing that you are tired.

We may view $1^{\leftarrow}$ as a *constructivity assumption* made by law-givers. A law-giver has the choice to declare a state of affairs φ to be ideal or to declare an action α leading to φ to be obligatory. In the first option, a kind of declarative, "implementation independent" law is issued that says that a state is desirable without explicitly putting any obligation on any action leading to that state. We have seen above that this option runs the danger that there is no action at all that leads to φ, putting the actors subject to the law in a hard predicament.

In the second option, a kind of constructive law is issued, in which an action is made obligatory. This option assumes that $1^{\leftarrow}$ is valid, i.e. that if it is obligatory that α leads to φ, then φ is obligatory. In this case, the motivation of making α obligatory is utilitarian, because α is made obligatory because of its results. One danger of this option is that α is performed mindlessly, without regard for its consequences, as a ceremony without contents. In more formal terms, α may have undesirable consequences in addition to the desirable consequence φ. The jogging example may be interpreted this way. The problem of undesirable derived consequences is not solved by any of the approaches presented in this paper and we take it up again in the discussion at the end of this paper.

2.2 Second attempt: all actions that lead to p are obligatory

We have observed that there may be desirable states for which there is no obligatory action α that leads to it, and that there are also obligatory actions that may lead to a state that is not always obligatory. Perhaps we can improve on Def.1 by requiring a closer connection between actions and their results before we let obligatoriness of the former lead to obligatoriness

of the latter. Def.2 does this by saying that φ is obligatory iff *all* actions that lead to it are obligatory.

Definition 2 *A state φ is obligatory iff all the actions that lead to the state of affairs φ are obligatory*

$$O\varphi \stackrel{\text{def}}{=} \text{for all } \alpha \text{ we have that } [\alpha]\varphi \rightarrow \hat{O}\alpha$$

This avoids the assumption that there always is an obligatory action that leads to a desirable state, and it also avoids the counterexample to $1^{\leftarrow}$, since we now require that *all* actions that lead to φ are obligatory before we regard φ to be desirable. However, it introduces other problems.

Counterexample 2.1 As a counterexample to $2^{\rightarrow}$, suppose that it is obligatory that the balance of a bank account is greater or equal to 0 (i.e., $O(\mathit{balance} \geq 0)$). Yet, not all actions that lead to a positive balance are obligatory. If the balance is already positive, then any deposit leads to a positive balance, but in this situation, deposits are not obligatory.

This counterexample illustrates a difference between ought-to-be and ought-to-do: The ought-to-be statement can be valid in all possible states of the world, but the corresponding ought-to-do statement applies only in a state where the ought-to-be is violated. Actions that are obligatory because they produce a desirable state of affairs, are only obligatory when that state of affairs does not hold. In our approach we can make these conditions explicit by specifying e.g. a conditional ought-to-do of the form $\neg\varphi \rightarrow \hat{O}\alpha$ rather than just $\hat{O}\alpha$.

Counterexample 2.2 Suppose we have a state φ such that no action leads to φ. Then, vacuously, all actions that lead to φ are obligatory, so the right-hand side of definition 2 is satisfied. Yet, to conclude from this that φ is ideal is absurd.

2.3 Third attempt: It is forbidden to undo state φ

The previous attempts express some positive connection between an ought-to-be and a corresponding ought-to-do. Perhaps we should try to find a negative connection, that stays closer to the classical view that something is obligated if it is forbidden to undo it. The following definition of ought-to-be is an attempt in this direction.

Definition 3 *A state is obligatory iff it is forbidden to undo it*

$$O\varphi \stackrel{\text{def}}{=} \text{for all } \alpha \text{ we have that } [\alpha]\neg\varphi \rightarrow \hat{F}\alpha$$

In other words, a state φ is obligatory iff all actions that lead to $\neg\varphi$ are forbidden. This corresponds to the classical definition $O\varphi \equiv F\neg\varphi$. We have found no counterexample to $3^{\rightarrow}$: if a state-of-affairs φ is obligatory, then every action that results in making φ untrue should — intuitively speaking — indeed be forbidden. On the other hand, we can again find a counterexample to the other direction.

Counterexample 3.1 The direction $3^{\leftarrow}$ is counterintuitive in the case of states φ for which there are no actions that lead to $\neg\varphi$. This is a variation on the counterexample against $2^{\leftarrow}$.

2.4 Fourth attempt: all actions necessary and sufficient to reach state φ are obligatory

In all examples so far, we have seen that the connection between φ and the action(s) that lead to it is not always "tight" enough, so that obligatoriness of the one could not be justifiably be inherited by the other. As a remedy we might require that the acceptable actions are those and only those which represent necessary conditions for obtaining φ. We formalize the concept of a necessary condition as follows:

An action α is said to be a necessary condition ($\alpha \rightsquigarrow \varphi$) for the obtaining of the state-of-affairs φ if the following equivalence holds

$$\alpha \rightsquigarrow \varphi \qquad \text{iff} \qquad [\alpha]\varphi \wedge [\bar{\alpha}]\neg\varphi^{5}$$

Here the notation $\overline{\alpha}$ means the non-performance of α. An exact definition of the non-performance of an action can be found in [DM90]. Thus, $\alpha \rightsquigarrow \varphi$ means α necessarily leads to φ and not doing α necessarily leads to $\neg\varphi$. There is no other action than α that brings about φ. We use this concept to update Def.2 as follows:

[5] In the present remark we limit us to necessary conditions for the obtaining of state-of-affairs, but we could as well consider necessary conditions for actions. In that case, the above definition has to be adapted:

$$\beta \rightsquigarrow \alpha \quad \text{iff} \quad [\beta](\langle\alpha\rangle\top) \wedge [\bar{\beta}]([\alpha]\,\bot)$$

i.e., action β is a necessary condition for performing action α iff α can be succesfully performed exactly when β has been performed.

Definition 4 *A state φ is obligatory iff all actions required to bring it about are obligatory as well*

$$O\varphi \stackrel{\text{def}}{=} \text{ for all } \alpha(\alpha \rightsquigarrow \varphi) \rightarrow \hat{O}\alpha$$

This move would rule out all cases not really necessary for obtaining a given end. As before, the left-to-right implication seems intuitively valid to us. However, there are troubles with $4^{\leftarrow}$ just as there are troubles with $3^{\leftarrow}$.

Counterexample 4.1 Exactly as Def. 2, this definition cannot cope with situations where φ cannot be reached by an action, so we can again use counterexample 2.2.

2.5 Discussion

Looking at the sequence of proposed reductions of ought-to-be to ought-to-do, we see that the implications from left-to-right improve but that the implications from right-to-left all are subject to counterexamples. The last two left-to-right implications have no counterexamples at all, and in the formal semantics given later, we will be able to prove them. Concentrating now on the right-to-left implications, it is interesting to consider them in a different sequence from the one they were presented above. Read in this direction, the natural starting point is definition 3, because this is the direct analogon to the classical SDL definition of ideal states (abbreviating "for all" to $\forall$):

$$O\varphi \stackrel{\text{def}}{=} \forall\alpha([\alpha]\neg\varphi \rightarrow \hat{F}\alpha).$$

This is counterintuitive for the cases where there are no actions at all that lead to $\neg\varphi$. Definition 2 is a variation of definition 3, with $\neg\varphi$ replaced by φ and $\hat{F}(\alpha)$ by $\hat{O}(\alpha)$:

$$O\varphi \stackrel{\text{def}}{=} \forall\alpha([\alpha]\varphi \rightarrow \hat{O}\alpha).$$

Here, $2^{\leftarrow}$ is counterintuitive when there are no actions that lead to φ. By combining the two definitions, we get definition 4:

$$O\varphi \stackrel{\text{def}}{=} \forall\alpha((\alpha \rightsquigarrow \varphi) \rightarrow \hat{O}\alpha).$$

The counterexample to $4^{\leftarrow}$ again exploits the fact that there may be unreachable states φ. This situation is excluded by definition 1:

$$O\varphi \stackrel{\text{def}}{=} \exists\alpha([\alpha]\varphi \wedge \hat{O}\alpha).$$

However, $1^{\leftarrow}$ is invalidated by the fact that, even if the results of obligatory actions are desirable, they may not be desirable in all possible states of the world. In addition, as we will see in the discussion at the end of the paper, there is the problem that not all derived results of an obligatory action may be obligatory.

We have analyzed the definitions according to our intuitive comprehension of the notions involved there, i.e., the notion of obligatory or forbidden action and desirable state of affairs, but we have not gone so far to the point of abstracting from the logical context in which those notions had to be related to each other, i.e., propositional logic. The relations between ought-to-be and ought-to-do are after all expressed by means of the standard propositional connectives and of first-order quantifiers, and we all know that, for instance, material implication does not express any common sense notion of implication, at all. Undoubtedly, such problems originate in our reluctance to abandon classical logic. The point is however that in non-classical logics we would not be able to investigate whether there exist *truth functional relations* between ought-to-be and ought-to-do, and this is in a certain sense what we are looking for here. That is why we are going to elaborate a *classical* (from the point of view of logic) framework where we may reason about both ought-to-do and ought-to-be without reducing the one to the other so that, in effect, we will obtain a framework in which these two forms of ought are *integrated*.

In the next section, we present Anderson's reduction of ought-to-be and Meyer's reduction of ough-to-do. We then integrate the two reductions in one logic, give a semantics and a number of axioms, and discuss the status of some well-known paradoxes in the combined logic. After that, we return to the attempted reductions discussed above and show that each of the above counterexamples corresponds to a formal counterexample in our semantics. In addition, the two implications for which no counterexample was found, are proven valid in the combined logic.

