Deontic Equilibrium Logic with Paraconsistent Modality

Pedro Cabalar[1][(0000−0001−7440−0953]], Ovidio Manteiga[1][0009−0003−0107−0000]], and Concepción Vidal[1,2][0000−0002−5561−6406]]

[1] University of A Coruña, Spain
{cabalar,ovidio.manteiga,concepcion.vidalm}@udc.es
[2] CITIC, University of A Coruña, Spain

Abstract. Equilibrium Logic constitutes a well-established logical characterisation of Answer Set Programming (ASP) that has been recently extended to the deontic formalism called DELX. Focused on obligations of literals (built with explicit negation), DELX allows a certain degree of paraconsistency: a literal can be both obligatory and forbidden when one of those two norms is violated. This is convenient for contrary-to-duty situations, where a secondary obligation can coexist with a contradictory, primary one when the latter is violated. Unlike modal approaches, any DELX theory can be reduced to a normal form where deontic operators are only applied on atoms. This allows a direct translation into ASP and was exploited in the recent tool for normative reasoning `deolingo`. However, the simplicity and flexibility of DELX come at the price of its impossibility of a proper representation of disjunctive obligations like $\mathbf{O}(p \vee q)$ that become DELX-equivalent to $\mathbf{O}p \vee \mathbf{O}q$. A similar problem arises with conjunctive prohibitions. In this paper, we study a modal deontic extension of equilibrium logic, MDEL, that avoids these problems by treating obligations as proper modal operators, while preserving the same restricted paraconsistency of DELX. We prove some basic properties of the new formalism and illustrate its use with several examples. We show that the monotonic basis of MDEL is strictly weaker than the one of DELX and that, as a result, any DELX equilibrium model of a theory is also an MDEL equilibrium model. Finally, we prove that DELX reasoning is formally correct in MDEL for the syntactic fragment where modal operators are exclusively applied on explicit literals.

1 Introduction

The concerns about the legal and ethical implications of the widespread use of Artificial Intelligence have renewed interest in reasoning about norms and regulations. The formalisation of normative concepts such as obligation, permission or prohibition is the topic of study of *Deontic* systems [6,5] that evolved from the so-called *Standard Deontic Logic* (SDL) [22] (a.k.a. modal logic KD) to overcome some of its limitations. In the literature, several prototypical scenarios, also called 'paradoxes', were used to illustrate multiple features of deontic reasoning that could not be covered by SDL. For instance, the following well-known scenario can be used to illustrate both the need of *defeasible norms* versus the potential inconsistency in *contrary-to-duty* (CTD) obligations (namely, when a secondary norm is derived from the violation of a primary one):

Example 1 (Cottage Fence [21]). A holiday cottage regulation consists of the norms:
(i) There must be no fence (f).
(ii) If, even so, there is a fence, it must be a white (w) fence.
(iii) If the cottage is by the sea (s), there may be a fence. □

Note how (iii) is an explicit permission to override the prohibition of fences (i), so the latter behaves as a defeasible norm that is *not effective* when the exception (being by the sea) is applicable. On the other hand, (ii) is a CTD obligation acquired when (i) is violated (a fence has been placed). In such a case, the secondary obligation of putting a white fence also implies the obligation to put a fence contradicts the prohibition of fences (i) that must be *still effective*. While it has been clear that defeasibility [15,18,13] requires introducing some form of *Non-Monotonic Reasoning* (NMR), the representation of CTDs has generated more controversy, leading to an heterogenenous umbrella of technical solutions to avoid the potential inconsistency between a primary and a secondary obligation. Some solutions establish formal differences between the two kinds of obligations, either by using different operators, establishing priorities or adding some kind of temporal precedence, but these normally imply further formalisation effort. Another alternative is to remove the **D** axiom $\neg(\mathbf{O}p \wedge \mathbf{O}\neg p)$ so that, for instance, the obligation of a fence $\mathbf{O}f$ and its prohibition $\mathbf{O}\neg f$ can coexist, forming in this way some sort of paraconsistency [4]. The usual problem with this kind of solution is that dilemmas (situations where we have equal instructions to do one thing *and* the opposite) are no longer inconsistent.

The recent *Deontic Equilibrium Logic with Explicit Negation* (DELX) [2] provided a solution to these and other problems in deontic reasoning by extending *Answer Set Programming* (ASP) [17] and in particular, its logical characterisation *Equilibrium Logic* [20], to incorporate deontic operators. DELX exploits the use of the two negations already present in ASP: on the one hand, defeasible norms are naturally represented using the standard ASP default negation (we will denote here as '$\neg$'). On the other hand, obligations in DELX are focused on literals built with an *explicit negation* [1] (written here as '$\sim$') allowing some degree of paraconsistency: $\mathbf{O}f$ and $\mathbf{O}{\sim}f$ are not inconsistent if one of them is violated. An important advantage of DELX is that any theory can be reduced to a normal form where deontic operators are only applied to atoms. This enables a direct translation into ASP without extra computational cost, something exploited by the deontic reasoning tool `deolingo` [3] built on top of the popular ASP solver `clingo` [7]. However, this normal form reduction is possible thanks to a quite radical treatment of disjunctive obligations. For instance, the imperative "fight in the army (a) or perform alternative civil (c) service" [16] formulated as $\mathbf{O}(a \vee c)$, is equivalent in DELX to the disjunction $\mathbf{O}a \vee \mathbf{O}c$ that, intuitively, is a stronger statement forcing one of the two atomic obligations to hold. The same counterintuitive behaviour arises with a prohibition of a conjunction:

Example 2. It is forbidden to drive (d) a car exceeding a limit of alcohol (a) in blood. □

If we formulate this prohibition as expected $\mathbf{O}{\sim}(d \wedge a)$ and we assume $\sim$ satisfies De Morgan laws, we again get a disjunctive obligation $\mathbf{O}({\sim}a \vee {\sim}d)$ that, in DELX, collapses to $\mathbf{O}{\sim}a \vee \mathbf{O}{\sim}d$ meaning that either we forbid driving or we forbid drinking alcohol, rather than forbidding the *simultaneous occurrence* of the two facts.

In this paper, we propose a new deontic extension of equilibrium logic, MDEL, where deontic connectives become proper modal operators under a Kripke semantics, while preserving the same paraconsistent treatment of obligations for literals formed with explicit negation. With respect to DELX, this new extension shows a more cautious behaviour for disjunctive obligations (analogously, conjunctive prohibitions) but, as a counterpart, its Kripke semantics does not allow a simple translation into ASP any more. We accompany the new definition with the proof of several basic properties. The main result is that MDEL collapses to DELX for the syntactic fragment where modal operators are exclusively applied on explicit literals.

