Sanction Semantics and Contrary-to-Duty Obligations

Louwe B. Kuijer

University of Groningen
l.b.kuijer@rug.nl

Abstract. In this paper I show that one cannot faithfully represent contrary-to-duty obligations in logics with sanction semantics. In order to do so I first provide a number of desiderata that a logic should satisfy in order to represent contrary-to-duty obligations using sanction semantics. I then show that no logic satisfying all desiderata can faithfully represent contrary-to-duty obligations. Finally I show that when dropping any one of the desiderata there is a logic that satisfies all others and can represent some contrary-to-duty obligations faithfully.

1 Introduction

A well known problem in deontic logic is that of contrary-to-duty (CTD) obligations. CTD obligations are obligations that apply when breaking another obligation; the CTD obligation is essentially a mitigating factor for the broken obligation. The term was introduced in [1], where it was also shown that formalizing CTD obligations is quite hard. Over the years a number of solutions to the problem of CTD obligations have been proposed, some more successful than others. A reasonably complete overview can be found in [2] and [3].

One common approach in deontic logic is to capture obligations in terms of a 'sanction'. The idea, introduced in [4], is that you have an obligation to do ϕ, denoted $O(\phi)$, if and only if there is some kind of 'sanction', denoted $\mathbb{S}$, if you do not do ϕ. This 'sanction' can be an actual punishment but it need not be, in its most general form the 'sanction' merely represents the fact that 'there is wrongdoing'. Many deontic logics are either explicitly based on such a sanction or can be described by it. The best known example of such logics is the so-called Standard Deontic Logic (SDL) which can be given possible world semantics using sanctions. Other examples include [5–9]. The strength of reducing obligations to a sanction lies in its simplicity and intuitive plausibility; 'you have an obligation to do ϕ iff there is wrongdoing if you do not do ϕ' sounds like a tautology.

Sanction based deontic logics are not however very successful in faithfully representing CTD obligations. This is not surprising, as in a sanction based logic there are only two possible degrees of goodness/badness: $\mathbb{S}$ and $\neg\mathbb{S}$. Since a CTD obligation is a mitigating factor for breaking an obligation one would expect that faithfully representing it requires at least three degrees of goodness/badness: no

broken obligation, mitigated broken obligation and unmitigated broken obligation. This suggests that a straightforward reduction of obligations to sanction cannot faithfully represent CTD obligations.

Several attempts have been made to represent CTD obligations in deontic logics that use a more complicated reduction of obligations to sanction, such as the logics SA [5], deontic modal action logic (DMAL) [6] and $\mathrm{PD_eL}$ [9]. Unfortunately these logics all have problems that prevent them from faithfully representing CTD obligations. A description of the problems with these logics is given in Sect. 2.1 as it requires a more formal description of CTD obligations.

The inability of these logics to faithfully represent CTD obligations suggests that it may be impossible to do so at all using sanction based logics even when using complicated reductions of obligations to sanction. In this paper I show that it is indeed impossible to faithfully represent CTD obligations in a logic with sanction based semantics. In order to do this I define a number of desiderata that any semantics should satisfy in order to represent CTD obligations using sanction semantics. I then show that it is not possible to faithfully represent CTD obligations with semantics satisfying these desiderata.

The structure of the paper is as follows. In Sect. 2 I give a definition of a CTD obligation that is to be modeled, the well known gentle murder scenario. In Sect. 3 I give some technical preliminaries that are needed to formulate the desiderata. The desiderata are defined in Sect. 4. In Sect. 5 I prove that there is no logic that satisfies all the desiderata and faithfully models the gentle murder scenario. Finally, in Sect. 6 I show that when dropping any one desideratum there is a logic that satisfies the remaining desiderata and faithfully models the gentle murder scenario.

2 The Gentle Murder Scenario

In stead of giving a general form for all CTD obligations and then checking whether the general form can be represented in a logic we will consider a specific CTD obligation. If a logic is incapable of faithfully representing the specific CTD obligation it is certainly incapable of representing CTD obligations in general.

The CTD obligation we consider is due to [10]. Consider the following situation: at some point in time you have the choice whether or not to murder. If you do murder you simultaneously have the choice whether to murder gently or un-gently. I hope we can all agree that you have an obligation not to murder. Let us write m for 'you murder', then we can represent this obligation as $O(\neg m)$.[1] If you would decide to murder anyway then you are doing something wrong, but you can slightly mitigate your action by murdering gently; you have a CTD obligation to murder gently if you murder. Let us write g for 'you murder gently'

[1] Here I let obligations apply to actions, such as murdering. But I could equivalently let obligations apply to states of affairs, such as the one where someone is murdered.

and $O(g|m)$ for the CTD obligation to murder gently. I refer to this situation as the gentle murder scenario.[2]

If the scenario included only these two obligations it would be trivially solvable, for example by saying that everything is obligatory. In order to faithfully model the scenario a few more statements must hold: a gentle murder is still a very bad thing so there must be an obligation $O(\neg g)$, there is no obligation to murder so $\neg O(m)$ and there is no CTD obligation to murder un-gently if you murder so $\neg O(\neg g|m)$. Let Ψ be the set of statements that should hold in the gentle murder scenario, $\Psi = \{O(\neg m), \neg O(m), O(\neg g), O(g|m), \neg O(\neg g|m)\}$.[3]

2.1 Problems with some existing formalizations of CTD obligations

As mentioned in the introduction three attempts to formalize CTD obligations using sanction based semantics are the logics SA [5], DMAL [6] and $\text{PD}_\text{e}\text{L}$ [9]. These three logics are especially instructive because they each suffer from a different problem in formalizing CTD obligations.

$\text{PD}_\text{e}\text{L}$ cannot represent the gentle murder scenario at all. A CTD obligation $O(g|m)$ can only be represented in $\text{PD}_\text{e}\text{L}$ as "if you m then you should subsequently g", but murdering and murdering gently do not happen in sequence.

In SA an obligation $O(\neg m)$ implies that $\mathbb{S}$ holds in both the $m \wedge \neg g$ and $m \wedge g$ cases. This leaves no way to deontically distinguish $m \wedge \neg g$ and $m \wedge g$ so a CTD obligation $O(g|m)$ usually[4] cannot occur without an opposite CTD obligation $O(\neg g|m)$.