3 Integrated framework for ought-to-do and ought-to-be

3.1 Anderson reduction of ought-to-be sentences to alethic modal ones

In this section we (re)consider Anderson's approach to deontic logic [AM57], based on a reduction to alethic modal logic. Anderson takes a modal logic of type KT (in the terminology of Chellas [Che80]) with modality $\Box$, read as "necessarily". (In the following we also use the usual dual operator $\Diamond$ ("possibly"), defined as $\neg\Box\neg$.) A special atom V is employed, interpreted by Anderson rather freely as expressing some form of 'sanction' or 'something bad'. Anderson then defines the (ought-to-be) deontic operators of obligation (O), prohibition (F) and permission (P) by the following reductions:

Definition 5

$$O\varphi \equiv \Box(\neg\varphi \rightarrow V)$$

$$F\varphi \equiv \Box(\varphi \rightarrow V)$$

$$P\varphi \equiv \neg F\varphi\,(\equiv \Diamond\varphi \wedge \neg V)$$

Thus, Anderson reduction of $O\varphi$ has to be read as "necessarily, $\neg\varphi$ implies a sanction (that things go wrong)". Furthermore, Anderson assumes that V is subject to three conditions:

1) $\Diamond\neg V$ is assumed (or proved) valid

2) $\neg V$ cannot be proved valid

3) V cannot be proved valid

For our purposes, these conditions are less important. For instance, in a very restricted context, say a robot world, it may well be that there is no "good" world, so that in this context $\Box V$ $(= \neg\Diamond\neg V)$ is true. For Anderson trying to give an account that is sound for general and universal deontic contexts (in ethics, for example), this would be unacceptable.

Viewed model-theoretically, models for the deontic system are just a special kind of Kripke models for modal sentences. These are usually defined as follows.

Definition 6 *A Kripke model M for a modal logic L is given by* $\mathcal{M} = \langle W, \pi, R, \models\rangle$
where

(1) *W a set of possible worlds;*

(2) *$\pi : \Pi \times W \rightarrow \{\mathbf{1}, \mathbf{0}\}$ is a function that assigns a truth value to propositional variables (Π is a set of propositional variables) in a possible world;*

(3) *$R : W \rightarrow \wp(W)$ a function that associates to world w, the set of possible worlds accessible from w;*

(4) *$\models = \{(w, \varphi) | w \in W$ and $\varphi \in \Pi\}$ the usual truth relation between worlds and sentences.*

Definition 7 *Let $\mathcal{M} = \langle W, \pi, R, \models \rangle$ be a Kripke model.*

- *φ is* true *in a world $w \in W$, denoted $\mathcal{M}, w \models \Box\varphi$, if for all $w' \in W$ such that wRw', we have $\mathcal{M}, w' \models \varphi$. (Other clauses, pertaining to the propositional connectives, are as usual.)*

- *A formula φ is* valid *in a model $\mathcal{M} = (W, \pi, R, \models)$, denoted $M \models \varphi$, if $M, w \models \varphi$ for all $w \in W$.*

- *A formula φ is* valid, *denoted $\models \varphi$, if $M \models \varphi$ for all Kripke models M.*

It can be shown that in this way, one obtains a deontic logic, which we call ADL, extending the standard deontic logic SDL (the modal logic KD in Chellas' classification [Che80]) regarding the modality O. That is, ADL contains all validities of SDL and more. Since ADL extends SDL, it inherits the well-known paradoxes of SDL. However, when one reads the atom V as *violation of a norm*, most of these paradoxes disappear. So, for instance, $O\varphi$ reads "necessarily, $\neg\varphi$ implies violation". A world where V is true might be viewed as a "bad" or "non-ideal" world, and w world in which V is false as a "desirable" or "ideal" world. Hence $O\varphi$ can also be interpreted as "necessarily, non-violation implies φ" or equivalently, "necessarily, ideal worlds satisfy φ". This is very close to the standard interpretation of $O\varphi$ in SDL.

However, as was shown in [McA81], ADL contains validities beyond those of SDL, that are counter-intuitive at first sight, such as the following:

- $O\varphi \rightarrow OO\varphi$ (If φ is obligatory, it ought to be obligatory)

- $\Box(\Box\varphi \rightarrow O\varphi)$ (necessarily, if φ is necessaruly true then it is obligatory)

- $\Box(O\varphi \rightarrow \Box O\varphi)$ (Necessarily, if φ is obligatory, then it is necessarily obligatory)

We shall return to these shortly. Here it suffices to say that these, too, are not so counterintuitive in the reading we have in mind. This is the reason why in spite of the criticism against Anderson's reduction in the literature, we will nevertheless incorporate it in our system for representing ought-to-be constraints.

However, in order to be able to give a proper formalization of ought-to-do constraints as well, we will deviate from Anderson's formulation at least in one important aspect. We have seen that he interprets V either in terms of sanction or in terms of an unspecified bad thing. We propose a more specific reading in terms of "violation", or better, we use V as a signal that a norm has been violated. We read $O\varphi$ as "necessarily, $\neg\varphi$ implies violation of the normative system to which $O\varphi$ belongs", where with normative system we simply intend a set of deontic constraints, thus not necessarily (or preferably not) moral norms. There are two important reasons that justify our reading of V:

1. It avoids the counterintuitive reading of a few derivations of ADL. These readings are counterintuitive when V is interpretated as sanction (cfr. [Cas60]). Consider, for example, FV, i.e. $\Box(V \rightarrow V)$. This means in Anderson's system "sanctions are forbidden" and this is evidently counterintuitive. In our own interpretation, it means "violations are forbidden", i.e. "violations are violations" and this is not counterintuitive. The formula $P\varphi \equiv \Diamond(\varphi \wedge \neg V)$ also is counterintuitive, as illustrated by the following example, taken from [Cas60], 46:

   > Let V, i.e. the sanction mentioned in the Penal Code, be 'you will be put in jail for 10 years'; and p be 'you will be put in jail for 9 years'. Clearly, it is logically possible to put you in jail for 9, but not 10 years. Thus, it follows logically that it is permitted to put you in jail for 9 years — without ado!

   Since, now, V means the signalling of *some* violation within a normative system, we cannot read it as 'you will put in jail for 10 years'. Furthermore, φ cannot retain the interpretation Castañeda gives to it. We have two cases: either (a) φ has to do with *some* violation of a norm or (b) it has not. In both cases we have that the conjunction $\varphi \wedge \neg V$ is intuitively false. For (a) this is immediate, since we say

that a violation is and is not at the same time the case. For (b), we may say that putting you in jail for not having violated anything is in violation of the human rights (obviously, we are here assuming that all normative systems somehow share the same basic rights, in this case, the right not to be imprisoned without conviction) — and therefore $\varphi \wedge \neg V$ is again contradictory.

2. Our reading offers an almost tautological and therefore scarcely objectionable reading of deontic formulas: if I do not fulfil a norm, then I violate it, or if φ is not the case, then the relevant constraint has been violated.

3.2 Meyer's reduction of ought-to-do sentences to dynamic ones

As argued in [McA81], Anderson's reduction to alethic modal logic is not quite satisfactory, at least not for the representation of ought-to-do constraints. After analyzing the reason why this is so, Meyer [Mey88] proposed another reduction, in this case to propositional dynamic logic. A consequence of the use of dynamic logic is the distinction between propositions (assertions) and actions (practitions, cf. [Cas81]). Meyer's reduction uses Anderson's violation atom V to indicate that an action has occurred that violates one of the deontic constraints.

(Propositional) Dynamic Logic (PDL, cf. [Har84]) consists of a two sorted propositional language (we have a set Π of propositional variables and a set $\underline{A}$ of actions), extended with modal operator $[\alpha]$ for every action $\alpha \in \underline{A}$. We call A the *alphabet of actions* and keep it fixed throughout the paper.

A formula $[\alpha]\phi$ is read as "the performance (execution) of the action denoted by α leads necessarily to a state (possible world) in which ϕ holds". In this approach, α is forbidden ($\hat{F}\alpha$), permitted ($\hat{P}\alpha$), and obligated ($\hat{O}\alpha$) are reduced to dynamic formulas as follows:

Definition 8

$$\hat{F}\alpha \equiv [\alpha]V$$

$$\hat{P}\alpha \equiv \neg\hat{F}\alpha\,(\equiv \langle\alpha\rangle\neg V)$$

$$\hat{O}\alpha \equiv \hat{F}\overline{\alpha}\,(\equiv [\overline{\alpha}]V)$$

Here for the reduction of the obligation operator $\hat{O}$, we employed the *negation* of an action term α, denoted $\overline{\alpha}$. The concept of action negation is discussed in [Mey89], [DM90] and [WM93a]; here, it will suffice to consider $\overline{\alpha}$ as a term denoting *any* choice of actions *not* involving the action denoted by α. This can be formalized as a kind of complementation operator [DM90, Mey88]. Since in this paper we do not need to bother about the structure of action terms beyond negation ($\bar{\cdot}$) and choice ($+$), we only give a concise treatment of terms $\overline{\alpha}$ and $\alpha_1 + \alpha_2$. As pointed out above, we assume our language has a fixed alphabet $\underline{\mathcal{A}}$ of atomic action constants, which are names of the basic actions that are considered relevant in a specific context. We keep the alphabet of action names fixed throughout the paper. In addition, we select a set $\mathcal{A}$ of elementary actions and associate with each atomic action name $\underline{a} \in \underline{\mathcal{A}}$ an elementary action $a \in \mathcal{A}$. We call $\mathcal{A}$ the universe of actions and keep it fixed throughout the paper as well.

In order to give an adequate semantics of negated action terms, we interpret action terms in an *open* sense, as explained below. The interpretation $[\![\alpha]\!]$ of an action term α will be given in two steps:

1. First, we interpret the action term as a so-called step of elementary actions that it involves. We will refer to this as the *step semantics* of action terms.

2. Next, we interpret an action term by specifying its effect on the state of the world. We call this the *state-transition* semantics of action terms.

The step semantics $[\![\alpha]\!]_S$ of an action term is a set of so-called *steps*. A step is a non-empty finite set of elementary actions, denoted as $[a_1, \ldots, a_n]$. Each step is a set of actions that occur simultaneously in a state transition of the world. A choice between steps is represented by a set of steps, where each step in the set represents one possible option. The step semantics of an atomic action name $\underline{a} \in \underline{\mathcal{A}}$ is now defined as the set of all steps that contain the action $a \in \mathcal{A}$ associated with the action name $\underline{a} \in \underline{\mathcal{A}}$:

$$[\![\underline{a}]\!]_S = \{S \subseteq \mathcal{A} \mid a \in S\}.$$

In other words, if the action denoted by the action term $\underline{a}$ occurs, then this means that *any* step is taken in which this action occurs. The step semantics of $\underline{a}$ can thus be paraphrased as "a occurs and any finite number of other actions may occur simultaneously". This agrees with the usual intention when we say that an action occurs: by saying that we do not intend that no other action occurs, but we leave open what other actions currently occur.