The rest of the paper is organised as follows. The next section revisits the logic DELX from [2] and its monotonic basis DHTX. Then, Section 3 introduces the new formalism, MDEL, in terms of its monotonic basis, Modal Deontic Here-and-There (MDHT). Section 4 provides some properties connecting both MDEL and MDHT to the non-modal variants DELX and DHTX, respectively. Finally, Section 5 concludes the paper. We include an appendix with the proofs of the theorems for reviewing purposes.

2 Background: Revisiting DELX

In this section, we revisit the definition of *Deontic Equilibrium Logic with eXplicit Negation* (DELX) [2] and its monotonic basis, *Deontic Here-and-There with eXplicit negation* (DHTX). Given a (countable) set of propositional atoms At, we define a *(deontic) formula* φ as any expression following the grammar:

$$\varphi ::= \bot \mid a \mid \varphi \wedge \varphi \mid \varphi \vee \varphi \mid \varphi \rightarrow \varphi \mid {\sim}\varphi \mid \mathbf{O}\varphi$$

where a is any atom in At. Operator '$\sim$' represents *explicit negation* whereas $\mathbf{O}$ stands for *obligation*. We define the derived operators $\neg\varphi := \varphi \rightarrow \bot$, $\top := \neg\bot$, $\mathbf{F}\varphi := \mathbf{O}{\sim}\varphi$, $\mathbf{P}\varphi := {\sim}\mathbf{F}\varphi$. The last two expressions are respectively read as "φ is *forbidden*" and "φ is *explicitly permitted*". Operator $\neg\varphi$ corresponds to *default negation* and is usually represented in ASP as '$not\ \varphi$'. A *(deontic) theory* is a set of (deontic) formulas. Given any set of atoms A, we define the corresponding set of *(explicit) literals* as $lit(A) := A \cup \{{\sim}p \mid p \in A\}$. A set $S \subseteq lit(A)$ is said to be *consistent* if it contains no pair $\{p, {\sim}p\}$ for any $p \in A$.

Definition 1 (DHTX-interpretation). *A DHTX-*interpretation $\mathcal{I}$ *is a pair* $\mathcal{I} = \langle \mathcal{I}^h, \mathcal{I}^t \rangle$ *of consistent sets of explicit literals where:*

(i) $\mathcal{I}^h \subseteq \mathcal{I}^t \subseteq lit(At')$ *for the extended signature:*
$At' := At \cup \{\mathbf{O}p \mid p \in At\} \cup \{\mathbf{F}p \mid p \in At\}$
(ii) $\{\mathbf{O}p, \mathbf{F}p\} \subseteq \mathcal{I}^t$ *implies* $\{p, {\sim}p\} \cap \mathcal{I}^t \neq \emptyset$*, for any* $p \in At$. □

The intuition about the two sets, $\mathcal{I}^h$ and $\mathcal{I}^t$, is that literals in $\mathcal{I}^h$ are considered to be "*founded*" whereas literals in $\mathcal{I}^t \setminus \mathcal{I}^h$ have less strength and are just "*assumed*" but not founded. The rest of literals in $lit(At') \setminus \mathcal{I}^t$ are just *not assumed*. For instance, take the signature $\{f, w, s\}$ for Example 1 and consider the interpretation I_0:

$$\mathcal{I}_0^h = \{s, {\sim}\mathbf{F}f\} \qquad \mathcal{I}_0^t = \mathcal{I}_0^h \cup \{\mathbf{O}w, {\sim}f\} \tag{1}$$

Then, literals s and $\sim\mathbf{F}f$ are founded, whereas $\mathbf{O}w$ and $\sim f$ are only assumed. On the other hand, the rest of possible literals not included in T, such as f, $\sim s$, w, $\sim w$, $\mathbf{F}f$, $\mathbf{O}f$, $\sim\mathbf{O}f$, etc, are not assumed. As we can see, in general, some literals can be assumed but not founded, as happened to $\mathbf{O}w$ and $\sim f$ above. The only situation in which this never happens is when $\mathcal{I}^h = \mathcal{I}^t$: in that case, we say that interpretation $\mathcal{I}$ is *total*. For simplicity, we will sometimes abbreviate $\langle \mathcal{I}^t, \mathcal{I}^t \rangle$ just as $\mathcal{I}^t$ when there is no ambiguity.

Condition (ii) allows the coexistence of an obligation $\mathbf{O}p$ and a prohibition $\mathbf{F}p$ of the same atom if one of them is violated, namely, either $p \in \mathcal{I}^t$ violates the prohibition or $\sim p \in \mathcal{I}^t$ violates the obligation. This condition corresponds to the axiom schemata:

$$\mathbf{O}p \wedge \mathbf{F}p \wedge \neg p \wedge \neg\sim p \rightarrow \bot \qquad (\mathbf{wD})$$

that, for every atom p, forces a weaker version of the **D** axiom $\mathbf{O}p \wedge \mathbf{F}p \rightarrow \bot$. As we will see later, $\mathbf{F}p$ is DHTX-equivalent to $\mathbf{O}\sim p$ so DHTX introduces a kind of paraconsistency where, in general, $\mathbf{O}p \wedge \mathbf{O}\sim p \equiv \mathbf{O}(p \wedge \sim p) \not\equiv \mathbf{O}\bot \equiv \bot$. Axiom (**wD**) enforces consistency when none of the two obligations is violated, namely $\neg p \wedge \neg\sim p$ having no information about p. In that case, we face an unresolved *dilemma* situation, that is, we have both the obligation and the prohibition of a fact, something that becomes contradictory, as customary in the deontic reasoning literature.

The use of explicit negation leads to a three-valued semantics where, in addition to satisfaction of formulas $\models$, we also have a dual falsification relation $\dashv$. Besides, formulas will be evaluated with respect to one of the two HT "coordinates" $x \in \{h, t\}$ and will depend on a *context* $c \in \{\mathbf{r}, \mathbf{o}, \mathbf{f}\}$ standing for *real*, *obligatory* and *forbidden*, respectively. Each context c has a dual $\overline{c}$ defined as $\overline{\mathbf{r}} := \mathbf{r}$, $\overline{\mathbf{o}} := \mathbf{f}$ and $\overline{\mathbf{f}} := \mathbf{o}$. The satisfaction $\models$ and falsification $\dashv$ relations for DHTX are shown in Figure 1. In the figure, M just stands for the interpretation $\mathcal{I} = \langle \mathcal{I}^h, \mathcal{I}^t \rangle$ but we keep that different name so that, later on, M will correspond to a richer structure that also contains $\mathcal{I}$.