In DMAL a CTD obligation $O(g|m)$ implies that $\mathbb{S}$ does not hold after either $\neg m$ or g. This leaves no way to deontically distinguish $\neg m$ and g. In particular, the obligation $O(g|m)$ implies a permission $P(g)$ to murder gently.

3 The Semantic Approach

The approach to logic taken here is a semantic one and more precisely a possible world semantics one. For this purpose the following definition will suffice.

Definition 1. *A* logic $\mathcal{L}$ *is a triple* $\mathcal{L} = (\mathfrak{M}, \Phi, \models)$ *where* $\mathfrak{M}$ *is a class of models,* Φ *is a set of formulas and* $\models$ *is a satisfaction relation such that*

- *there are a countable subset* $P \subseteq \Phi$ *of propositional variables and one designated variable* $\mathbb{S} \in P$*,*
- Φ *is closed under the unary operator* $\neg$ *and the binary operator* $\rightarrow$*,*

[2] The situation in combination with a few other statements is usually referred to as the paradox of the gentle murderer or Forrester's paradox. The paradox however depends on a factual statement that m holds. I make no such assumption here so there is no paradox.

[3] Note that there is no requirement for $\neg O(g)$ to hold. This leaves open the possibility for the logic to satisfy some form of detachment from $O(g|m)$.

[4] The unusual case is if it is impossible to murder gently.

- *every* $\mathcal{M} \in \mathfrak{M}$ *is a triple* $\mathcal{M} = (W, R, v)$ *where* W *is a set of possible worlds,* R *is a set and* $v : P \to \wp(W)$ *is a valuation function,*
- $\models$ *is a relation between model-world pairs* $(\mathcal{M}, w)$ *and formulas* ϕ*, where* $\mathcal{M} = (W, R, v) \in \mathfrak{M}$*,* $w \in W$ *and* $\phi \in \Phi$*,*
- *for every* $\mathcal{M}$*,* w *and* $p \in P$ *it holds that* $\mathcal{M}, w \models p$ *iff* $w \in v(p)$*,*
- *for every* $\mathcal{M}$*,* w *and* $\phi \in \Phi$ *it holds that* $\mathcal{M}, w \models \neg\phi$ *iff* $\mathcal{M}, w \not\models \phi$*.*
- *for every* $\mathcal{M}$*,* w *and* $\phi_1, \phi_2 \in \Phi$ *it holds that if* $\mathcal{M}, w \models \phi_1 \to \phi_2$ *and* $\mathcal{M}, w \models \phi_1$ *then* $\mathcal{M}, w \models \phi_2$*.*

The set R is left unspecified and can be used to encode additional structure such as an accessibility relation or a set of agents. The exact contents of R are irrelevant for present purposes. We say that ϕ *holds* or *is true* in w on $\mathcal{M}$ if $\mathcal{M}, w \models \phi$ and use a few common notations.

Definition 2. *For any* $\mathcal{M} = (W, R, V) \in \mathfrak{M}$*,* $w \in W$*,* $\phi \in \Phi$ *and* $\Gamma \subseteq \Phi$*:*

- *The set* $[\![\phi]\!]_{\mathcal{M}}$ *is given by* $[\![\phi]\!]_{\mathcal{M}} := \{w \in W \mid \mathcal{M}, w \models \phi\}$*.*
- Γ *holds in* w *on* $\mathcal{M}$*, denoted* $\mathcal{M}, w \models \Gamma$ *if* $\mathcal{M}, w \models \phi$ *for all* $\phi \in \Gamma$*.*
- *The set* $[\![\Gamma]\!]_{\mathcal{M}}$ *is given by* $[\![\Gamma]\!]_{\mathcal{M}} := \{w \in W \mid \mathcal{M}, w \models \Gamma\}$*.*
- ϕ *is* valid on $\mathcal{M}$*, denoted* $\mathcal{M} \models \phi$*, if* $\mathcal{M}, w \models \phi$ *for all* $w \in W$*.*

Now let us consider what it means for a logic $\mathcal{L} = (\mathfrak{M}, \Phi, \models)$ to be capable of modeling the gentle murder scenario. Firstly the logic should be capable of representing the obligations under consideration. I formalize this as $\Psi \subseteq \Phi$, but the formulas in Ψ may be considered as abbreviations for other formulas. For example, we could use $O(g|m)$ as an abbreviation for $O(m \to g)$ or $m \to O(g)$. Secondly, there should be a model $\mathcal{M}^O = (W^O, R^O, v^O) \in \mathfrak{M}$ representing the scenario and a world $w^O \in W^O$ where the obligations hold, so $\mathcal{M}^O, w^O \models \Psi$. Furthermore, the propositional variables m and g should represent murder and gentle murder respectively in this model $\mathcal{M}^O$. We cannot enforce such a meaning, except for the part that a gentle murder is still a murder so $\mathcal{M}^O \models g \to m$.

However, if the logic has this one 'canonical' representation of the gentle murder scenario then it also has other representations of the scenario. We could for example write n instead of m for 'you murder' and still have a representation. Each such representation can be seen as a tuple $(\mathcal{M}, w, \Gamma, \chi)$ where $\mathcal{M}$ is a model, w is a world in the model, Γ is a set of formulas corresponding to Ψ and χ is a formula corresponding to $g \to m$ such that $\mathcal{M}, w \models \Gamma$ and $\mathcal{M} \models \chi$.

We will need to consider both the class G of such tuples and the 'canonical' representation $(\mathcal{M}^O, w^O, \Psi, g \to m) \in G$.