The step semantics of a choice $\alpha_1 + \alpha_2$ is simply the union of the step semantics of α_1 and α_2:

$$[\![\alpha_1 + \alpha_2]\!]_S = [\![\alpha_1]\!]_S \cup [\![\alpha_2]\!]_S.$$

This means that $\alpha_1 + \alpha_2$ occurs if and only if α_1 or α_2 occurs.

Finally, the step semantics of a negated action term $\overline{\alpha}$ is obtained as the set-theoretic complement of the set of steps denoted by α, where the complement is taken with respect to the set of all steps:

$$[\![\overline{\alpha}]\!]_s = STEPS \setminus [\![\alpha]\!],$$

where $STEPS$ is the set of all nonempty finite subsets of $\mathcal{A}$. This means that the negated action term $\overline{\alpha}$ denotes all those steps in which the action denoted by α does not occur. This concludes our informal exposition of the step semantics of action terms.

Turning to the state-transition semantics of action terms, we now define the effect of each action on the possible worlds of a Kripke structure. With each action $a \in \mathcal{A}$ we associate a function $eff(a) : W \to W$, that describes the effect state-transforming effect of a. For convenience, we may consider eff as a function

$$\mathcal{A} \to (W \to W).$$

So $eff(a)(w) = w'$ says that the event a occurring in world w results in a world w'. Now we lift the function eff to steps as follows. For a step $S = [a_1, \ldots, a_n]$ we define

$$eff(S) = eff(a_1) \circ \cdots \circ eff(a_n),$$

where $\circ$ denotes function composition. Again, $eff(S)$ is a function $W \to W$, describing the state-transforming effects of the step S. In order for this definition to be meaningful, we need to impose the notion of compatible steps. A step $S = [a_1, \ldots, a_n]$ is *compatible* if

$$eff(a_{i_1}) \circ \cdots \circ eff(a_{i_n}) = eff(a_1) \circ \cdots \circ eff(a_n)$$

for every permutation $(i_1, \ldots, i_n)$ of $(1, \ldots, n)$. This simply means that the actions in the step may occur in any arbitrary order without changing the result. For non-compatible steps we simply leave $eff(S)$ undefined.

We lift eff further to sets $T$of steps as follows:

$$eff(T)(w) = \{eff(S)(w) \mid S \in T \text{ compatible } \}.$$

So $eff(T)$ is a function of type $W \to \wp(W)$, where $\wp$ stands for powerset. Note that if $T = \{S\}$ where S is not compatible, then $eff(T)(w) = \emptyset$.

Finally, we define the state-transition semantics of an action term α. For any $w \in W$,

$$[\![\alpha]\!](w) = eff([\![\alpha]\!]_S)(w).$$

Thus, $[\![\alpha]\!]$ is a function of the type $W \to \wp(W)$, describing the state-transforming effect of the action denoted by α, where there might be multiple outcomes collected in a subset of W.

Alternatively and equivalently, we may define the effect of α by means of an accessibility relation R_α, given by

$$R_\alpha(w, w') \Leftrightarrow w' \in [\![\alpha]\!](w).$$

This is more in line with the way semantics is defined modal/dynamic logic.

3.3 Integration

The system that we will adopt integrates both Anderson's and Meyer's reductions in a modal logic, where, for convenience, we take an S5-type necessity operator $\Box$ as a basis. This will simplify our models for the logic below, although by taking S5 rather than KT as our basis, extra validities are obtained, which in the traditional interpretation of deontic logic are generally considered problematic. We will return to this in the sequel.

Thus, the system we are going to assume as basic is characterized as a mixed modal-dynamic logic with the following axioms

Axioms 3.1

$$\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi) \quad \text{(K)}$$

$$\Box\varphi \to \varphi \quad \text{(T)}$$

$$\neg\Box\varphi \to \Box\neg\Box\varphi \quad (5)$$

$$\Diamond\neg V \quad \text{(D)}$$

$$[\alpha](\varphi \to \psi) \to ([\alpha]\varphi \to [\alpha]\psi) \quad \text{(AK)}$$

$$\varphi,\ \varphi \to \psi / \psi \quad \text{(MP)}$$

$$\varphi / \Box\varphi \quad \text{(N)}$$

$$\varphi / [\alpha]\varphi \quad \text{(AN)}$$

$$F\varphi \leftrightarrow \Box(\varphi \rightarrow V) \qquad (\mathrm{F})$$

$$O\varphi \leftrightarrow \Box(\neg\varphi \rightarrow V) \qquad (\mathrm{O})$$

$$P\varphi \leftrightarrow \neg F\varphi \qquad (\mathrm{P})$$

$$\hat{F}\alpha \leftrightarrow [\alpha]V \qquad (\hat{\mathrm{F}})$$

$$\hat{O}\alpha \leftrightarrow [\overline{\alpha}]V \qquad (\hat{\mathrm{O}})$$

$$\hat{P}\alpha \leftrightarrow \neg\hat{F}\alpha \qquad (\hat{\mathrm{P}})$$

$$\Box\varphi \rightarrow [\alpha]\varphi \qquad (\Box[\alpha])$$

$$\neg[\alpha]\bot \wedge [\alpha]\Box\varphi \rightarrow \Box\varphi \qquad ([\alpha]\Box)$$

It is important for the counterexamples given later that $\mathcal{A}$ is a subset of the universe of actions A.

The rules are, as usual, necessitation for both operators (N and AN, respectively) and modus ponens (MP). We call the system containing the above axioms and rules $PDeL^{AM}$, Propositional deontic Logic with Anderson's and Meyer's reductions.

3.4 Semantics

Definition 9 *A Kripke model M for PDL is given by* $\mathcal{M} = \langle \mathcal{A}, W, \pi, [\![\alpha]\!], \models, \mathrm{opt} \rangle$
where

(1) $\mathcal{A} \subseteq A$ *is a set of basic actions such that for each action* $\alpha \in \mathcal{A}$ *there is also its negation* $\overline{\alpha} \in \mathcal{A}$*;*

(2) W *a set of possible worlds;*

(3) $\pi : \Pi \times W \rightarrow \{\mathbf{1}, \mathbf{0}\}$ *is a function that assigns a truth value to propositional variables in a possible world;*

(4) $[\![a]\!] : \mathcal{A} \times W \rightarrow \wp(W)$ *a function that associates to action* $a \in \mathcal{A}$ *and world* w*, the set of possible worlds to which the performance of a leads;*

(5) $\models = \{(w, \varphi) | w \in W \text{ and } \varphi \in \Pi\}$ *the usual truth relation between worlds and sentences;*

(6) $\mathrm{opt} \subseteq W$ *represents the set of "best" elements of* W*, such that* $w \in \mathrm{opt}$ *iff* $w \models \neg V$[6]*. The set opt is assumed to be non-empty.*

[6] See also [Åqv88], 107–113, for further details.

Note that $\mathcal{A}$ contains action *literals*, i.e. for each action α also its negation $\overline{\alpha}$. Of course, there should be a relation between $[\![\alpha]\!]$ and $[\![\overline{\alpha}]\!]$, that depends upon the formalization of action negation. For example, we may want to require that $[\![\overline{\alpha}]\!] = [\![\alpha]\!]$. Action negation is studied in detail elsewhere [DM90, Mey88, WWMD91].

Definition 10 *Let* $\mathcal{M} = \langle \mathcal{A}, W, \pi, [\![\alpha]\!], \models \rangle$ *be a Kripke model.*

$\mathcal{M}, w \models \varphi$ *is inductively defined as follows.*

$$
\begin{array}{lcl}
\mathcal{M}, w \models p & \Leftrightarrow & \pi(p, w) = \mathbf{1} \\
\mathcal{M}, w \models \neg\varphi & \Leftrightarrow & \text{not } (\mathcal{M}, w \models \varphi) \\
\mathcal{M}, w \models \varphi \wedge \psi & \Leftrightarrow & \mathcal{M}, w \models \varphi \text{ and } \mathcal{M}, w \models \psi \\
\mathcal{M}, w \models [\alpha]\varphi & \Leftrightarrow & \forall w'[w' \in [\![\alpha]\!](w) \Rightarrow \mathcal{M}, w' \models \varphi] \\
\mathcal{M}, w \models <\alpha>\varphi & \Leftrightarrow & \exists w'[w' \in [\![\alpha]\!](w) \wedge \mathcal{M}, w' \models \varphi] \\
\mathcal{M}, w \models V & \Leftrightarrow & w \notin \text{opt} \\
\mathcal{M}, w \models \Box\varphi & \Leftrightarrow & \forall w'[w' \in W \Rightarrow \mathcal{M}, w' \models \varphi]
\end{array}
$$

Validity is defined as before.

The approach of reducing standard deontic logic to dynamic logic has the advantage of allowing *integration* of a logic for ought-to-be sentences a logic for ought-to-do sentences. The integration requires no particular technical step besides, of course, the definition of modal and dynamic ought in terms of Anderson and Meyer reduction, respectively. Semantically we adopt an S5-semantics adapted in order to cope with the additional axiom $\Diamond\neg V$ for modal formulas and retain the standard semantics of PDL for dynamic ones.

3.5 Some properties and deontic paradoxes analyzed in $PDeL^{AM}$

The following proposition uses some operators that have not been defined here but are treated in detail elsewhere [Mey88, WWMD91]. If α_1 and α_2 are actions, then

- $\alpha_1 + \alpha_2$ is the process that does α_1 or α_2,
- $\alpha_1 \& \alpha_2$ is the process that does α_1 and α_2 simultaneously, and
- $\alpha_1; \alpha_2$ is the process that first does α_1 and then does α_2.

Similar operators have also been studied by Segerberg [Seg82].