A DHTX-interpretation $\mathcal{I}$ is a *model* of a theory Γ, written $\mathcal{I} \models \Gamma$, if $\mathcal{I}, \mathbf{r}, h \models \varphi$ for all $\varphi \in \Gamma$. A formula φ is a DHTX-*tautology* (or DHTX-*valid*), written $\models \varphi$, if any DHTX-interpretation is a model of φ. DHTX is the logic induced by all DHTX-tautologies.

Definition 2 (DHTX-Equilibrium model). *A total DHTX-interpretation $\mathcal{I}$ (when $\mathcal{I}^h = \mathcal{I}^t$) is a DHTX-*equilibrium model *of a theory Γ if $\mathcal{I} \models \Gamma$ and there is no model $\mathcal{J} \models \Gamma$ such that $\mathcal{J}^t = \mathcal{I}^t$ and $\mathcal{J}^h \subset \mathcal{I}^h$.*

DELX is the logic induced by the set of DHTX-equilibrium models of a deontic theory.

Example 3 (Example 1 continued). Example 1 can be represented as:

$$\neg\sim\mathbf{F}f \rightarrow \mathbf{F}f \qquad (2)$$

$$f \wedge \mathbf{F}f \rightarrow \mathbf{O}w \qquad (3)$$

$$s \rightarrow \sim\mathbf{F}f \qquad (4)$$

$$(w \rightarrow f) \ \wedge \ (\mathbf{O}w \rightarrow \mathbf{O}f) \qquad (5)$$

Formulas (2)-(4) respectively correspond to norms (i)-(iii) while (5) captures the commonsense knowledge that having a white fence implies there is a fence and that the

φ	$M, c, x \models \varphi$ **when**	$M, c, x \dashv \varphi$ **when**
$\top$ ($\bot$)	always (never)	never (always)
$\alpha \wedge \beta$	$M, c, x \models \alpha$ and $M, c, x \models \beta$	$M, c, x \dashv \alpha$ or $M, c, x \dashv \beta$
$\alpha \vee \beta$	$M, c, x \models \alpha$ or $M, c, x \models \beta$	$M, c, x \dashv \alpha$ and $M, c, x \dashv \beta$
$\alpha \rightarrow \beta$	for $y \in \{x, t\}$: $M, c, y \not\models \alpha$ or $M, c, y \models \beta$	$M, c, t \models \alpha$ and $M, c, x \dashv \beta$
$\sim\alpha$	$M, \overline{c}, x \dashv \alpha$	$M, \overline{c}, x \models \alpha$
p	$p \in \mathcal{I}^x$ if $c = \mathbf{r}$ $\mathbf{O}p \in \mathcal{I}^x$ if $c = \mathbf{o}$ $\sim\mathbf{F}p \in \mathcal{I}^x$ if $c = \mathbf{f}$	$\sim p \in \mathcal{I}^x$ if $c = \mathbf{r}$ $\sim\mathbf{O}p \in \mathcal{I}^x$ if $c = \mathbf{o}$ $\mathbf{F}p \in \mathcal{I}^x$ if $c = \mathbf{f}$
$\mathbf{O}\alpha$	$M, \mathbf{o}, x \models \alpha$	$M, \mathbf{o}, x \dashv \alpha$

Fig. 1. DHTX satisfaction and falsification relations where $M = \mathcal{I}$.

obligation of a white fence implies the obligation of a having a fence. We may observe that (2)-(5) plus the axiom ($\mathbf{wD}$) looks like an ASP program (with explicit negation), assuming we read formulas like '$\mathbf{O}a$' and '$\mathbf{F}a$' as ASP atoms. In fact, Cabalar et al. [2] proved that, in general, any DHTX theory is reducible to a syntactic fragment, called *Deontic Atom Normal Form* (DANF), where deontic operators are only applied on atomic expressions of the forms $\mathbf{O}a$ and $\mathbf{F}a$ where a is an atom. Once in DANF, the formula is then reducible to a logic program so that the answer sets of the latter correspond to the DHTX-equilibrium models of the former. Using this connection, it is easy to see that the only equilibrium model of (2)-(5) is a total interpretation $\mathcal{I}$ formed by $\mathcal{I}^h = \mathcal{I}^t = \{\mathbf{F}f\}$. If we add the fact f stating that we put a fence, then we get a new, unique equilibrium model formed by $\{\mathbf{F}f, f, \mathbf{O}w, \mathbf{O}f\}$ where we have derived the obligation to put a white fence $\mathbf{O}w$ and so, the obligation to put a fence $\mathbf{O}f$ too. The potential contradiction between $\mathbf{F}f$ and $\mathbf{O}f$ is cancelled by the fact that $\mathbf{F}f$ has been violated by f. Finally, if we extend the theory once more, stating that we are by the sea (2)-(5)$\cup\{f, s\}$, we obtain a unique equilibrium model $\{f, s, \sim\mathbf{F}f\}$ where we have derived the explicit permission $\sim\mathbf{F}f$ to put a fence, retracting the default prohibition $\mathbf{F}f$ derived before from (2). □

Cabalar et al. [2] illustrated the expressiveness of DELX with multiple examples combining CTD with dilemmas, defeasible obligations, explicit versus implicit permissions and defeasible deontic detachment. However, all these cases were focused on obligations of explicit literals, and left apart more complex formulas such as disjunctive obligations. Let us consider a classical example from the literature:

Example 4 (From [16]). An agent is constrained by the following two imperatives:

(i) Fight in the army (a) or perform alternative civil (c) service.
(ii) Do not fight in the army.

Intuitively, (i) corresponds to the disjunctive obligation $\mathbf{O}(a \vee c)$ whereas the second is a prohibition $\mathbf{F}a$ or equivalently $\mathbf{O}{\sim}a$. If we are given (i) alone, since $\mathbf{O}(a \vee c) \equiv \mathbf{O}a \vee \mathbf{O}c$ in DHTX, we obtain two equilibrium models $\{\mathbf{O}a\}$ and $\{\mathbf{O}c\}$ where we force

(exclusively) one of the two obligations to occur. If we add (ii) to the theory, then $\mathbf{F}a$ forms a dilemma with $\mathbf{O}a$ according to ($\mathbf{wD}$) since none of the two obligations has been violated. As a result, we obtain a unique equilibrium model $\{\mathbf{F}a, \mathbf{O}c\}$ where we derive the obligation to perform civil service, as expected. Let us elaborate the example by adding the following two premises:

(iii) If you fight in the army and violate (ii) then do not use firearms (f).

(iv) You fight in the army.