Definition 3. *Let* $\mathcal{L}$ *be a logic,* $\mathcal{M}^O$ *a model,* w^O *a world in the model and* G *a class of tuples* $(\mathcal{M}, w, \Gamma, \psi)$ *where* $\mathcal{M} = (W, R, v) \in \mathfrak{M}$*,* $w \in W$*,* $\Gamma \subseteq \Phi$*,* $\psi \in \Phi$ *and* $\mathcal{M} \models \psi$*. Then*

- *the tuple* $(\mathcal{L}, \mathcal{M}^O, w^O, G)$ models the gentle murder scenario *if* $(\mathcal{M}^O, w^O, \Psi, g \to m) \in G$ *and*
- *the tuple* $(\mathcal{L}, \mathcal{M}^O, w^O, G)$ faithfully models the gentle murder scenario *if it models the gentle murder scenario and furthermore* $\mathcal{M}, w \models \Gamma$ *for every* $(\mathcal{M}, w, \Gamma, \psi) \in G$*.*

Definition 4. *A logic* $\mathcal{L}$ (faithfully) models the gentle murder scenario *if there are* $\mathcal{M}^O, w^O$ *and* G *such that* $(\mathcal{L}, \mathcal{M}^O, w^O, G)$ *(faithfully) models the gentle murder scenario.*

4 Desiderata

In order for a logic to represent CTD obligations with sanction semantics it should have certain properties, given here as desiderata. The desiderata should hold in general, but in some cases formally defining the desideratum in general is very hard. In order to give a reasonably simple definition I therefore restrict some of the desiderata to the representations G of the gentle murder scenario or even the canonical representation $(\mathcal{M}^O, w^O, \Psi, g \to m)$ of the gentle murder scenario. If it is not possible to faithfully model the gentle murder scenario while satisfying the restricted form of the desiderata it is also impossible to do so while satisfying the general form of the desiderata.

Invariance under propositional renaming. We use the propositional variables m and g for the statements 'you murder' and 'you murder gently'. But of course we could use any other two variables. If not committing murder is obligatory it should remain obligatory if we use the variable r instead of m for 'you murder'. More generally, renaming any propositional variable other than the designated variable $\mathbb{S}$ should not change the truth value of any formula as long as the appropriate substitution is applied to the formula.
Likewise, replacing a propositional variable other than $\mathbb{S}$ by its negation should not change anything. If we use m for 'you murder' there is an obligation $O(\neg m)$ not to murder. Then if we instead use m for 'you do not murder' (and therefore $\neg m$ for 'you murder') there should still be an obligation not to murder, although it is then denoted $O(m)$.
In order to formalize this let us first introduce a notation $[p/q]$ and $[p/\neg p]$ for renaming p to q or to $\neg p$ respectively.

Definition 5. *For* $\mathcal{M} = (W, R, v) \in \mathfrak{M}$ *and* $p, q \in P$ *define* $v[p/q]$ *and* $v[p/\neg p]$ *by*

$$v[p/q](r) := \begin{cases} v(q) \text{ if } r = p \\ v(p) \text{ if } r = q \\ v(r) \text{ otherwise} \end{cases}$$

$$v[p/\neg p](r) := \begin{cases} v(r) & \text{if } r \neq p \\ W \setminus v(p) & \text{if } r = p \end{cases}$$

and $\mathcal{M}[p/\phi]$ *by*

$$\mathcal{M}[p/\phi] := (W, R, v[p/\phi])$$

for $\phi \in \{q, \neg p\}$. *Furthermore, for* $\phi \in \Phi$ *define* $\phi[p/q]$ *to be the formula obtained by simultaneously replacing all occurrences of* p *in* ϕ *by* q *and all occurrences of* q *by* p *and define* $\phi[p/\neg p]$ *to be the formula obtained by replacing all occurrences of* p *in* ϕ *by* $\neg p$.

Using this notation we can easily give a formalization of the desideratum.

Desideratum 1 (Invariance under propositional renaming). *For any* $\mathcal{M} = (W, R, v) \in \mathfrak{M}$*, any* $w \in W$*, any* $\phi \in \Phi$ *and any* $p, q \in P \setminus \{\mathbb{S}\}$ *it holds that*

$$\mathcal{M}, w \models \phi \Leftrightarrow \mathcal{M}[p/q], w \models \phi[p/q] \text{ and}$$

$$\mathcal{M}, w \models \phi \Leftrightarrow \mathcal{M}[p/\neg p], w \models \phi[p/\neg p]$$

Note that since $\models$ is a relation between pairs $(\mathcal{M}, w)$ where $\mathcal{M} \in \mathfrak{M}$ and formulas $\phi \in \Phi$ the desideratum implies that $\mathcal{M}[p/q]$, $\mathcal{M}[p/\neg p] \in \mathfrak{M}$ and $\phi[p/q], \phi[p/\neg p] \in \Phi$. Similar claims are implicit in the other desiderata.

Determinacy of sanction. The sanction $\mathbb{S}$ should represent the presence or absence of wrongdoing. Whether there is wrongdoing in a world should be fully determined by the truth values of the deontically relevant formulas in that world. What exactly the deontically relevant formulas are is determined by what one is modeling.

In the gentle murder scenario the formulas m and g are obviously deontically relevant, but one could argue that there are other relevant formulas such as for example a formula c corresponding to 'you covet your neighbor's house'. Such a formula c can undeniably have deontic relevance in some systems of rules. It should however be possible to represent the rule system that is described in the gentle murder scenario, which only has rules about murdering and murdering gently.

The rule system that only contains rules about murdering and murdering gently can be seen as a canonical rule system for the gentle murder scenario, so it seems reasonable to require the model $\mathcal{M}^O$ to correspond to this particular system. The desideratum thus becomes a requirement that the value of $\mathbb{S}$ in a world of $\mathcal{M}^O$ is fully determined by the values m and g on that world.

Desideratum 2 (Determinacy of sanction). *For every* $w_1, w_2 \in W^O$ *such that* $w_1 \in v^O(m) \Leftrightarrow w_2 \in v^O(m)$ *and* $w_1 \in v^O(g) \Leftrightarrow w_2 \in v^O(g)$ *it holds that* $w_1 \in v^O(\mathbb{S}) \Leftrightarrow w_2 \in v^O(\mathbb{S})$.

Range of outcomes. When considering a CTD obligation there is a number of possible outcomes. The gentle murder scenario for example has three outcomes: an un-gentle murder (m and $\neg g$), a gentle murder (m and g) and no murder ($\neg m$ and $\neg g$). The possible outcomes should be represented in the model of the scenario.