Proposition 1 *The following formulas are theorems in $PDeL^{AM}$. The dashes indicate that there is no syntactic counterpart for an ought-to-be or ought-to-do formula.*

1a	—	*1b*	$\vdash \hat{F}(\alpha_1;\alpha_2) \leftrightarrow [\alpha_1]\hat{F}(\alpha_2)$
2a	$\vdash F(\varphi_1)\wedge F(\varphi_2) \leftrightarrow F(\varphi_1\vee\varphi_2)$	*2b*	$\vdash \hat{F}(\alpha_1)\wedge\hat{F}(\alpha_2) \leftrightarrow \hat{F}(\alpha_1+\alpha_2)$
3a	$\vdash F(\varphi_1)\vee F(\varphi_2) \rightarrow F(\varphi_1\wedge\varphi_2)$	*3b*	$\vdash \hat{F}(\alpha_1)\vee\hat{F}(\alpha_2) \rightarrow \hat{F}(\alpha_1\&\alpha_2)$
4a	—	*4b*	$\vdash \hat{O}(\alpha_1;\alpha_2) \leftrightarrow \hat{O}(\alpha_1) \wedge [\alpha_1]\hat{O}(\alpha_2)$
5a	$\vdash O(\varphi_1)\vee O(\varphi_2) \rightarrow O(\varphi_1\vee\varphi_2)$	*5b*	$\vdash \hat{O}(\alpha_1)\vee\hat{O}(\alpha_2) \rightarrow \hat{O}(\alpha_1+\alpha_2)$
6a	$\vdash O(\varphi_1)\wedge O(\varphi_2) \leftrightarrow O(\varphi_1\wedge\varphi_2)$	*6b*	$\vdash \hat{O}(\alpha_1)\wedge\hat{O}(\alpha_2) \leftrightarrow \hat{O}(\alpha_1\&\alpha_2)$
7a	$\vdash O(\varphi\rightarrow\psi)\rightarrow(O\varphi\rightarrow O\psi)$	*7b*	—
8a	$\vdash \Box(\varphi\rightarrow\psi)\rightarrow\Box(O\varphi\rightarrow O\psi)$	*8b*	—
9a	$\varphi\rightarrow\psi \vdash O\varphi\rightarrow O\psi$	*9b*	—
10a	$\vdash O\varphi\rightarrow OO\varphi$	*10b*	$\nvdash \hat{O}\alpha\rightarrow O\hat{O}(\alpha)$ *but* $\hat{O}(\alpha)\vdash O(\hat{O}(\alpha))$
11a	$\vdash \Box\varphi\rightarrow O\varphi$	*11b*	—
12a	$\vdash O\varphi\rightarrow\Box O\varphi$	*12b*	$\nvdash \hat{O}\alpha\rightarrow\Box\hat{O}\alpha$
13a	$\vdash O\varphi\rightarrow\neg O\neg\varphi$	*13b*	$\nvdash \hat{O}\alpha\rightarrow\neg\hat{O}\overline{\alpha}$
14a	$\vdash \neg(O\varphi\wedge O\neg\varphi)$	*14b*	$\nvdash \neg(\hat{O}(\alpha)\wedge\hat{O}(\overline{\alpha}))$
15a	$\vdash O\varphi\rightarrow P\varphi$	*15b*	$\nvdash \hat{O}(\alpha)\rightarrow\hat{P}\alpha$
16a	$\vdash \neg O\varphi\rightarrow O\neg O\varphi$	*16b*	$\nvdash \neg\hat{O}\alpha\rightarrow O\neg\hat{O}\alpha$

The theorems and non-theorems in this proposition follow the interesting discussion of Anderson's reduction of deontic logic to alethic logic by McArthur [McA81].

1-6 The ought-to-be statements are standard theorems that should hold in any deontic logic [McA81]. The sequential ought-to-do statements in *4b* (i.e. $\hat{O}(\alpha_1;\alpha_2)$) have no counterpart in a language for ought-to-be.

7a This is the K-rule that must be part of any normal modal logic (cf. [Che80]). There is no corresponding axiom in our language for ought-to-do. However, a possible future extension is to add an *involvement* operator, say $\Rightarrow$, such that $\alpha \Rightarrow \beta$ means "doing α involves doing β" Axiomatization of this operator is a topic of current research.

8a This is the Good Samaritan paradox. An intuitive reading of the theorem in a deontic language that does not distinguish actions from states is "necessarily, if the good Samaritan helps Jones who was robbed, then necessarily, if Jones ought to be helped, then Jones ought to be robbed." In $PDeL^{AM}$, such a reading is incorrect, because we would then represent actions by letters that are supposed to represent states. We think that if [8a] is read as a statement about ideal or desirable states, it is intuitively harmless: "necessarily, if the state φ obtains only if state ψ obtains, then necessarily, if φ is desirable, then ψ is desirable".

9a Again, we think this theorem is intuitively harmless if read, as it should be, as a theorem about ideal/desirable states.

10 The ought-to-be theorem $O\varphi \rightarrow OO\varphi$ says that we cannot express "meta-ought-to-be" statements as distinct from unnested ough-to-be statements. There is no strict ought-to-do counterpart of this, since $\hat{O}\hat{O}\alpha$ is ungrammatical. Interestingly, we can express mixed modalities, as shown in [10b]. It is *not* a theorem that if α is obligatory, then it is ideal/desirable that it is obligatory: $\nvdash \hat{O}\alpha \rightarrow O\hat{O}(\alpha)$. However, if we can prove the validity of $\hat{O}\alpha$, then this is so in all possible states of the world, so this is so in particular in the ideal/desirable states of the world: $\hat{O}(\alpha) \vdash O(\hat{O}(\alpha))$.

11a Read as a statement about ought-to-do, $\Box\varphi \rightarrow O\varphi$ is counterintuitive, but read, as it should be, as a statement about states, it is harmless: "If φ is necessarily true, then it is true in all ideal/desirable states of the world".

12 [12a] says that there is only one set of ideal/desirable states. If φ is an ideal state, then it is in all states of the world the case that φ is an ideal state. In other words, this theorem says that we cannot revise our idea of what is ideal depending on the current state of the world. A similar statement holds for obligatory actions.

13–15 The ought-to-be theorems state that if a state is ideal, then it not at the same time ideal/desirable that the state does not hold [13a]; in other words, desires are consistent, because it will not be the case that a state is desirable and undesirable at the same time [14a]; and yet another way of saying this is that if a state is desirable, it is permitted [15a]. A consequence of our reading of O as "desirable" is that we must read P as "compatible with the desired state of affairs" or, more briefly, as "compatible with our desires". There are no analogous statements for the ought-to-do case. It is very well possible to be in an obligation to perform α as well as its complement; and if you ought to do α, there may very well be reasons why α is forbidden at the same time. The ought-to-be theorems [13–15] all follow from the axiom $\Diamond\neg V$. In applications where we find the ought-to-be theorems [13–15] not satisfied, we can just drop this axiom. Take for example a robot that must work under adverse conditions. Later in this paper, we will introduce more than one violation state, so that this possibility can be modeled more realistically.

16 Due to our choice to take S5 for the $\Box$-modality, we also "gain" the theorem $\neg O\varphi \rightarrow O\neg O\varphi$.

Proof Since in S5 we have that $\vdash \Diamond\psi \rightarrow \Box\Diamond\psi$. We have as an instance that

$$\vdash \Diamond(\neg\varphi \wedge \neg V) \rightarrow \Box\Diamond(\neg\varphi \wedge \neg V),$$

and so, by the fact that $\psi_1 \rightarrow \psi_2$ entails $\vdash \Box\psi_1 \rightarrow \Box\psi_2$, we obtain the following derivation:

$$\begin{array}{l}
\vdash \Diamond(\neg\varphi \wedge \neg V) \rightarrow \Box(\Diamond(\neg\varphi \wedge \neg V) \vee V) \\
\vdash \neg\Box(\neg\varphi \rightarrow V) \rightarrow \Box(\neg\Box(\neg\varphi \rightarrow V) \vee V) \\
\vdash \neg\Box(\neg\varphi \rightarrow V) \rightarrow \Box(\Box(\neg\varphi \rightarrow V) \rightarrow V) \\
\vdash \neg\Box(\neg\varphi \rightarrow V) \rightarrow \Box(O\varphi \rightarrow V) \\
\neg O\varphi \rightarrow O\neg O\varphi.
\end{array}$$

■

This means that our logic for ought-to-be satisfies what Chellas calls "Deontic S5" or KD45. This is generally thought of as too strong a system for obligation and is rejected on those grounds [Che80]. It is also not true in ADL, since Anderson takes KT for $\Box$ instead of our choice of S5, for which KT $\nvdash \Diamond\psi \rightarrow \Box\Diamond\psi$. However, as we already argued in relation with [10], we are not interested in "meta-obligations"

(obligations of obligations), and thus we do not regard [16] as a problem. In fact, [10] and [16] together, deontic S5, yields that (negations of) obligations of (negations of) obligations can always be reduced to (negations of) obligations, so that "meta-obligations" simply do not enter the picture. Note, moreover, that (as in [10]) again we do not have a counterpart of 16a for ought-to-do's, nor is the formula $\neg\hat{O}\alpha \rightarrow O\neg\hat{O}\alpha$ a theorem.

4 The relation between ought-to-be and ought-to-do revisited: a formal appraisal

We are now in a position to go back to the attempted reductions of formulas of ought-to-be to ought-to-do listed in section 2 and evaluate them by means of the formal semantics of the combined system presented in the previous section. For each counterexample to a definition in section 2, we give a model that corresponds to it. In addition, a number of reductions to which we did not find counterexamples in section 2, will be proven valid in our formal semantics.

We will use models where the set $\mathcal{A}$ is finite (and possibly reduced to a single action), so that a first order analysis of the definitions is feasible without greater complications.