Our new theory would contain the formulas $\{\mathbf{O}(a \vee c), \mathbf{F}a, (\mathbf{F}a \wedge a \rightarrow \mathbf{F}f), a\}$ where $(\mathbf{F}a \wedge a \rightarrow \mathbf{F}f)$ states a CTD. In this case, the dilemma between $\mathbf{O}a$ with $\mathbf{F}a$ does no longer hold, because the prohibition has been violated. As a result, the two options from (i) are enabled and we obtain the two equilibrium models $\{\mathbf{O}a, \mathbf{F}a, a, \mathbf{F}f\}$ and $\{\mathbf{O}c, \mathbf{F}a, a, \mathbf{F}f\}$. □

Even though the previous example shows a reasonable behaviour in practice, the fact that $\mathbf{O}(a \vee c)$ is eventually treated as $\mathbf{O}a \vee \mathbf{O}c$ seems a rather radical assumption. This becomes perhaps clearer in the following case:

Example 5 (Example 2 revisited). Consider again the prohibition to drive drunk, that can be stated as $\mathbf{F}(d \wedge a)$. This formula is DHTX-equivalent to $\mathbf{F}d \vee \mathbf{F}a$ and produces the two equilibrium models $\{\mathbf{F}d\}$ and $\{\mathbf{F}a\}$, meaning that either we completely forbid driving or we completely forbid drinking alcohol (above the limit), a statement that is stronger than the original prohibition. If, for instance, we add evidence of driving without exceeding the alcohol limit $d \wedge {\sim}a$, we obtain an equilibrium model $\{\mathbf{F}a, d, {\sim}a\}$ where we fulfil a prohibition (no alcohol) but another one $\{\mathbf{F}d, d, {\sim}a\}$ where we make a violation of the prohibition to drive. □

In what follows, we study a possible way to *weaken* the DELX behaviour for disjunctive obligations (through a Kripke-based modal semantics) to better accommodate this kind of compound obligations.

3 Modal Deontic HT

We proceed now to define *modal deontic*-HT (MDHT). The syntax of MDHT extends DHTX formulas with a new operator $\widehat{\mathbf{O}}\varphi$ that will act as a dual of $\mathbf{O}\varphi$. We also define $\widehat{\mathbf{P}}\varphi := {\sim}\widehat{\mathbf{O}}{\sim}\varphi$ which will become a kind of stronger explicit permission. Given the propositional signature At, we define its corresponding *deontic signature* as $At_D := At \cup \{\overline{p} \mid p \in At\}$. Intuitively, atoms $\overline{p}$ will be used for prohibitions whereas literals ${\sim}\overline{p}$ will constitute explicit permissions. An HT-*world* I is a pair $\langle I^h, I^t\rangle$ of consistent sets of explicit literals $I^h \subseteq I^t \subseteq lit(At_D)$. As a notation remark, we will use non-calligraphic capital letters I, J for denoting HT-worlds formed over At_D, as a distinction from calligraphic $\mathcal{I}$, $\mathcal{J}$ for DHTX-interpretations over At' seen before.

Definition 3 (MDHT-interpretation). *An* MDHT-interpretation *is a pair $\langle \mathbb{W}, I\rangle$ where I is a distinguished HT-world (the "*actual*" world) for signature At and $\mathbb{W} \neq \emptyset$ is a set of HT-worlds for signature At_D such that, for all $I, J \in \mathbb{W}$, if $I^t = J^t$ then $I^h = J^h$.*

To increase readability, given a cluster $\mathbb{W} = \{J_1, J_2, \ldots, J_n\}$ we will remove the commas and use square brackets, writing $[J_1\ J_2\ \ldots\ J_n]$ instead. The intuition of $\langle \mathbb{W}, I\rangle$ is that the actual world I exclusively contains literals in $lit(At)$, that is, of the forms p or $\sim p$, meaning that we have evidence for p true or false, respectively. On the other hand, the set $\mathbb{W}$ will act as a cluster (like in modal logic S5) of HT-worlds over literals $lit(At_D)$, that is, of any of the forms p, $\sim p$, $\overline{p}$ or $\sim\overline{p}$, but their meaning now is slightly different: p will represent the obligation of p; $\overline{p}$ its prohibition; $\sim\overline{p}$ an explicit permission for p; and $\sim p$ an explicit permission for $\sim p$. This meaning will be combined with quantification: an obligation (or prohibition) must be satisfied in all worlds (as usual with modal *necessity* operators), whereas a permission must hold in at least one (as usual with modal possibility operators). Besides, as we can see, both the real world and each point J in the cluster have the two components J^h and J^t that, as before, are used for separating founded literals (those in J^h) from just assumed ones (those in $J^t \setminus J^h$).

Example 6. As an example of interpretation, take $M_0 = \langle \mathbb{W}_0, I_0\rangle$ with $I_0^h = \{s\}$, $I_0^t = \{s, \sim f\}$ and the cluster:

$$\mathbb{W}_0 = [\quad \langle\{\sim\overline{f}\}, \{w, \sim\overline{f}\}\rangle \quad \langle\emptyset, \{w\}\rangle \quad \langle\{w, \overline{s}\}, \{w, \overline{s}\}\rangle \quad] \tag{6}$$

Let us respectively call J_1, J_2, J_3 to the three worlds above. In this cluster, the obligation for w is assumed, since $w \in J_i^t$ for all worlds $i = 1, 2, 3$. However, this obligation is not founded, since $w \notin J_1^h$ and $w \notin J_2^h$. On the other hand, the permission of f is founded, since we have at least one world $\sim\overline{f} \in J_1^h$. The last condition of Def.3 implies that we cannot have new worlds in $\mathbb{W}$ like $\langle\cdot, \{w, \sim\overline{f}\}\rangle$, $\langle\cdot, \{w\}\rangle$ or $\langle\cdot, \{w, \overline{s}\}\rangle$ because those three t-components already occur in $\mathbb{W}$. □

An MDHT-interpretation $M = \langle \mathbb{W}, I\rangle$ is said to be *total* if all its HT-worlds are total, that is, $J^h = J^t$ for all $J \in \mathbb{W} \cup \{I\}$. Given any MDHT-interpretation $M = \langle \mathbb{W}, I\rangle$, we define its corresponding total interpretation as $M^t := \langle \mathbb{W}^t, \langle I^t, I^t\rangle\rangle$ where $\mathbb{W}^t := \{\langle J^t, J^t\rangle \mid J \in \mathbb{W}\}$. Informally speaking, M^t overrides all the h-components in all the worlds in M to make them coincide with their corresponding t-components. As an example, $M_0^t = \langle \mathbb{W}_0^t, \langle I_0^t, I_0^t\rangle\rangle$ with I_0^t from (6) and the cluster:

$$\mathbb{W}_0^t := [\quad \langle\{w, \sim\overline{f}\}, \{w, \sim\overline{f}\}\rangle \quad \langle\{w\}, \{w\}\rangle \quad \langle\{w, \overline{s}\}, \{w, \overline{s}\}\rangle \quad]$$

As we did in DHTX, we will sometimes write total worlds $\langle J^t, J^t\rangle$ simply as J^t when there is no ambiguity: for instance, the above cluster becomes $[\ \{w, \sim\overline{f}\}\ \{w\}\ \{w, \overline{s}\}\ ]$. It is easy to see that M is total iff $M^t = M$.