We could require the model to have exactly one world for each outcome, but that seems too strong a requirement. Consider for example the use of a new variable r for 'it is raining'. The combination of m, g and r cannot occur in the same world as m, g and $\neg r$ but they are part of the same outcome, the gentle murder. A better requirement is therefore that there is at least one world in each outcome. For the gentle murder scenario this gives the following desideratum.

Desideratum 3 (Range of outcomes). *There exist* $w_1, w_2, w_3 \in W^O$ *such that* $\mathcal{M}^O, w_1 \models \{m, \neg g\}$, $\mathcal{M}^O, w_2 \models \{m, g\}$ *and* $\mathcal{M}^O, w_3 \models \{\neg m, \neg g\}$.

Invariance under act renaming. We use the propositional variables m and g for 'you murder' and 'you murder gently'. But we could describe the same situation using different acts, for example by writing m for 'you murder' and u for 'you murder *un*-gently'. Changing the names of acts in such a way does not change the situation that is described, so the logic should be insensitive to such renaming.

It is important to note that act renaming is not the same as propositional renaming, $\neg u$ is not equivalent to g as g is necessarily a murder while $\neg u$ need not be. It is not clear whether under such conditions it should in general hold that a conditional obligation $O(g|m)$ implies a conditional obligation $O(\neg u|m)$. However, in this particular case it is clear that a conditional obligation $O(\neg u|m)$ should hold; if you murder you have an obligation not to do so un-gently. Similarly there is no obligation to murder un-gently if you murder $\neg O(u|m)$. The obligation $O(\neg g)$ not to murder gently does not however change into an obligation $O(u)$ to murder un-gently, but into an obligation $O(\neg u)$ not to murder un-gently.

This kind of act renaming can also be done without changing the variable used. If g represents murdering gently and we want to change it to representing murdering un-gently we should change the value of g where m holds, but not where $\neg m$ holds. Let us denote by $v[p/\neg p|\phi]$ the valuation obtained by changing the value of p on $[\![\phi]\!]_{\mathcal{M}}$ worlds while keeping it constant on $[\![\neg\phi]\!]_{\mathcal{M}}$.

Definition 6. *For* $\mathcal{M} = (W, R, v)$, $p \in P$ *and* $\phi \in \Phi$ *define* $v[p/\neg p|\phi]$ *by*

$$v[p/\neg p|\phi](r) := \begin{cases} v(r) & \text{if } r \neq p \\ ((W \setminus [\![\phi]\!]_{\mathcal{M}}) \cap v(p)) \cup ([\![\phi]\!]_{\mathcal{M}} \cap (W \setminus v(p))) & \text{if } r = p \end{cases}$$

Furthermore, define $\mathcal{M}[p/\neg p|\phi]$ *by* $\mathcal{M}[p/\neg p|\phi] := (W, R, v[p/\neg p|\phi])$.

The desideratum can then be given as follows.

Desideratum 4 (Invariance under act renaming). *Let*

$$\Psi[g/\neg g|m] = \{O(\neg m), \neg O(m), O(\neg g), O(\neg g|m), \neg O(g|m)\}.$$

Then for any $\mathcal{M}$ *and* w *such that* $(\mathcal{M}, w, \Psi, g \to m) \in G$ *it holds that* $(\mathcal{M}[g/\neg g|m], w, \Psi[g/\neg g|m], g \to m) \in G$.

Invariance under outcome renaming. When using sanction semantics the moral status of an outcome depends only on the value of $\mathbb{S}$ in the outcome as opposed to for example a preference order between the outcomes. As such, the outcomes should "be treated the same way" when determining the relevant obligations. One way of stating this is that if we interchange the values of the propositional variables on different outcomes this should have no influence on the obligations in effect.

Unfortunately, there is a problem with interchanging the values on different outcomes. Since different outcomes may contain different numbers of

worlds it can be impossible to completely interchange them. The values of the relevant variables are however constant in a given outcome, so we can interchange the values of the relevant variables. This may result in the change of some morally irrelevant facts in the outcomes, but this doesn't matter as these facts are morally irrelevant.

Definition 7. *For $\mathcal{M} = (W, R, v)$ and $\Gamma \subseteq \{m, g, \mathbb{S}\}$ define $W_\Gamma^\mathcal{M}$ to be the set of worlds in which the variables in Γ are true and those in $\{m, g, \mathbb{S}\} \setminus \Gamma$ are false,*

$$W_\Gamma^\mathcal{M} := \{w \in W \mid \forall p \in \Gamma : w \in v(p) \text{ and } \forall p \in \{m, g, \mathbb{S}\} \setminus \Gamma : w \notin v(p)\}.$$

Furthermore, for $\Gamma, \Theta \subseteq \{m, g, \mathbb{S}\}$ define $v[W_\Gamma^\mathcal{M}/W_\Theta^\mathcal{M}]$ to be the valuation obtained from v by interchanging the valuations of m, g and $\mathbb{S}$ on the $W_\Gamma^\mathcal{M}$ and $W_\Theta^\mathcal{M}$ worlds,

$$v[W_\Gamma^\mathcal{M}/W_\Theta^\mathcal{M}](p) := \begin{cases} v(p) & \text{if } p \notin \Gamma \text{ and } p \notin \Theta \\ (v(p) \cup W_\Gamma^\mathcal{M}) \setminus W_\Theta^\mathcal{M} & \text{if } p \notin \Gamma \text{ and } p \in \Theta \\ (v(p) \cup W_\Theta^\mathcal{M}) \setminus W_\Gamma^\mathcal{M} & \text{if } p \in \Gamma \text{ and } p \notin \Theta \\ v(p) & \text{if } p \in \Gamma \text{ and } p \in \Theta \end{cases}$$

and

$$\mathcal{M}[W_\Gamma^\mathcal{M}/W_\Theta^\mathcal{M}] := (W, R, v[W_\Gamma^\mathcal{M}/W_\Theta^\mathcal{M}])$$

Now we can formalize the desideratum for the gentle murder scenario.