4.1 First attempt: there is an obligatory action that leads to state p

Definition 11 *A state is obligatory iff it is the result of an obligatory action α. Formally, this is now expressed as: For all models $\mathcal{M} = \langle \mathcal{A}, W, \pi, [\![\alpha]\!], \models, \text{opt}\rangle$ and for all $w \in W$:*

$$w \models O\varphi \text{ iff there is an action term } \alpha \in \underline{\mathcal{A}} \text{ such that } w \models [\alpha]\varphi \wedge \hat{O}\alpha \quad (1)$$

As we have pointed out in section 2.1, this formula lacks plausibility when interpreted as a modelling of real situations, since $1^{\rightarrow}$ assumes that given any state of affairs there is at least one action for bringing it about and $1^{\leftarrow}$ assumes that actions have a range that includes all possible worlds. Our semantics can easily cope with both cases. To counter $1^{\rightarrow}$, we can think of a model where $O\varphi$ is everywhere the case but where nevertheless there is not always an action for bringing about φ. Take, e.g., the model $\mathcal{M}$ defined as follows (figure 1):

(1) $\mathcal{A} = \{a, \overline{a}\}$;
(2) $W = \{w_0, w_1, w_2\}$;
(3) for all j and k, $w_j R w_k$;
(4) $[\![a]\!](w_0) = \{w_1, w_2\}$:
(5) $w_0 \in [\![\overline{a}]\!](w_0)$;
(6) $w_j \models (\neg\varphi \to V)$ for $j = 0 \ldots 2$;
(7) $w_0, w_2 \models \neg\varphi$, $w_1 \models \varphi$.

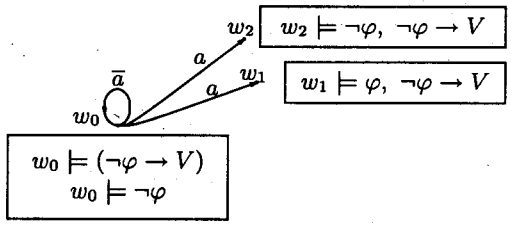

Figure 1: Counterexample to $1^{\rightarrow}$.

The model yields $w_0 \not\models \Box(\neg\varphi \to V) \to ([\alpha]\varphi \wedge [\overline{\alpha}]V)$ for α is a and $\overline{\alpha}$ is $\overline{a}$, thus refuting $1^{\rightarrow}$.

As a counterexample to $1^{\leftarrow}$, consider the following model (figure 2).

(1) $\mathcal{A} = \{a, \overline{a}\}$;
(2) $W = \{w_0, \ldots, w_3\}$;
(3) for all j and k, $w_j R w_k$;
(4) $[\![a]\!](w_0) = \{w_1, w_2\}$;
(5) $w_3 \in [\![\overline{a}]\!](w_0)$;
(6) $w_1 \models \varphi$ and $w_2 \models \varphi$;
(7) $w_3 \models V$;
(8) $w_0 \models \neg\varphi \wedge \neg V$;
(9) $w_1, w_2, w_3 \models \neg\varphi \to V$.

And, hence, we have that $w_0 \not\models ([a]\varphi \wedge [\overline{a}]V) \to \Box(\neg\varphi \to V)$ and a refutation of $1^{\leftarrow}$.

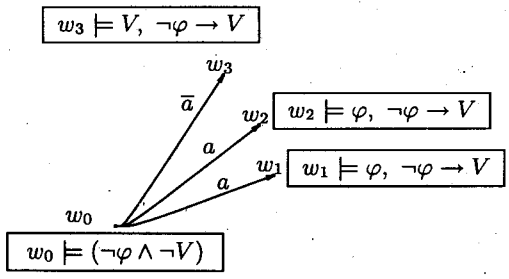

Figure 2: Counterexample to $1^{\leftarrow}$.

4.2 Second attempt: all actions that lead to φ are obligatory

Definition 12 *A state φ is obligatory iff all the actions that lead to the state of affairs φ are obligatory: For all models $\mathcal{M} = \langle \mathcal{A}, W, \pi, [\![\alpha]\!], \models, \mathrm{opt} \rangle$ and for all $w \in W$:*

$$w \models O\varphi \text{ iff for all } \alpha \in \underline{\mathcal{A}} \text{ we have } w \models [\alpha]\varphi \to \hat{O}\alpha \qquad (2)$$

In section 2.2, we had rejected the left to right implication ($2^{\rightarrow}$). A formal countermodel that substantiates this rejection is a model where, for instance, we have a world $w_0 \in [\![\overline{\alpha}]\!](w_0)$ such that $w_0 \models \neg V$, whereas the truth conditions for $O\varphi$ and $[\alpha]\varphi$ are satisfied (figure 3). This corresponds

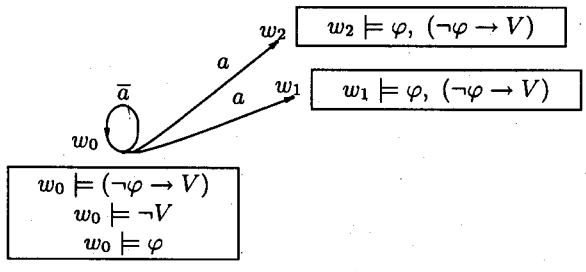

Figure 3: Counterexample to $2^{\rightarrow}$.

exactly to our informal counterargument in section 2.2: although by not performing α (say not doing any deposit on one's bank account) one stays within a desirable state of affairs (i.e., not being in the red), one cannot conclude from this that not depositing anything is obligatory! In the countermodel we indeed have that $O\varphi$ and $[a]\varphi$, but not $\hat{O}a$ ($= [\overline{a}]V$).

In section 2.2 we have already seen how the right to left implication ($2^{\leftarrow}$) may be rejected: simply take an action that does not lead to φ and assume that φ is not desirable. Formally this may be represented as follows (figure 4).

(1) $\mathcal{A} = \{a\}$;

(2) $W = \{w_0, w_1\}$;

(3) for all j and k, $w_j R w_k$;

(4) $[\![a]\!](w_0) = \{w_1\}$;

(5) $w_0 \models \neg\varphi \wedge \neg V$;

(6) $w_1 \models \neg\varphi$.

This suffices for rejecting $2^{\leftarrow}$. Since no action leads to φ, the right-hand

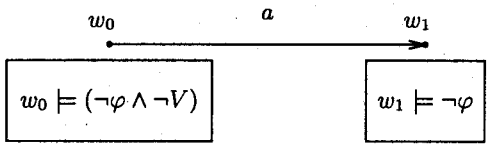

Figure 4: Counterexample to $2^{\leftarrow}$ and $4^{\leftarrow}$.

side of the definition is satisfied, but the left-hand side is not true. (This example uses the fact that there are no actions at all in the model that lead to φ, but this is not crucial.)

4.3 Third attempt: it is forbidden to undo state φ

Definition 13 *A state is obligatory iff to undo it is forbidden: Formally, this is now expressed as: For all models* $\mathcal{M} = \langle \mathcal{A}, W, \pi, [\![\alpha]\!], \models, \mathrm{opt} \rangle$ *and for all* $w \in W$*:*

$$w \models O\varphi \text{ iff for all } \alpha \in \underline{\mathcal{A}} \text{ we have } w \models [\alpha]\neg\varphi \rightarrow \hat{F}\alpha \qquad (3)$$

The left to right implication ($3^{\rightarrow}$) is derivable in every model for modal logics having ($\Box[\alpha]$) among their axioms.

Proposition 2 *For any* $\alpha \in \mathcal{A}$ *(the universe of actions),*

$$\vdash O\varphi \rightarrow ([\alpha]\neg\varphi \rightarrow \hat{F}\alpha)$$

Proof

1: $O\varphi$	
2: $\Box(\neg\varphi \rightarrow V)$	from 1 by axiom (O)
3: $[\alpha](\neg\varphi \rightarrow V)$	from 2 by axiom ($\Box[\alpha]$)
4: $[\alpha]\neg\varphi \rightarrow [\alpha]V$	from 3 by axiom (AK)

■

The proposition says that of a state of affairs is ideal, then any action that leads to its negation is forbidden. It implies that for all $\alpha \in \underline{\mathcal{A}}$, for all $\mathcal{M} = \langle \mathcal{A}, W, \pi, [\![\alpha]\!], \models, \text{opt} \rangle$ and for all $w \in W$,

$$w \models O\varphi \rightarrow ([\alpha]\neg\varphi \rightarrow \hat{F}\alpha)$$

Let's see what we can do with proposition 2. Let h be the state of affairs "the holder of a season-ticket and the person who uses it are the same person", and *lend* be the action " lending one's season-ticket to someone else for a travel". Clearly, [lend]$\neg h$ holds. Furthermore, suppose that h is an obligatory state of affairs, i.e., Oh holds. Then, we have by proposition 2, the following.

$$Oh \rightarrow \forall\alpha([\alpha]\neg h \rightarrow \hat{F}\alpha)$$

Instantiating this, we get

$$Oh \rightarrow ([\text{lend}]\neg h \rightarrow \hat{F}(\text{lend}))$$

By modus ponens and interdefinability of deontic operators (i.e., $\hat{F} = \neg\hat{P}$), we obtain $\neg\hat{P}$(lend), i.e., it is not permitted to lend one's season ticket to someone else for travelling.

Looking at the other direction, we can again find a model in which the right to left implication ($3^{\leftarrow}$) is false. Simply modify the countermodel for $2^{\leftarrow}$ by requiring that $w_1 \models \varphi$ (figure 5). We get thus a world w_0 where the consequent $\Box(\neg\varphi \rightarrow V)$ is false and the antecedent is true (because of the falsity of $[a]\neg\varphi$), so that $3^{\leftarrow}$ is falsified.