The satisfaction $\models$ and falsification $\dashv$ relations for MDHT are shown in Figure 2. Note that we reuse Figure 1, but in the latter, we now take the liberty of reading M as the MDHT-interpretation $M = \langle \mathbb{W}, I\rangle$ instead of $M = \mathcal{I}$ we had before. We say that an MDHT-interpretation M is a *model* of a theory Γ, written $M \models \Gamma$ when $M, \mathbf{r}, h \models \varphi$ for every $\varphi \in \Gamma$. The following theorem proves an essential property in intermediate logics called *persistence*: in HT, it means that anything holding at h must still hold at t.

Theorem 1 (Persistence). *For any formula α, any MDHT-interpretation $M = \langle \mathbb{W}, I\rangle$ and any context $c \in \{\mathbf{r}, \mathbf{o}, \mathbf{f}\}$:*

- $M, c, h \models \alpha$ *implies* $M, c, t \models \alpha$.
- $M, c, h \dashv \alpha$ *implies* $M, c, t \dashv \alpha$.

φ	$M, c, x \models \varphi$ **when**	$M, c, x \mathrel{=\!\!|} \varphi$ **when**
$\bot \top \rightarrow$ $\wedge \vee \sim$	as in DHTX (Fig. 1)	as in DHTX (Fig. 1)
p	$p \in I^x$ if $c \neq \mathbf{f}$ $\sim\overline{p} \in I^x$ if $c = \mathbf{f}$	$\sim p \in I^x$ if $c \neq \mathbf{f}$ $\overline{p} \in I^x$ if $c = \mathbf{f}$
$\mathbf{O}\alpha$	$\forall J \in \mathbb{W}\colon \langle \mathbb{W}, J\rangle, \mathbf{o}, x \models \alpha$	$\exists J \in \mathbb{W}\colon \langle \mathbb{W}, J\rangle, \mathbf{o}, x \mathrel{=\!\!|} \alpha$
$\widehat{\mathbf{O}}\alpha$	$\exists J \in \mathbb{W}\colon \langle \mathbb{W}, J\rangle, \mathbf{o}, x \models \alpha$	$\forall J \in \mathbb{W}\colon \langle \mathbb{W}, J\rangle, \mathbf{o}, x \mathrel{=\!\!|} \alpha$
Derived		
$\neg\alpha$	$M, c, t \not\models \alpha$	$M, c, t \models \alpha$
$\mathbf{F}\alpha$	$\forall J \in \mathbb{W}\colon \langle \mathbb{W}, J\rangle, \mathbf{f}, x \mathrel{=\!\!|} \alpha$	$\exists J \in \mathbb{W}\colon \langle \mathbb{W}, J\rangle, \mathbf{f}, x \models \alpha$
$\mathbf{P}\alpha$	$\exists J \in \mathbb{W}\colon \langle \mathbb{W}, J\rangle, \mathbf{f}, x \models \alpha$	$\forall J \in \mathbb{W}\colon \langle \mathbb{W}, J\rangle, \mathbf{f}, x \mathrel{=\!\!|} \alpha$
$\widehat{\mathbf{P}}\alpha$	$\forall J \in \mathbb{W}\colon \langle \mathbb{W}, J\rangle, \mathbf{f}, x \models \alpha$	$\exists J \in \mathbb{W}\colon \langle \mathbb{W}, J\rangle, \mathbf{f}, x \mathrel{=\!\!|} \alpha$

Fig. 2. MDHT satisfaction and falsification relations, being $M = \langle \mathbb{W}, I\rangle$.

Proposition 1. *The semantics for derived operators $\neg$, $\mathbf{F}$, $\mathbf{P}$ and $\widehat{\mathbf{P}}$ correspond to the last four lines of the table in Figure 2.*

We say that two formulas φ and ψ are *equivalent*, written $\varphi \equiv \psi$, when both (1) $M, c, x \models \varphi$ iff $M, c, x \models \psi$ and (2) $M, c, x \mathrel{=\!\!|} \varphi$ iff $M, c, x \mathrel{=\!\!|} \psi$, for all MDHT-interpretation M and any $c \in \{\mathbf{r}, \mathbf{o}, \mathbf{f}\}$, $x \in \{h, t\}$.

Note that MDHT is also paraconsistent, as happened in DHTX: in general, $\mathbf{O}p \wedge \mathbf{F}p = \mathbf{O}(p \wedge \sim p) \not\equiv \bot$. Moreover, even when $\mathbf{O}(p \wedge \sim p)$ does not hold, we may still have now that $\widehat{\mathbf{O}}(p \wedge \sim p)$, namely, both the obligation of p and its prohibition $\overline{p}$ may hold in *some* world in the cluster $J \in \mathbb{W}$. Thus, we propose a stronger variant of ($\mathbf{wD}$) where a non-violated $\widehat{\mathbf{O}}(p \wedge \sim p)$ becomes a dilemma:

$$\widehat{\mathbf{O}}(p \wedge \sim p) \wedge \neg p \wedge \neg\sim p \rightarrow \bot \qquad (\mathbf{wD'})$$

Proposition 2. *An MDHT-interpretation $M = \langle \mathbb{W}, I\rangle$ for signature At is a model of ($\mathbf{wD'}$) iff, for all $p \in At$: $p \in I^t$ or $\sim p \in I^t$ or there is no $J \in \mathbb{W}$ s. t. $\{p, \overline{p}\} \subseteq J^t$.*

In what follows, we will assume that ($\mathbf{wD}$') is satisfied. It must be noticed, however, that this axiom only applies to atoms and is not extrapolable to arbitrary formulas. As a counterexample, note that the formula $\widehat{\mathbf{O}}(\varphi \wedge \sim\varphi) \wedge \neg\varphi \wedge \neg\sim\varphi$ is consistent[3] for $\varphi = (a \vee c)$ even under the presence of ($\mathbf{wD}$').

We can now define the equilibrium models in this extended modal setting.