Desideratum 5 (Invariance under outcome renaming). *For any $\Gamma, \Theta \subseteq \{m, g, \mathbb{S}\}$, $\mathcal{M} = (W, R, v)$ and w such that $(\mathcal{M}, w, \Psi, g \to m)$, $W_\Gamma^\mathcal{M} \neq \emptyset$ and $W_\Theta^\mathcal{M} \neq \emptyset$ it holds that $(\mathcal{M}[W_\Gamma^\mathcal{M}/W_\Theta^\mathcal{M}], w, \Psi, g \to m) \in G$.*

These desiderata are rather weak, so complicated and 'strange' semantics are allowed as long as they are based on the use of a sanction. Logics that satisfy the desiderata (for an appropriate choice of G, $\mathcal{M}^O$ and w^O) include the possible world semantics for SDL as well as the systems presented in for example [5, 7].

The main weakness of the desiderata is that they do not apply to dynamic deontic logics where m and g would be labels of transitions between possible worlds as opposed to propositional variables, such as the logics described in [6, 9]. This is mostly a matter of notation, the desiderata could be rephrased to apply to dynamic deontic logics and an impossibility result similar to the one obtained with the current desiderata could be reached. Including the dynamic version of the desiderata would however greatly complicate the notation of the desiderata without significant conceptual changes so I do not do so.

5 Impossibility Result

Theorem 1. *There are no logic $(\mathfrak{M}, \Phi, \models)$, class G of tuples, model $\mathcal{M}^O = (W^O, R^O, v^O) \in \mathfrak{M}$ and world $w^O \in W^O$ that satisfy desiderata 1 to 5 and faithfully model the gentle murder scenario.*

Proof. Suppose that there are such logic $(\mathfrak{M}, \Phi, \models)$, class G of tuples, model $\mathcal{M}^O = (W^O, R^O, v^O) \in \mathfrak{M}$ and world $w^O \in W^O$.

Let $X_1 = [\![\{m, \neg g\}]\!]_{\mathcal{M}^O}$, $X_2 = [\![\{m, g\}]\!]_{\mathcal{M}^O}$ and $X_3 = [\![\{\neg m, \neg g\}]\!]_{\mathcal{M}^O}$. We have $X_1 \cup X_2 \cup X_3 = W^O$ because $(\mathcal{M}^O, w^O, \Psi, g \to m) \in G$ and therefore $\mathcal{M}^O \models g \to m$. Furthermore, as a consequence of the **Range of outcomes** desideratum X_1, X_2 and X_3 are nonempty and by the **Determinacy of sanction** desideratum the value of $\mathbb{S}$ is constant inside each of the three sets, so X_1, X_2 and X_3 are of the form $W_\Gamma^{\mathcal{M}^O}$ for some $\Gamma \subseteq \{m, g, \mathbb{S}\}$, see Definition 7.

Suppose X_1 and X_2 have the same value for $\mathbb{S}$. Then interchanging the valuations of X_1 and X_2 results in the same model as changing the meaning of g to 'you murder un-gently'. That is, $\mathcal{M}^O[X_1/X_2] = \mathcal{M}^O[g/\neg g|m]$.

Example 1. If X_1 and X_2 have the same value for $\mathbb{S}$ it does not matter where $\mathbb{S}$ holds. In order to illustrate why $\mathcal{M}^O[X_1/X_2] = \mathcal{M}^O[g/\neg g|m]$ it is however convenient to take a concrete example, so consider the case where $\mathbb{S}$ holds on X_1 and X_2 but not on X_3, see Fig. 1. Then $v^O(m) = v^O(\mathbb{S}) = X_1 \cup X_2$ and $v^O(g) = X_2$. If we switch the valuations of X_1 and X_2 we get $v^O[X_1/X_2](m) = v^O[X_1/X_2](\mathbb{S}) = X_1 \cup X_2$ and $v^O[X_1/X_2](g) = X_1$.

If we change the meaning of g to murdering un-gently we get $v^O[g/\neg g|m](m) = v^O[g/\neg g|m](\mathbb{S}) = X_1 \cup X_2$ and $v^O[g/\neg g|m](g) = X_1$. So we have $v^O[X_1/X_2] = v[g/\neg g|m]$ and therefore $\mathcal{M}^O[X_1/X_2] = \mathcal{M}^O[g/\neg g|m]$.

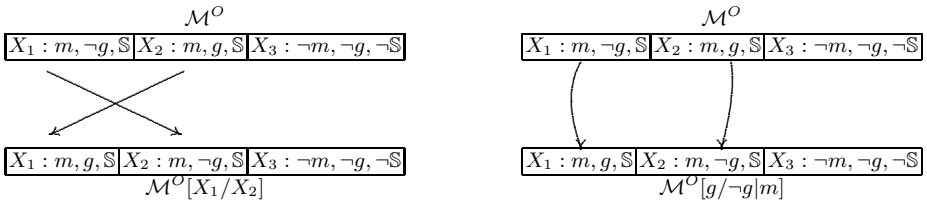

Fig. 1. If X_1 and X_2 have the same value for $\mathbb{S}$ then interchanging X_1 and X_2 has the same result as changing the meaning of g to 'you murder un-gently'.

By the **Invariance under outcome renaming** desideratum it holds that $(\mathcal{M}^O[X_1/X_2], w^O, \Psi, g \to m) \in G$. The gentle murder scenario is faithfully modeled, so $\mathcal{M}^O[X_1/X_2], w^O \models \Psi$ and in particular $\mathcal{M}^O[X_1/X_2], w^O \not\models O(\neg g|m)$. By **Invariance under act renaming** it also holds that $(\mathcal{M}^O[g/\neg g|m], w^O, \Psi[g/\neg g|m], g \to m) \in G$ so $\mathcal{M}^O[g/\neg g|m], w^O \models \Psi[g/\neg g|m]$ by the faithful modeling and in particular $\mathcal{M}^O[g/\neg g|m], w^O \models O(\neg g|m)$. But $\mathcal{M}[X_1/X_2] = \mathcal{M}[g/\neg g|m]$ and the formula cannot be both true and false. The assumption that X_1 and X_2 have the same value for $\mathbb{S}$ is therefore false.

Now suppose that X_1 and X_3 have the same value for $\mathbb{S}$. Then first interchanging the valuations of X_1 and X_3 and subsequently changing the meaning of g to murdering un-gently results in the same model as first changing the meaning of g to murdering un-gently, then renaming both m and g to their negations

and finally renaming m and g to each other. That is, $\mathcal{M}^O[X_1/X_3][g/\neg g|m] = \mathcal{M}^O[g/\neg g|m][g/\neg g][m/\neg m][g/m]$.