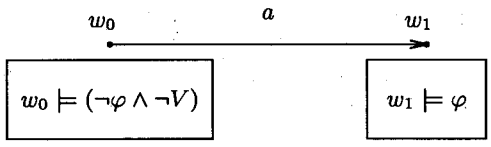

Figure 5: Counterexample to $3^{\leftarrow}$

4.4 Fourth attempt: all actions necessary and sufficient to reach state φ are obligatory

Definition 14 *A state φ is obligatory iff all the necessary and sufficient actions that lead to it are obligatory as well. Formally, for all models $\mathcal{M} = \langle \mathcal{A}, W, \pi, [\![\alpha]\!], \models, \mathrm{opt} \rangle$ and for all $w \in W$:*

$$w \models O\varphi \text{ iff for all actions } \alpha \in \underline{\mathcal{A}} \text{ we have } w \models (\alpha \rightsquigarrow \varphi) \rightarrow \hat{O}\alpha$$

This definition, though closely resembling 12, requires that given $O\varphi$, for $\hat{O}\alpha$ to hold besides $[\alpha]\varphi$ we should also have $[\overline{\alpha}]\neg\varphi$. This yields that in worlds belonging to the set $[\![\overline{\alpha}]\!](w_0)$, violation V is always the case. This explains why the left to right implication is always true in our semantics. More formally:

Proposition 3 *For any $\alpha \in A$,*

$$\vdash O\varphi \rightarrow ((\alpha \rightsquigarrow \varphi) \rightarrow \hat{O}\alpha)$$

Proof

1: $O\varphi$	
2: $\Box(\neg\varphi \rightarrow V)$	from 1 by axiom (O)
3: $[\overline{\alpha}](\neg\varphi \rightarrow V)$	from 2 by axiom ($\Box[\alpha]$)
4: $[\overline{\alpha}]\neg\varphi \rightarrow [\overline{\alpha}]V$	from 3 by axiom (AK)
5: $([\overline{\alpha}]\neg\varphi \rightarrow [\overline{\alpha}]V) \rightarrow (([\alpha]\varphi \wedge [\overline{\alpha}]\neg\varphi) \rightarrow [\overline{\alpha}]V)$	by propositional logic
6: $([\alpha]\varphi \wedge [\overline{\alpha}]\neg\varphi) \rightarrow [\overline{\alpha}]V$	from 4, 5 by modus ponens

■

An apparently counterintuive example application of this proposition is the following. In an ideal state of affairs, John is rich. The only action in the considered universe of actions that will make John rich is killing and robbing his neighbour. Therefore, John must kill and rob his neighbour. Wih respect to the fixed universe of possible actions available to John, this is indeed a valid deduction. Of course, the conflict with intuition arises because on moral grounds, we refuse to restrict ourselves to such a (brutally) limited universe of possible actions. In the kind of application of the logic that we have in mind, the specification of the behavior of computer systems, or of the real world as registered or controlled by computer systems, it is true that the system indeed does have a restricted universe of possible actions at its disposal.

The inverse of the implication in the proposition is, of course, false. Just take the countermodel for $2^{\leftarrow}$.

5 Application to normative system specification

5.1 Specification of the bank account example

We now return to the bank account example mentioned in the introduction. We assume in this example that we can use integer arithmetic. Although we have left this open until now, we can simply assume that first-order logic with equality is part of our language, and that an equational theory of the integers has been specified in this language. The theory of how this can be done is standard [EM85]. Now, let b be a variable whose value is the balance of an account, w(m) be the action "withdraw at least m from the account", and d(m) be the action "deposit at least m on the account", and let n and m be constants. In terms of our theory of actions, we take as atomic action names the set $\underline{\mathcal{A}} = \{\underline{w}_1, \underline{w}_2, \ldots, \underline{d}_0, \underline{d}_1, \underline{d}_2, \ldots\}$ and the corresponding set of actions $\mathcal{A} = \{w_1, w_2, \ldots, d_0 d_1, d_2, \ldots\}$. Intuitively, w_i is the action of withdrawing exactly i from the account, and d_i is the action of depositing exactly i on the account. Now the action term $\underline{w(m)}$ can be defined as a choice $\underline{w_m} + \underline{w_{m+1}} + \cdots$, which we will write as $\underline{\Sigma_{n \geq m} w_m}$. This formalizes the meaning "withdraw at least m from the account". Analogously, $\underline{d(m)}$ is defined as $\Sigma_{n \geq m} \underline{d_m}$. Following our approach of section 3, the step semantics of the atomic action terms $\underline{w_i}$ and $\underline{d_i}$ is defined as follows:

- $[\![\underline{w_i}]\!]_S = \{S \subseteq \mathcal{A} \mid w_i \in S\}$
- $[\![\underline{d_i}]\!]_S = \{S \subseteq \mathcal{A} \mid d_i \in S\}$

The effcts of the events w_i and d_i are given as follows: If $\pi(b,u) = n$ (the balance in state u is n), then we have

- $eff(w_i)(u) = u'$ with $\pi(b,u') = n - i$,
- $eff(d_i)(u) = u'$ with $\pi(b,u') = n + i$.

We know that when we are interested in the state-transition semantics of action terms, it is sufficient to consider compatible steps only. It is easy to see that the only compatible steps in our present universe of actions are singleton steps of the form $[w_i]$ and $[d_i]$. Determination of the state-transition semantics of action terms $w(m)$, $d(m)$, $\overline{w(m)}$ and $\overline{d(m)}$ is now easy:

- $$[\![w(m)]\!](u) = eff([\![\sum_{n \geq m} \underline{w_n}]\!])(u)$$
$$= eff(\bigcup_{n \geq m} [\![\underline{w_n}]\!])(u)$$
$$= \bigcup_{n \geq m} eff([\![\underline{w_n}]\!])(u)$$
$$= \bigcup_{n \geq m} eff(\{[w_n]\})(u)$$

- $$[\![d(m)]\!](u) = \bigcup_{n \geq m} eff([\![\underline{d_n}]\!])(u)$$
$$= \bigcup_{n \geq m} eff(\{[d_n]\})(u)$$

- $$[\![\overline{w(m)}]\!](u) = eff(STEPS \setminus w(m))(u)$$
$$= eff(STEPS \setminus [\![\sum_{n \geq m} \underline{w_n}]\!])(u)$$

$$= eff(STEPS \setminus \bigcup_{n \geq m} [\![\underline{w_n}]\!])(u)$$

$$= eff(\bigcap_{n \geq m} STEPS \setminus [\![\underline{w_n}]\!])(u)$$

$$= eff(\bigcup_{n<m} \{[w_n]\} \cup \bigcup_{l \geq 0} \{[d_l]\})(u)$$

$$= \bigcup_{n<m} eff(\{[w_n]\})(u) \cup \bigcup_{l \geq 0} eff(\{[d_l]\})(u).$$

This says that the negation of "withdraw at least m" is to withdraw less than m or deposit anything.

- $[\![\overline{d(m)}]\!](u) = \bigcup_{n<m} eff(\{[d_n]\})(u) \cup \bigcup_{l \geq 1} eff(\{[w_l]\})(u)$. This says that the negation of "deposit at least m" is to deposit less than n or withdraw anything.

Combined with the idea that a world with $b < 0$ is non-ideal, we may specify our informal constraints by means of the following axioms, which are validated by the above semantics.

$$O(b \geq 0) \tag{1}$$

$$(b = n \wedge m \geq 0) \rightarrow [\mathrm{w}(m)](b \leq n - m) \tag{2}$$

$$(b = n \wedge m \geq 0) \rightarrow [\overline{\mathrm{w}(m)}](b > n - m) \tag{3}$$

$$(b = n \wedge m \geq 0) \rightarrow [\mathrm{d}(m)](b \geq n + m) \tag{4}$$

$$(b = n \wedge m \geq 0) \rightarrow [\overline{\mathrm{d}(m)}](b < n + m) \tag{5}$$

Note that $d(m)$ is defined as a deposit of *at least* m, but possibly more. Thus, a deposit of 10 Guilders is also a deposit of 5 Guilders — in both cases, we have a deposit of *at least* 5 Guilders. Similarly, $w(m)$ is a withdrawal of *at most* m. A consequence of these formalizations is that $\hat{O}(d(m))$ is the obligation to deposit at least m. This will play an important role in the proofs later. These formalizations are straightforward in this example. In general we may consider other choices.

We can now prove one of the results listed in the introduction as desired, viz. that if the balance is negative, it is forbidden to withdraw money.

Proposition 4 *Let $n, m \geq 0$, then (1) + (2) imply the following formula.*

$$(b = n \wedge n - m < 0) \rightarrow \hat{F}(\mathrm{w}(m))$$

Proof

1: $(b = n \wedge n - m < 0) \rightarrow [\mathrm{w}(m)](b \leq n - m)$	from axiom (2) of the specification
2: $(b = n \wedge n - m < 0) \rightarrow [\mathrm{w}(m)](b < 0)$	from 1
3: $(b = n \wedge n - m < 0) \rightarrow [\mathrm{w}(m)]\neg(b \geq 0)$	from 2
4: $\Box(\neg(b \geq 0) \rightarrow V)$	(*) from axiom (1) of the specification and Axiom (O)
5: $[w(m)](\neg(b \geq 0) \rightarrow V)$	from 4 by Axiom ($\Box[\alpha]$)
6: $[w(m)](\neg(b \geq 0)) \rightarrow [\mathrm{w}(m)]V$	from 5 by Axiom (K),
7: $(b = n \wedge n - m < 0) \rightarrow [\mathrm{w}(m)]V$	from 3,6 by double MP
8: $(b = n \wedge n - m < 0) \rightarrow \hat{F}(\mathrm{w}(m))$	from Def. 8

■

In addition, we can prove that if an account has a negative balance, it is obligatory to deposit an amount that is sufficient to make it non-negative again.

Proposition 5

$$(b = n \wedge n < 0) \rightarrow \hat{O}(\mathrm{d}(-n))$$

Proof

1:$(b = n \wedge -n \geq 0) \rightarrow [\overline{\mathrm{d}(-n)}](b < n + (-n))$	from axiom (5) of the specification
2:$\Box(\neg(b \geq 0) \rightarrow V)$	from axiom (1) of the specification and Axiom (O)
3:$[\overline{d(-n)}](\neg(b \geq 0) \rightarrow V)$	from 2 by Axiom ($\Box[\alpha]$)
4:$[\overline{d(-n)}](\neg(b \geq 0)) \rightarrow [\overline{\mathrm{d}(-n)}]V$	from 3 by axiom K
5:$(b = n \wedge -n \geq 0) \rightarrow [\overline{\mathrm{d}(-n)}]V$	from 1, 4 by MP

■

We may now wonder whether it is obligatory or permitted to deposit more than simply $-n$ if the balance is $-n$. Somewhat surprisingly, neither the obligation nor the permission to deposit more than necessary can be derived in the current system. First, look at the following formula:

$$\not\vdash b = n \wedge n < 0 \wedge m > -n \rightarrow \hat{O}(d(m)) \tag{7a}$$

The derivation of the formula is blocked because we can only derive $b = n \wedge n < 0 \wedge m > -n \rightarrow [\overline{\mathrm{d}(m)}](b < n + m)$, and from this we cannot derive $b = n \wedge n < 0 \wedge m > -n \rightarrow [\overline{\mathrm{d}(m)}](b < 0)$, since from $b < n + m$ it does not follow that $b < 0$ (even though $n + m > 0$). Hence, we cannot apply axiom (1) of the specification in order to conclude $\hat{O}(d(m))$.