Definition 4 (MDHT-Equilibrium model). *A total MDHT-interpretation M, where $M = \langle \mathbb{W}, I\rangle$ and $M^t = M$, is an (MDHT)*-equilibrium model *of a theory Γ if it is a model of Γ and there is no other model M' of Γ such that $(M')^t = M$.*

[3] As an example of a model, we may take a total interpretation $M = \langle \{J\}, I\rangle$ with $I^h = I^t = \{\sim a\}$ and $J^h = J^t = \{a, \overline{a}, \overline{c}\}$.

MDEL is the logic induced by the set of MDHT-equilibrium models of a deontic theory.

Example 7 (Example 2 revisited). Suppose our theory only contains the formula $\mathbf{F}(d \wedge a) = \mathbf{O}{\sim}(d \wedge a)$. Following Figure 2, this formula forces clusters $\mathbb{W}$ where all the worlds $J \in \mathbb{W}$ must satisfy either $\overline{d} \in J^h$ or $\overline{a} \in J^h$. It is not hard to see that, for deciding equilibrium models, only these two literals $\overline{d}$ and $\overline{a}$ are relevant in the set of worlds: the addition of any other literal $d, {\sim}d, {\sim}\overline{d}, a, {\sim}a, {\sim}\overline{a}$, in any J^t for $J \in \mathbb{W}$ or in I^t will never be in equilibrium, because we can always form a smaller M' with $(M')^t = M$ that removes these atoms making them unfounded. As a result, equilibrium models will have an empty real world $I^t = \emptyset$ (no literal is assumed). Similarly, the assumed literals for J^t in the cluster can only be either $\{\overline{a}\}$ or $\{\overline{d}\}$ since adding the two atoms $\{\overline{a}, \overline{d}\}$ together is unfounded (we can always remove one of them in J^h). This only leaves us three equilibrium models $M_1 = \langle \mathbb{W}_1, I\rangle$, $M_2 = \langle \mathbb{W}_2, I\rangle$ and $M_3 = \langle \mathbb{W}_3, I\rangle$ with $I^h = I^t = \emptyset$ and the clusters $\mathbb{W}_1 = [\ \{\overline{a}\}\ ], \mathbb{W}_2 = [\ \{\overline{d}\}\ ], \mathbb{W}_3 = [\ \{\overline{a}\}\ \{\overline{d}\}\ ]$. Interestingly, M_1 and M_2 correspond to the two equilibrium models we obtained in DELX, $\{\mathbf{F}a\}$ and $\{\mathbf{F}d\}$ respectively: in fact, we will see later that any DHTX-equilibrium model corresponds to an MDHT-equilibrium model. The novelty in MDEL is the third equilibrium model M_3. Notice that the equilibrium model minimisation works in a similar way as in ASP: the rest of literals are disregarded due to a lack of evidence and, in fact, the disjunction between $\overline{a}$ and $\overline{d}$ is also minimised, acting as an exclusive-or. If we add the information $d \wedge {\sim}a$ to our theory, we get again three equilibrium models formed by $\mathbb{W}_1$, $\mathbb{W}_2$ and $\mathbb{W}_3$ as before with a common actual world $I^h = I^t = \{d, {\sim}a\}$. □

In the previous example, we may observe that, when we add $d \wedge {\sim}a$, we still get an equilibrium model (formed by $\mathbb{W}_2$) where we violate the prohibition $\mathbf{F}d$. However, the situation now is weaker, since we also have now an equilibrium model with cluster $\mathbb{W}_3$ that contains "less information" since $\mathbb{W}_2 \subseteq \mathbb{W}_3$ (in fact, $\mathbb{W}_1 \subseteq \mathbb{W}_3$ too). In general, we may be sometimes interested in a preference relation among MDHT-equilibrium models using the following order:

Definition 5 (Relation $\preceq$). *Given two total MDHT-interpretations $M = \langle \mathbb{W}, I\rangle$ and $M' = \langle \mathbb{W}', I'\rangle$ we say that M is* less-preferred *than M', written $M \preceq M'$, when $I = I'$ and for all $J \in \mathbb{W}$ there exists some $J' \in \mathbb{W}'$ such that $(J')^t \subseteq J^t$.*

In the previous example, the only $\preceq$-maximal equilibrium model is $\mathbb{W}_3$ and $I^t = \{d, {\sim}a\}$. In that model, we are not violating any prohibition: moreover, we do fulfil the compound prohibition $\mathbf{F}(d \wedge a)$ because ${\sim}a$ holds and so $(d \wedge a)$ is falsified. We now illustrate relation $\preceq$ applied on permissions:

Example 8. Take the following pair of explicit permissions:

(i) A license gives you permission to drive (d).
(ii) Being adult, you can drink alcohol (a).

For simplicity, suppose you are adult and have a driving license, so $\mathbf{P}d \wedge \mathbf{P}a$ applies. The four equilibrium models of this conjunction are $M_i = \langle \mathbb{W}_i, \emptyset\rangle$ with $i = 1..4$ and:

$$\mathbb{W}_1 = [\ \{{\sim}\overline{d}, {\sim}\overline{a}\}\ ] \qquad \mathbb{W}_2 = [\ \{{\sim}\overline{d}, {\sim}\overline{a}\}\ \emptyset\ ]$$
$$\mathbb{W}_3 = [\ \{{\sim}\overline{d}\}\ \{{\sim}\overline{a}\}\ ] \qquad \mathbb{W}_4 = [\ \{{\sim}\overline{d}\}\ \{{\sim}\overline{a}\}\ \emptyset\ ]$$

In this case, we also have a unique $\preceq$-maximal equilibrium model formed by $\mathbb{W}_4$. Note how $\mathbb{W}_4$ does not allow concluding an explicit permission to drive and drink simultaneously. Such a permission would correspond to the formula $\mathbf{P}(d \wedge a)$ whose equilibrium models are $\mathbb{W}_1$ and $\mathbb{W}_2$, being the latter the only $\preceq$-maximal one in this case. Finally, if we combine the two permissions $\mathbf{P}a$ and $\mathbf{P}d$ with the prohibition to drive drunk $\mathbf{F}(d \wedge a)$ we get four equilibrium models M_5, M_6, M_7 and M_8 formed by a common actual world $I = \emptyset$ and the respective clusters:

$$\mathbb{W}_5 = [\ \{\overline{a}, \sim\overline{d}\}\ \{\overline{d}, \sim\overline{a}\}\ ] \qquad \mathbb{W}_6 = [\ \{\overline{a}, \sim\overline{d}\}\ \{\overline{d}, \sim\overline{a}\}\ \{\overline{a}\}\ ]$$
$$\mathbb{W}_7 = [\ \{\overline{a}, \sim\overline{d}\}\ \{\overline{d}, \sim\overline{a}\}\ \{\overline{d}\}\ ] \qquad \mathbb{W}_8 = [\ \{\overline{a}, \sim\overline{d}\}\ \{\overline{d}, \sim\overline{a}\}\ \{\overline{a}\}\ \{\overline{d}\}\ ]$$

being $\mathbb{W}_8$ the only $\preceq$-maximal one. On the other hand, if we take the formula $\mathbf{F}(d \wedge a) \wedge \mathbf{P}(d \wedge a)$ instead, we get no equilibrium models as expected – note that $\mathbf{P}(d \wedge a)$ amounts to $\sim\mathbf{F}(d \wedge a)$ by definition. □