Example 2. As another concrete example consider the case where $\mathbb{S}$ holds on X_1 and X_3 but not on X_2, see Fig. 2. Then $v^O(m) = X_1 \cup X_2$, $v^O(g) = X_2$ and $v^O(\mathbb{S}) = X_1 \cup X_3$. Interchanging the valuations for X_1 and X_3 we get

$$v^O(p)[X_1/X_3] = \begin{cases} X_2 \cup X_3 & \text{if } p = m \\ X_2 & \text{if } p = g \\ X_1 \cup X_3 & \text{if } p = \mathbb{S} \\ v^O(p) & \text{otherwise.} \end{cases}$$

If we subsequently change the meaning of g to murdering un-gently we get

$$v^O(p)[X_1/X_3][g/\neg g|m] = \begin{cases} X_2 \cup X_3 & \text{if } p = m \\ X_3 & \text{if } p = g \\ X_1 \cup X_3 & \text{if } p = \mathbb{S} \\ v^O(p) & \text{otherwise.} \end{cases}$$

If we start at v^O and change the meaning of g to murdering un-gently we get

$$v^O[g/\neg g|m](p) = \begin{cases} X_1 \cup X_2 & \text{if } p = m \\ X_1 & \text{if } p = g \\ X_1 \cup X_3 & \text{if } p = \mathbb{S} \\ v^O(p) & \text{otherwise.} \end{cases}$$

If we then rename m and g to their negations we get

$$v^O[g/\neg g|m][g/\neg g][m/\neg m](p) = \begin{cases} X_3 & \text{if } p = m \\ X_2 \cup X_3 & \text{if } p = g \\ X_1 \cup X_3 & \text{if } p = \mathbb{S} \\ v^O(p) & \text{otherwise.} \end{cases}$$

Subsequently renaming m and g to each other gives

$$v^O[g/\neg g|m][g/\neg g][m/\neg m][g/m](p) = \begin{cases} X_2 \cup X_3 & \text{if } p = m \\ X_3 & \text{if } p = g \\ X_1 \cup X_3 & \text{if } p = \mathbb{S} \\ v^O(p) & \text{otherwise,} \end{cases}$$

so $v^O[X_1/X_3][g/\neg g|m] = v^O[g/\neg g|m][g/\neg g][m/\neg m][g/m]$.

By the **Invariance under outcome renaming** desideratum it holds that $(\mathcal{M}^O[X_1/X_3], w^O, \Psi, g \to m) \in G$, and then by the **Invariance under act renaming** desideratum that $(\mathcal{M}^O[X_1/X_3][g/\neg g|m], w^O, \Psi[g/\neg g|m], g \to m) \in G$. This implies that $\mathcal{M}^O[X_1/X_3][g/\neg g|m], w^O \models \Psi[g/\neg g|m]$ so in particular $\mathcal{M}^O[X_1/X_3][g/\neg g|m], w^O \models O(\neg g)$.

However, by the **Invariance under act renaming** desideratum it also holds that $(\mathcal{M}^O[g/\neg g|m], w^O, \Psi[g/\neg g|m], g \to m) \in G$. So $\mathcal{M}^O[g/\neg g|m], w^O \models$

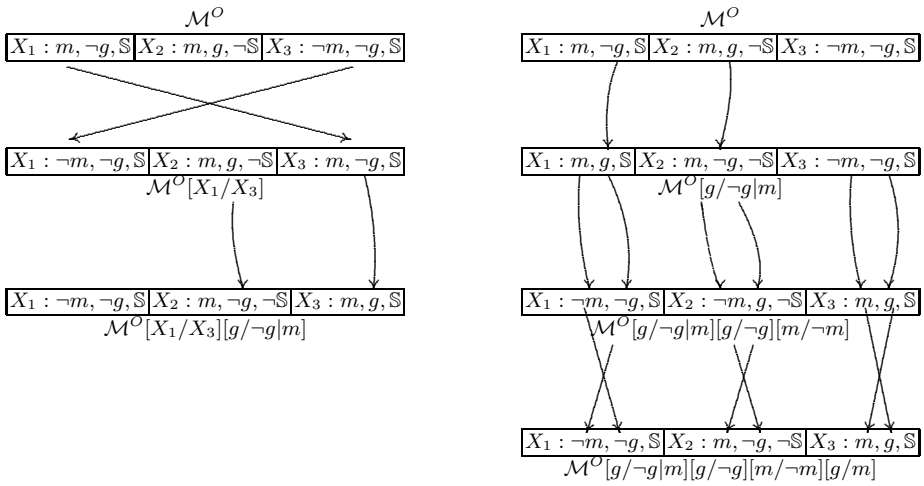

Fig. 2. If X_1 and X_3 have the same value for $\mathbb{S}$ then interchanging X_1 and X_3 followed by changing the meaning of g to murdering un-gently has the same result as first changing the meaning of g to murdering un-gently, then renaming m and g to their negations and finally renaming m and g to each other.

$\Psi[g/\neg g|m]$ and in particular $\mathcal{M}^O[g/\neg g|m], w^O \not\models O(m)$. By repeated application of the **Invariance under propositional renaming** desideratum it can then be seen that $\mathcal{M}^O[g/\neg g|m][g/\neg g][m/\neg m][g/m], w^O \not\models O(m)[g/\neg g][m/\neg m][g/m]$ so $\mathcal{M}^O[g/\neg g|m][g/\neg g][m/\neg m][g/m], w^O \not\models O(\neg g)$.

But $\mathcal{M}^O[X_1/X_3][g/\neg g|m] = \mathcal{M}^O[g/\neg g|m][g/\neg g][m/\neg m][g/m]$, and $O(\neg g)$ cannot be both true and false. The assumption that X_1 and X_3 have the same value for $\mathbb{S}$ must therefore be false.

Finally, suppose X_2 and X_3 have the same value for $\mathbb{S}$. Then first interchanging X_2 and X_3 and then renaming m and g to each other has the same result as renaming m and g to their negations. That is, $\mathcal{M}^O[X_2/X_3][m/g] = \mathcal{M}^O[g/\neg g][m/\neg m]$.