As an aside, we may remark that also the negation of the obligation cannot be derived. Thus, the following formula cannot be proved either:

$$(b = n \wedge n < 0) \rightarrow \neg\hat{O}(d(m)) \tag{7b}$$

We shall show this in the next section in a slightly refined setting.

Turning to permission, we note that the following formula cannot be proved:

$$b = n \wedge n < 0 \wedge m > -n \rightarrow \hat{P}(d(m)) \tag{7c}$$

In order to derive this formula, we would have to make two extra assumptions, $\Box\langle d(m)\rangle\top$ and $Op \leftrightarrow \Box(\neg p \leftrightarrow V)$. The former is required in order to express that depositing is always possible and is not controversial. The latter is required since we need $\neg(b \geq 0)$ to be not only a *sufficient* condition for the obtaining of the violation (as (*) states in proposition 4) but also a *necessary* one. Since V signalizes a generic violation, it would be incorrect, given V, to assert that this or that particular norm has been violated. For this purpose, we need a refinement of our formalization. We give such a refinement in the following section.

5.2 Refinement of the logic

The refinement concerns the formalization of the violation atom and consequently of deontic formulas. In particular, we will relate a norm and its violation to a certain piece of legislation where the norm is considered.

We divide the space of norms in classes, the typical representant of which is i. We say that two norms are of the same type i if and only if their

violations have both the form V_i. The index i refers to a piece of legislation, such as a paragraph of a code of laws, which we do not specify further. We thus specialize the idea of violation to the idea of violation of a part of the law. Here another advantage of speaking of violation rather than of sanction shows itself. In the case where we would interpret V as "sanction" we would have assumed that the piece i of legislation had mentioned sanctions for the trespassing of the relevant norms. Now that we interpret V as violation, such an assumption is not necessary.

The reason for adopting this device is more of conceptual than technical character. As we have already pointed out earlier, until now we have spoken of violation in a very general way. When we speak of nonfulfilment of a norm we say that the whole normative system to which the norm belongs has been violated. This is so only by way of approximation, since we know exactly where the violation has arisen. We indicate this by the atom V_i but in addition keep our general atom V, which still has the meaning "a violation has occurred". Obviously, violation of a particular norm implies what we might call a generic violation of the normative system to which that particular norm belongs. Thus, in the system introduced below, we have

$$V_i \rightarrow V. \tag{8}$$

We achieve this by defining V as the (finite) disjunction of all violations that can take place within a given normative system.

Definition 15 *Let $i \in \mathbb{N}$ refer to a generic piece of legislation in a given normative system S. We define the following*

$$V =_{\text{def}} \bigvee_i V_i$$

Note that under this rewriting of violation, for every norm (obligations and prohibitions) of type i, we have a permission of the same type. This depends upon the validity in our logic of the equivalence $\Box(\neg p \rightarrow V_i) \equiv \neg\Diamond(\neg p \wedge \neg V_i)$. This implies that the piece of legislation to which i refers has to be interpreted as stating both which norms fall under ought-to-be and which under ought-to-do, and also what is permitted (or compatible) with these norms.

The introduction of a flagged violation atom requires a minor change in the deontic axioms of our system.

Axioms 5.1

$$F_i p \leftrightarrow \Box(p \rightarrow V_i) \qquad (\mathrm{F_i})$$

$$O_i p \leftrightarrow \Box(\neg p \rightarrow V_i) \qquad (\mathrm{O_i})$$

$$P_i p \leftrightarrow \neg F_i p \qquad (\mathrm{P_i})$$

$$\hat{F}_i \alpha \leftrightarrow [\alpha] V_i \qquad (\hat{\mathrm{F}}_\mathrm{i})$$

$$\hat{O}_i \alpha \leftrightarrow [\overline{\alpha}] V_i \qquad (\hat{\mathrm{O}}_\mathrm{i})$$

$$\hat{P}_i \alpha \leftrightarrow \neg \hat{F}_i \alpha \qquad (\hat{\mathrm{P}}_\mathrm{i})$$

Proposition 6 *In $PDeL^{AM}$ with flagged violations, we have the following (non)theorems.*

1. $\vdash F(V)$
2. $\vdash F(V_i)$
3. $\nvdash F_i(V)$

We cannot derive $F_i(V)$ because we cannot derive $\Box(V_i \rightarrow V_j)$ for $i \neq j$. Using flagged violations, we can still derive propositions 4 and 5, but now in a more informative way. If we replace axiom (1) with $O_1(b \geq 0)$, then we can prove that under certain conditions, we have $\hat{F}_1(w(m))$ (i.e. withdrawal of m would cause a violation of axiom (1)) and that under other conditions, we have $\hat{O}_1(d(-n))$ (i.e. depositing of $-n$ is obligatory under rule (1)).

We now turn to the derivation of (7c), which says that it is permitted to deposit more than is necessary to make a balance positive:

$$b = n \wedge n < 0 \wedge m > -n \rightarrow \hat{P}(d(m)) \qquad (7\mathrm{c})$$

On assumption we need for this is that if φ is obligatory, then $\Box(\neg\varphi \leftrightarrow V)$. As stated earlier, this cannot be true for the general violation atom V. We now note that, by contrast, for particular violations V_i, this is a plausible assumption. We now define an obligation operator that says the $\neg\varphi$ is the *only* way in which the violation flag V_i can be raised.

Definition 16

$$O'_i p \stackrel{\mathrm{def}}{=} \Box(\neg p \leftrightarrow V_i)$$

The specification of bank account behavior is now changed by replacing (1) by

$$O'_i(b \geq 0) \qquad (1')$$

We also add the intuitively plausible axiom that money can always be deposited ($m \geq 0$):

$$\Box\langle d(m)\rangle\top \tag{9}$$

More precisely, this formula states that in any world it is possible to perform an action $d(m)$, i.e. it is always possible to deposit some amount $m \geq 0$. This is obvious, but here we see the need for specifying this explicitly. Now we can prove formula (7).

Proposition 7 *Let* $m > -n$,

$$b = n \wedge n < 0 \rightarrow \hat{P}_i(d(m)) \tag{7}$$

Proof

1: $\Box(b \geq 0 \leftrightarrow \neg V_i)$	from Def. 16 and (1')
2: $b = n \wedge n < 0 \rightarrow [d(m)](b \geq 0)$	from axiom (4) of the specification
3: $b = n \wedge n < 0 \rightarrow [d(m)]\neg V_i$	from 1,2 by Axiom ($\Box[\alpha]$)
4: $b = n \wedge n < 0 \rightarrow \langle d(m)\rangle\neg V_i$	from 3 by (9)

■

Note how we are forced to make all assumptions explicit when we formalize the specification of bank account behavior.

We conclude with a few remarks about the non-derivability of formula

$$(b = n \wedge n \leq 0 \wedge m > -n) \rightarrow \neg\hat{O}_i(d(m)) \tag{7b}$$

In order to prove this we should be able to prove the following (simply apply axiom 5.1).

$$(b = n \wedge n \leq 0 \wedge m > -n) \rightarrow \langle\overline{d(m)}\rangle(b \geq 0) \tag{10}$$

We only know that $[\overline{d(m)}](b < n + m)$ holds (with $n + m > 0$), that is to say that by not-depositing m we get into a state (world) where the balance b is less than some positive number. However, to derive (10), we need to know whether $\langle\overline{d(m)}\rangle b > 0$, i.e. despositing m may result in a state where the balance is positive. This does *not* follow from the previous assertion: $[\overline{d(m)}]b < n + m$ allows for the situation where after performing $\overline{d(m)}$ the

balance is negative in all cases. Of course, this situation is not true in this particular case. In order to make this explicit, we must refine our specification with the following axioms:

$$b = n \wedge m \geq 0 \rightarrow \langle d(m) \rangle (b = k), \ for\ all\ k \geq n + m$$
$$b = n \wedge m \geq 0 \rightarrow \langle \overline{d(m)} \rangle (b = k), \ for\ all\ k < n + m.$$

These are clearly valid with respect to our semantics, as the reader may verify him/herself. Now we have sufficient information to derive what we want to derive. The second axiom gives us what we need to derive (10). Again this need for additional specifications is not really surprising. We should simply be aware that when we give a specification of constraints originally expressed by means of natural language formulations, we may easily forget a good deal of the assumptions implicit in the use of those formulations.

6 Discussion and conclusion

We have shown that the distinction between ought-to-do and ought-to-be is relevant for at least some kinds of system specification and that a few candidates for playing the role of link between the two notions are intuitively not valid. However, by maintaining a certain degree of generality and using Anderson's and Meyer's reduction for ought-to-be and ought-to-do, respectively, to an alethic modal (dynamic) logic, we have seen that it is possible to express both kinds of norms in one integrated system without reducing one of them to the other. In the integrated system, no specific relations between them are assumed other than those that follow immediately from both reductions to alethic modal logic. We also presented a refinement of the system, in which violations are indexed by the piece of legislation that is violated, and in which we assume an equivalence between the non-occurrence of an obligatory state and the corresponding violation state.

In both integrated systems, without and with flagged violation states, the notions of ought-to-do and ought-to-be remain rather less intrinsically (logically) related than one might expect. The only relations that do hold follow from Anderson's and Meyer's reductions to an alethic modal logic in which a violation atom plays a prominent role, since the proposed equivalential definitions are only partially valid in our system. In this sense the relation between ought-to-be and ought-to-do remains rather extensional.

We have used a formalism that heavily abstracts from what happens in real-life situations. Our system is a very general one that can cover many concrete cases. This results in having less valid assertions about the relationship between ought-to-be and ought-to-do. For example, our formalization of the intuitive counterexamples to the definitions that reduce ought-to-be to ought-to-do hinge on the non-availability of certain actions. In applications where this non-availability is impossible — where all actions can always be performed, for example — these counterexamples are not realistic. In such applications, one of the reductions of ought-to-be to ought-to-do may be very well be valid, even though in *different* contexts there are counterexamples to it. If we restrict ourselves to such applications, we may safely add one of the reductions we discussed in this paper without fear of inconsistency. Our logic can thus be viewed as a basic platform on which more specific structure can be built to specify concrete situations.