Example 9 (Example 4 continued). Consider the first two premises (i)-(ii) from Example 4 modelled as the theory $\{\mathbf{O}(a \vee c), \mathbf{F}a\}$. Due to axiom (**wD'**), this theory produces a unique MDHT-equilibrium model $\langle [\ \{\overline{a}, c\}\ ], \emptyset \rangle$ where c becomes obligatory, as happened in DELX. If we extend the theory by covering premises (i)-(iv), that is $\{\mathbf{O}(a \vee c), \mathbf{F}a, (\mathbf{F}a \wedge a \rightarrow \mathbf{F}f), a\}$, then the dilemma is removed because $\mathbf{F}a$ is violated and we obtain now three MDHT-equilibrium models: $\langle [\ \{\overline{a}, c, \overline{f}\}\ ], \{a\} \rangle$, $\langle [\ \{\overline{a}, a, \overline{f}\}\ ], \{a\} \rangle$ and $\langle [\ \{\overline{a}, c, \overline{f}\}\ \{\overline{a}, a, \overline{f}\}\ ], \{a\} \rangle$. The first two cases correspond to the two DELX-equilibrium models we saw in Example 4, whereas the last one is the only $\preceq$-maximal equilibrium model and keeps $\mathbf{O}(a \vee c)$ without concluding neither $\mathbf{O}a$ nor $\mathbf{O}c$. Note how, in this model, paraconsistency only occurs in world $\{\overline{a}, a, \overline{f}\}$, where both the obligation a and prohibition $\overline{a}$ can coexist because of the violation of $\mathbf{F}a$. □

One interesting observation in the previous examples is that the DELX-equilibrium models seem to correspond to its MDHT-equilibrium models whose clusters are singletons. We will see next that, in fact, this is a general property.

4 Some properties of MDHT

In this section, we study some properties of MDHT, starting with its relation to DHTX. In what follows, and for simplicity, we will use the same symbols $\models$ and $\models\!\!\!=$ for evaluation of formulas both in DHTX and MDHT: ambiguity is resolved by the form of the interpretations used on the left, $\mathcal{I}$ with HT-interpretation over $lit(At')$ for DHTX and $M = \langle \mathbb{W}, I \rangle$ with $\mathbb{W} \cup \{I\}$ HT-worlds over $lit(At_D)$ for MDHT. We start identifying a kind of MDHT-interpretations that will have a strong relation to DHTX.

Definition 6 (Singleton MDHT-interpretation). *An MDHT-interpretation M, where $M = \langle \mathbb{W}, I \rangle$, is said to be* singleton *if its cluster is a singleton, that is* $|\mathbb{W}| = 1$. □

Notice that M is a singleton iff M^t is a singleton.

Theorem 2. *An MDHT-interpretation M is singleton iff M is a model of the axiom schemata, for every $\ell \in lit(At)$:*

$$(\widehat{\mathbf{O}}\ell \rightarrow \mathbf{O}\ell) \wedge (\sim\mathbf{O}\ell \rightarrow \sim\widehat{\mathbf{O}}\ell) \qquad \textbf{(S)}$$

Under axiom (**S**), we can easily observe that $\widehat{\mathbf{O}}\varphi \equiv \mathbf{O}\varphi$ since the existential and universal quantification used in both operators collapse to a same condition when we have a singleton cluster. Therefore, if (**S**) holds, we also get that (**wD'**) collapses to (**wD**).

Definition 7 (Modal mapping μ). *Given a DHTX-interpretation $\mathcal{I}$ for signature At, we define $\mu(\mathcal{I})$ as the singleton MDHT-interpretation $\mu(\mathcal{I}) = \langle\{J\}, I\rangle$ where:*

$$\begin{aligned} I^x &:= \{p \mid p \in lit(At) \cap \mathcal{I}^x\} \\ J^x &:= \{p \mid \mathbf{O}p \in \mathcal{I}^x\} \cup \{\overline{p} \mid \mathbf{F}p \in \mathcal{I}^x\} \\ &\quad \cup \{\sim p \mid \sim\mathbf{O}p \in \mathcal{I}^x\} \cup \{\sim\overline{p} \mid \sim\mathbf{F}p \in \mathcal{I}^x\} \end{aligned}$$

for every $x \in \{h, t\}$. □

For instance, given $\mathcal{I}_0$ from equation (1), we get $\mu(\mathcal{I}_0) = \langle\{J_0\}, I_0\rangle$ where $I_0 = \langle\{s\}, \{s, \sim f\}\rangle$ as in Example 6 and $J_0 = \langle\{\sim\overline{f}\}, \{\sim\overline{f}, w\}\rangle$. It is very easy to see that function μ is bijective, so it has an inverse μ^-: given any singleton $M = \langle\{J\}, I\rangle$, $\mu^-(M)$ corresponds to $\mathcal{I} = \langle\mathcal{I}^h, \mathcal{I}^t\rangle$ where:

$$\begin{aligned} \mathcal{I}^x := I^x \cup\ &\{\mathbf{O}p \mid p \in At, p \in J^x\} \cup\ \{\mathbf{F}p \mid p \in At, \overline{p} \in J^x\} \\ &\cup\ \{\sim\mathbf{O}p \mid p \in At, \sim p \in J^x\} \cup\ \{\sim\mathbf{F}p \mid p \in At, \sim\overline{p} \in J^x\} \end{aligned}$$

Theorem 3. *Let φ be any DHTX-formula (that is, not containing $\widehat{\mathbf{O}}$) over signature At, let $\mathcal{I}$ be a DHTX-interpretation over At and let $\mu(\mathcal{I}) = \langle\{J\}, I\rangle$ its corresponding singleton MDHT-interpretation. Then, for all $x \in \{h, t\}$ and all $c \in \{\mathbf{o}, \mathbf{f}\}$:*