Example 3. As a concrete example consider the case where $\mathbb{S}$ holds on X_2 and X_3 but not on X_1, see Fig. 3. Then $v^O(m) = X_1 \cup X_2$, $v^O(g) = X_2$ and $v^O(\mathbb{S}) = X_2 \cup X_3$. Interchanging X_2 and X_3 we get

$$v^O[X_2/X_3](p) = \begin{cases} X_1 \cup X_3 & \text{if } p = m \\ X_3 & \text{if } p = g \\ X_2 \cup X_3 & \text{if } p = \mathbb{S} \\ v^O(p) & \text{otherwise.} \end{cases}$$

Subsequently renaming m and g to each other gives

$$v^O[X_2/X_3][m/g](p) = \begin{cases} X_3 & \text{if } p = m \\ X_1 \cup X_3 & \text{if } p = g \\ X_2 \cup X_3 & \text{if } p = \mathbb{S} \\ v^O(p) & \text{otherwise.} \end{cases}$$

If on the other hand we start at v^O and rename m and g to their negations we get

$$v^O[g/\neg g][m/\neg m](p) = \begin{cases} X_3 & \text{if } p = m \\ X_1 \cup X_3 & \text{if } p = g \\ X_2 \cup X_3 & \text{if } p = \mathbb{S} \\ v^O(p) & \text{otherwise,} \end{cases}$$

so $v^O[X_2/X_3][m/g] = v^O[g/\neg g][m/\neg m]$.

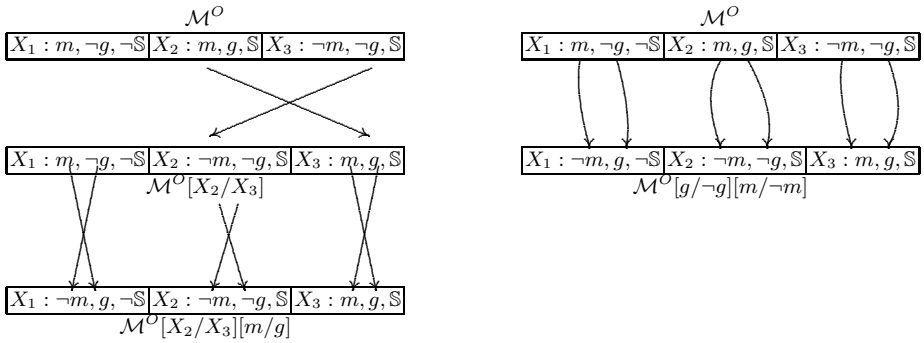

Fig. 3. If X_2 and X_3 have the same value for $\mathbb{S}$ then interchanging X_1 and X_3 followed by renaming m and g to each other has the same result as renaming m and g to their negations.

By the **Invariance under outcome renaming** desideratum it holds that $(\mathcal{M}^O[X_2/X_3], w^O, \Psi, m \to g) \in G$. This implies that $\mathcal{M}^O[X_2/X_3], w^O \models \Psi$ so in particular $\mathcal{M}^O[X_2/X_3], w^O \models O(\neg g)$. By the **Invariance under propositional renaming** desideratum it then holds that $\mathcal{M}^O[X_2/X_3][m/g], w^O \models O(\neg g)[m/g]$ so $\mathcal{M}^O[X_2/X_3][m/g], w^O \models O(\neg m)$.

However, $(\mathcal{M}^O, w^O, \Psi, m \to g) \in G$ so $\mathcal{M}^O, w^O \models \Psi$ and in particular $\mathcal{M}^O, w^O \not\models O(m)$. This implies $\mathcal{M}^O[g/\neg g][m/\neg m], w^O \not\models O(m)[g/\neg g][m/\neg m]$ by **Invariance under propositional renaming**, so $\mathcal{M}^O[g/\neg g][m/\neg m], w^O \not\models O(\neg m)$.

But $\mathcal{M}^O[X_2/X_3][m/g] = \mathcal{M}^O[g/\neg g][m/\neg m]$ and $O(\neg m)$ cannot be both true and false. The assumption that X_2 and X_3 have the same value for $\mathbb{S}$ must therefore be false.

We have obtained the results that X_1 and X_2 cannot have the same value for $\mathbb{S}$, that X_1 and X_3 cannot have the same value for $\mathbb{S}$ and that X_2 and X_3 cannot have the same value for $\mathbb{S}$. This cannot happen since there are only two possible values for $\mathbb{S}$. The assumption that there are a logic $(\mathfrak{M}, \Phi, \models)$, class G of tuples, model $\mathcal{M}^O = (W^O, R^O, v^O) \in \mathfrak{M}$ and world $w^O \in W^O$ such that $(\mathcal{M}^O, w^O, \Psi, g \to m) \in G$ and desiderata 1 to 5 are satisfied must therefore be false, which proves the theorem. □

6 Relaxing the Desiderata

Having established that we cannot find semantics that faithfully model the gentle murder scenario and satisfy all the desiderata it seems worthwhile to consider what happens if we drop one of the desiderata. Dropping any of the desiderata allows us to faithfully model the gentle murder scenario while satisfying the remaining desiderata, although for most of the desiderata the logic in question is not very useful.

6.1 Dropping Invariance under Propositional Renaming

If we drop the **Invariance under propositional renaming** desideratum we can model the gentle murder scenario by giving special treatment to m, letting $O(\neg m)$ and $\neg O(m)$ always be true and using $\mathbb{S}$ only to determine the moral value of g. The semantics for $O(\phi|m)$ could then for example be $\mathcal{M}, w \models O(\phi|m) \Leftrightarrow \mathcal{M} \models \neg\phi \rightarrow \mathbb{S}$. Such a logic does not seem very useful, however.

6.2 Dropping Determinacy of Sanction

If we drop the **Determinacy of sanction** desideratum we can model the gentle murder scenario by having $\mathbb{S}$ true on all m and $\neg g$ worlds, false on all $\neg m$ worlds and true on some but not all m and g worlds. Effectively this creates a third degree of badness in between 'always $\mathbb{S}$' and 'never $\mathbb{S}$'.