Like all logics, $PDeL^{AM}$ has its limitations. A number of important issues from deontic logic remain unresolved in $PDeL^{AM}$. We review some of these in the following paragraphs.

6.1 Conflict of duties

Consider the following example.

> Suppose you are in a situation where you ought to pay a debt but you have not enough money on your bank account to pay it. Of course, if you pay the debt, then you are in the red (φ) but even if paying the debt is obligatory, being in the red cannot be desirable; in fact, it is not at all.

Here we see a conflict between the obligation to pay a debt and that of being not in the red on a bank account. A kind of priority of obligations is needed in order to solve this conflict. This is related to a well-known topic in the area of AI and Law, viz. of defeasible reasoning and inconsistency handling. We do not deal with this issue in this paper.

6.2 Derived consequences of obligatory actions need not be obligatory

Another situation that $PDeL^{AM}$ is not fit to deal with consists of obligations that imply states of affairs having not desirable consequences. Consider, e.g., a case as follows: doing an action α or not doing it spans the whole space of

possible worlds (i.e., $\alpha+\overline{\alpha}$ is equal to the universal action having all possible worlds as its range); performing α necessarily results in a state of affairs φ. Action α is obligatory but still the result φ is not a desirable state of affairs.

Proposition 8 *In a model where each state is reachable by α or $\overline{\alpha}$, we have* $[\alpha]\varphi \wedge \hat{O}\alpha \rightarrow O\varphi$.

Proof

1: $[\alpha]\varphi$	
2: $[\alpha](\neg\varphi \rightarrow V)$	from 1 by modal reasoning
3: $\hat{O}\alpha$	
4: $[\overline{\alpha}]V$	from 3 by axiom ($\hat{\text{O}}$)
5: $[\overline{\alpha}](\neg\varphi \rightarrow V)$	from 4 by modal reasoning
6: $[\alpha+\overline{\alpha}](\neg\varphi \rightarrow V)$	from 2,5 by dynamic logic
7: $\Box(\neg\varphi \rightarrow V)$	from 6 by the universality assumption for $\alpha+\overline{\alpha}$
8: $O\varphi$	from 7 by axiom (O)

■

A concrete example of this situation might be the following (assuming that fasten or not fasten the seat belts is universal in the sense given above):

> Let φ be the state in which we have restricted freedom of movement when seated in a car. There is an obligatory action, viz. fastening your seatbelts, that leads to φ, but from this we do not want to conclude that φ is obligated. We have here a consequence of fastening the seat belts that is somehow not considered desirable.

6.3 Conditional obligations

Another aspect we mentioned already in earlier counterexamples is that obligations to do something often are conditional upon certain facts, e.g.,

fastening the seat belts upon driving a car. This is, of course, not considered in the definitions given above (see the short remark in section 4.2).

Viewed logically, however, the most important difficulty in our system is that relevant relations among the occurrence of certain facts and the holding of certain norms are simply neglected. The formula $O\varphi \rightarrow ([\alpha]\neg\varphi \rightarrow \hat{F}\alpha)$, which is provable in our system, should be understood as a sort of justification: as a justification of saying that α is forbidden, we may argue that α undoes a desirable state. However, our system works with material implication, so that the above formula merely says that $O\varphi$ is a sufficient condition for the truth of $[\alpha]\neg\varphi \rightarrow \hat{F}\alpha$. This is too weak to let $O\varphi$ function as part of a justification of $\hat{F}\alpha$.

Because of the shortcomings discussed above, we might look at still other ways to integrate ought-to-do and ought-to-be modalities. Firstly, we might strenghten the relation between actions and state of affairs, for instance, in the direction pointed out by Segerberg in [Seg89], explaining actions in terms of "bringing it about that ...". Secondly, as pointed out above, we might opt for an implication connective different from material implication. We plan to look at some of these possibilities in the future.

Acknowledgements: This paper benefited from detailed comments given by Bernd van Linder on an earlier version. We also want to thank the participants of the AAAI 1993 Spring Symposium, held at Stanford University, and those of the Workshop on Deontic and Nonmonotonic Logics, held in 1993 at EURIDIS, Rotterdam, for discussing earlier versions of this paper. The paper also has benefited from the constructive criticisms of the anonymous referees.

References

[AM57] Alan Ross Anderson and Omar K. Moore. The formal analysis of normative concepts. *The American Sociological Review*, 22:9–17, 1957. reprinted in: I.M.Copi–J.A.Gould (eds) *Contemporary Readings in Logical Theory* MacMillan, New York 1967.

[Åqv88] Lennart Åqvist. *Introduction to deontic logic and the theory of normative systems.* Bibliopolis, Napoli, 1987 [1988].

[BMT87] C. Biagioli, P. Mariani, and D. Tiscornia. ESPLEX: A rule and conceptual based model for representing statutes. In *The First International Conference on Artificial Intelligence and Law*, pages 240–251. ACM, May 1987.

[Cas60] Hector-Neri Castañeda. Obligation and modal logic. *Logique et Analyse*, 3:40–48, 1960.

[Cas70] Hector-Neri Castañeda. On the semantics of the ought–to–do. *Synthese*, 21:449–468, 1970.

[Cas81] Hector-Neri Castañeda. The paradoxes of deontic logic: the simplest solution to all of them in one fell swoop. In Risto Hilpinen, editor, *New Studies in Deontic Logic*, pages 37–85. D. Reidel, Dordrecht – Boston – London, 1981.

[Che80] Brian F. Chellas. *Modal Logic. An Introduction.* Cambridge University Press, Cambridge, 1980.

[Coe93] J. Coenen. Top-down development of layered fault-tolerant systems and its problems —a deontic perspective. *Annals of Mathematics and Artificial Intelligence*, 9:133–150, 1993.

[DM90] F.P.M. Dignum and J.-J.Ch. Meyer. Negations of transactions and their use in the specification of dynamic and deontic integrity constraints. In M.Z. Kwiatkowska, M.W. Shields, and R.M. Thomas, editors, *Semantics for Concurrency*, pages 61–80. Springer, 1990.

[EM85] H. Ehrig and B. Mahr. *Fundamentals of Algebraic Specification 1. Equations and Initial Semantics.* Springer, 1985. EATCS Monographs on Theoretical Computer Science, Vol. 6.

[Gar86] J. L. A. García. The *tunsollen*, the *seinsollen*, and the *soseinsollen*. *American Philosophical Quarterly*, 23:267–276, 1986.

[Gea81] Peter Geach. Whatever happened to deontic logic? *Philosophia*, 11:1–12, 1981.

[GMP89] J. Glasgow, G. MacEwen, and P. Panangaden. Security by permission in databases. In C.E. Landwehr, editor, *Database Security II: Status and Prospects*, pages 197–205. North-Holland, 1989. Results of the IFIP WG 11.3 Workshop on Database Security (October 1988), Kingston, Ontario, Canada.

[Har84] D. Harel. Dynamic logic. In D.M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic II*, pages 497–604. D. Reidel, 1984.

[Hum93] I.L. Humberstone. Zero–place operations and functional completeness, and the definition of new connectives. *History and Philosophy of Logic*, 14:39–66, 1993.

[KL88] S.O. Kimbrough and R.M. Lee. Logic modeling: A tool for management science. *Decision Support Systems*, 4:3–16, 1988.

[KM87] S. Khosla and T.S.E. Maibaum. The prescription and description of state based systems. In B. Banieqbal, H. Barringer, and A. Pnueli, editors, *Temporal Logic in Specification*, pages 243–294. Springer, 1987. Lecture Notes in Computer Science 398.

[Lee88a] R.M. Lee. Bureaucracies as deontic systems. *ACM Transactions on Office Information Systems*, 6:87–108, 1988.

[Lee88b] R.M. Lee. A logic model for electronic contracting. *Decision Support Systems*, 4:27–44, 1988.

[McA81] R.P. McArthur. Anderson's deontic logic and relevant implication. *Notre Dame Journal of Formal Logic*, 22:145–154, 1981.

[Mey88] J.-J.Ch. Meyer. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29:109–136, 1988.

[Mey89] J.-J.Ch. Meyer. Using programming concepts in deontic reasoning. In R. Bartsch, J.F.A.K. van Benthem, and P. van Emde Boas, editors, *Semantics and Contextual Expression*, pages 117–145. FORIS publications, Dordrecht/Riverton, 1989.

[ML85] N.H. Minsky and A.D. Lockman. Ensuring integrity by adding obligations to priviliges. In *8th IEEE International Conference on Software Engineering*, pages 92–102, 1985.

[MW93] J.-J.Ch. Meyer and R.J. Wieringa. Deontic logic: A concise overview. In J.-J.Ch. Meyer and R.J. Wieringa, editors, *Deontic Logic in Computer Science: Normative System Specification*, pages 3–16. Wiley, 1993.

[Seg82] K. Segerberg. A deontic logic of action. *Studia Logica*, 41:269–282, 1982.

[Seg89] Krister Segerberg. Bringing it about. *Journal of Philosophical Logic*, 18:327–347, 1989.

[Sta80] R. Stamper. LEGOL: Modelling legal rules by computer. In B. Niblett, editor, *Computer Science and Law*, pages 45–71. Cambridge University Press, 1980.

[WM93a] R.J. Wieringa and J.-J.Ch. Meyer. Actors, actions, and initiative in normative system specification. *Annals of Mathematics and Artificial Intelligence*, 7:289–346, 1993.

[WM93b] R.J. Wieringa and J.-J.Ch. Meyer. Applications of deontic logic in computer science: A concise overview. In J.-J.Ch. Meyer and R.J. Wieringa, editors, *Deontic Logic in Computer Science: Normative System Specification*, pages 17–40. Wiley, 1993.

[WMW89] R.J. Wieringa, J.-J. Ch. Meyer, and H. Weigand. Specifying dynamic and deontic integrity constraints. *Data and Knowledge Engineering*, 4:157–189, 1989.

[WWMD91] R.J. Wieringa, H. Weigand, J.-J. Ch. Meyer, and F. Dignum. The inheritance of dynamic and deontic integrity constraints. *Annals of Mathematics and Artificial Intelligence*, 3:393–428, 1991.