This solution does not however generalize to situations where more than three degrees of badness are needed such as situations with multiple mitigating factors. In order to create more than three degrees of badness we would have to give relevance to exactly how often $\mathbb{S}$ holds. But because of the **Invariance under outcome renaming** desideratum we can interchange any number of worlds of one outcome with any number of worlds of another outcome so the exact number of worlds in a given outcome cannot be relevant.

6.3 Dropping Range of Outcomes

Dropping the **Range of outcomes** desideratum allows us to model the gentle murder scenario, by using nonexistence of an outcome as a heavier sanction than $\mathbb{S}$. This leads to a model with two worlds, one with m, g and $\mathbb{S}$ and one with $\neg m, \neg g$ and $\neg\mathbb{S}$, with semantics given by $\mathcal{M}, w \models O(\phi) \Leftrightarrow \mathcal{M} \models \neg\phi \rightarrow \mathbb{S}$ and $\mathcal{M}, w \models O(\psi|\phi) \Leftrightarrow (\mathcal{M}, w \models O(\neg\phi)$ and $\mathcal{M} \not\models \neg(\psi \wedge \neg\phi))$.

Under these semantics all the obligations from Ψ are satisfied, but also some obligations one would prefer not to have in a model of the gentle murder scenario such as $O(m|g)$. The method also doesn't generalize well to more complicated contrary-to-duty obligations.

6.4 Dropping Invariance under Act Renaming

If we drop the **Invariance under act renaming** desideratum we can give deontic relevance to whether we discuss murdering gently or murdering un-gently.

Whether we use g for murdering gently or for murdering un-gently, $\neg g$ will hold in the $\neg m$ worlds. This $\neg g$ can then be set as the default action, which we can consider either a *pessimistic* default or an *optimistic* default. If it is a pessimistic default the contrary-to-duty obligation when murdering is to make the default false, if it is an optimistic default the contrary-to-duty obligation is to make the default true. The semantics of the pessimistic default could for example be given by $\mathcal{M}, w \models O(\phi|\psi) \Leftrightarrow \mathcal{M} \models O(\neg\psi) \wedge (\neg\psi \rightarrow \neg\phi)$.

This method of setting defaults allows us to faithfully model the gentle murder scenario and certain generalizations of it, but not every CTD obligation.

6.5 Dropping Invariance under Outcome Renaming

If we drop the **Invariance under outcome renaming** desideratum we can simply use the additional structure R of a model $\mathcal{M} = (W, R, v)$ to encode obligations, for example by letting R be a partial order on the worlds.

Contrary-to-duty obligations on models where a preference between the possible worlds is given by a partial order are well studied, see for example [11] for an overview.

6.6 Using multiple sanctions

One more way to formalize the gentle murder scenario is to use multiple sanctions $\mathbb{S}_1, \mathbb{S}_2, \cdots$. This would require modifications to the **Invariance under propositional renaming**, **Determinacy of sanction** and **Invariance under outcome renaming** desiderata, as the special status of $\mathbb{S}$ would have to be extended to all sanctions.

Any number greater than 1 of sanctions would allow us to faithfully represent the gentle murder scenario while satisfying all (modified) desiderata. A finite number of sanctions can only model a finite number of different degrees of badness however, and is therefore incapable of faithfully representing obligations with more than a certain number of mitigating factors.

Using an infinite number of sanctions would allow us to faithfully model every CTD obligation but lacks the simplicity that makes sanction semantics so attractive. In fact, the simplest way to represent arbitrary obligations using an infinite number of sanctions is probably to use a preference order on the sanctions and let the possible worlds inherit this order, thus reducing the use of sanctions to the use of a preference relation on the possible worlds.

7 Conclusion

A logic modeling the gentle murder scenario using sanction semantics can be expected to satisfy the **Invariance under propositional renaming**, **Determinacy of sanction**, **Range of outcomes**, **Invariance under act renaming**

and **Invariance under outcome renaming** desiderata. Several such logics exist; examples include SA of [5], XSTIT of [7] and a common semantics for SDL.

It is not possible for a logic to faithfully model the gentle murder scenario while satisfying all the desiderata. If we drop any one of the desiderata a logic can be found that faithfully models the gentle murder scenario while satisfying all remaining desiderata, although most such logics are not very useful. An exception is the logic using a preference relation on the possible worlds, which satisfies all desiderata except **Invariance under outcome renaming** and seems capable of faithfully modeling any CTD obligation. This logic can hardly be considered to be based on sanction semantics, however.

References

1. Chisholm, R.M.: Contrary-to-duty imperatives and deontic logic. Analysis **24** (1963) 33–36
2. Åqvist, L.: Deontic logic. In Gabbay, D.M., Guenthner, F., eds.: Handbook of Philosophical Logic. Volume 8. 2nd edn. Springer (2002)
3. Carmo, J., Jones, A.J.I.: Deontic logic and contrary-to-duties. In Gabbay, D.M., Guenthner, F., eds.: Handbook of Philosophical Logic. Volume 8. 2nd edn. Springer (2002)
4. Anderson, A.R., Moore, O.K.: The formal analysis of normative concepts. American Sociological Review **22** (1957) 9–17
5. Bartha, P.: Conditional obligation, deontic paradoxes, and the logic of agency. Annals of Mathematics and Artificial Intelligence **9** (1993) 1–23
6. Broersen, J.M.: Modal Action Logics for Reasoning about Reactive Systems. PhD thesis, Vrije Universiteit Amsterdam (2003)
7. Broersen, J.M.: Deontic epistemic stit logic distinguishing modes of mens rea. Journal of Applied Logic **9** (2011) 127–152
8. Lomuscio, A., Sergot, M.: Deontic interpreted systems. Studia Logica **75** (2003) 63–92
9. Meyer, J.C.: A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. Notre Dame Journal of Formal Logic **29** (1988) 109–136
10. Forrester, J.W.: Gentle murder, or the adverbial samaritan. The Journal of Philosophy **81** (1984) 193–197
11. van der Torre, L.: Reasoning about Obligations: Defeasibility in Preference-Based Deontic Logic. PhD thesis, Erasmus University Rotterdam (1997)