(i) $\mathcal{I}, \mathbf{r}, x \models \varphi \ \Leftrightarrow\ \langle\{J\}, I\rangle, \mathbf{r}, x \models \varphi$
(ii) $\mathcal{I}, \mathbf{r}, x \dashv \varphi \ \Leftrightarrow\ \langle\{J\}, I\rangle, \mathbf{r}, x \dashv \varphi$
(iii) $\mathcal{I}, c, x \models \varphi \ \Leftrightarrow\ \langle\{J\}, J\rangle, c, x \models \varphi$
(iv) $\mathcal{I}, c, x \dashv \varphi \ \Leftrightarrow\ \langle\{J\}, J\rangle, c, x \dashv \varphi$ □

As a consequence, $\mathcal{I}, \mathbf{r}, h \models \varphi$ iff $\mu(\mathcal{I}), \mathbf{r}, h \models \varphi$ and, since μ is bijective for singleton interpretations, we conclude that DHTX is strictly stronger than MDHT:

Corollary 1. *Logic DHTX corresponds to MDHT restricted to singleton interpretations. In other words:* DHTX = MDHT + (**S**). □

This can be used to prove that, in the non-monotonic setting, DELX is also strictly stronger than MDEL. In particular, DHTX-equilibrium models correspond exactly to MDHT-equilibrium models that are singleton, as stated below:

Theorem 4. *Let $\mathcal{I}$ be a total DHTX-interpretation and Γ a deontic theory (not containing $\widehat{\mathbf{O}}$). Then: $\mathcal{I}$ is a DHTX-equilibrium model of Γ iff $\mu(\mathcal{I})$ is a MDHT-equilibrium model of Γ.*

Recall that μ is bijective: for any singleton MDHT-equilibrium model M, this theorem guarantees that we can always get the corresponding DHTX-equilibrium model $\mathcal{I} = \mu^-(M)$.

In what follows, we will show that DHTX is a sound approximation of MDHT when restricted to formulas in the syntactic fragment of DANF. We recall that a deontic formula φ is in DANF if all the occurrences of deontic operators in φ are of the forms $\mathbf{O}a$

or $\mathbf{F}a$ where $a \in At$. Although DANF constitutes a normal form for DHTX, this is not the case in MDHT, though: for instance, the formula $\mathbf{O}(a \vee c)$ is not MDHT-reducible to DANF. We start defining, for any arbitrary MDHT-interpretation M, a particular singleton interpretation $\delta(M)$ assigned to M:

Definition 8. *Given an MDHT-interpretation $M = \langle \mathbb{W}, I \rangle$, we define the singleton interpretation $\delta(M) := \langle \{K\}, I \rangle$ where*

$$\begin{aligned} K^x := \quad & \{A \in J^x \mid A \in At_D, \textit{ for all } J^x \in \mathbb{W}\} \\ & \cup \{\sim A \in J^x \mid \sim A \in At_D, \textit{ for some } J^x \in \mathbb{W}\} \end{aligned}$$

for $x \in \{h, t\}$.

For instance, given M_0 defined before Equation (6), we get $\delta(M) = \langle \{K_0\}, I_0 \rangle$ with I_0 as in M_0 and $K_0 = \langle \{\sim\overline{f}\}, \{\sim\overline{f}, w\} \rangle$ which coincides with J_0 seen after Def. 7. Moreover, $\mu^-(\delta(\mathcal{I}))$ is $\mathcal{I}_1$ from Equation (1). It is obvious that, if M is already singleton, then $\delta(M) = M$. On the other hand, if M is any total interpretation $M^t = M$, then $\delta(M)$ is also total: $\delta(M)^t = \delta(M)$. The following result shows that the set of MDHT-models of a formula in DANF is completely characterised by its singleton models.

Theorem 5. *Let φ be a formula in DANF and M an MDHT-interpretation. Then: $M, \mathbf{r}, h \models \varphi$ iff $\delta(M), \mathbf{r}, h \models \varphi$.*

Corollary 2. *Checking MDHT-satisfiability (resp. MDHT-validity) of a formula φ in DANF can be reduced to checking its satisfiability (resp. validity) in MDHT+($\mathbf{S}$), or equivalently (Corollary 1) in DHTX.*

The correspondence at the non-monotonic level (i.e. with respect to equilibrium models) is stated in the following theorem:

Theorem 6. *If M is a MDHT-equilibrium model of a formula φ in DANF then $\delta(M)$ is also a MDHT-equilibrium model of φ.*

This result is weaker than in the monotonic case. Even if the formula is in DANF, we may obtain more MDHT-equilibrium models than in DHTX, as happens for arbitrary formulas after all. The novelty now is that if M is an equilibrium model, then the singleton $\delta(M)$ will also be in equilibrium. Moreover, these singleton equilibrium models are *sufficiently representative* in the sense that they suffice to cover the satisfaction/falsification of deontic atoms $\mathbf{O}a$ and $\mathbf{F}a$. As a result, singleton equilibrium models are enough to check equilibrium entailment when formulas are in DANF.

Theorem 7. *Let φ and ψ be two formulas in DANF. To check whether $M, \mathbf{r}, h \models \psi$ in all (resp. some) MDHT-equilibrium model of φ is equivalent to check it in all (resp. some) singleton MDHT-equilibrium model of φ, that is, all (resp. some) DHTX-equilibrium model of φ.*

Example 10 (Example 8 revisited). As an illustration, consider again the formula $\mathbf{P}d \wedge \mathbf{P}a$. This formula is an abbreviation of $\sim\mathbf{F}d \wedge \sim\mathbf{F}a$ that is in DANF. As we saw before, the formula has four MDHT-equilibrium models M_1, M_2, M_3 and M_4. We may observe

that only one of them, M_1 with cluster $\mathbb{W}_1 = [\{\sim\overline{d}, \sim\overline{a}\}]$ is singleton and so, due to Theorem 4, $\mu^-(M_1) = \{\sim\mathbf{F}d, \sim\mathbf{F}a\}$ is the unique DHTX-equilibrium model of $\mathbf{P}d \wedge \mathbf{P}a$. Notice that, in M_1, we may derive the formula $\mathbf{P}(d \wedge a)$, that is, we have an explicit permission for both driving and drinking alcohol at the same time. This is no surprise, since $\mathbf{P}d \wedge \mathbf{P}a \equiv \mathbf{P}(d \wedge a)$ in DHTX. Remember, however, that the only $\preceq$-maximal MDHT-equilibrium model was M_4 where this simultaneous permission is not derived. Now, since φ is in DANF and M_1 is the only singleton equilibrium model, Theorem 6 further guarantees that $\delta(M_i) = M_1$ with $i = 1..4$, that is, for the four equilibrium models of φ. Finally, Theorem 7 tells us that, if we are only interested in equilibrium entailment of formulas in DANF, such as $\sim\mathbf{F}a$ or $\sim\mathbf{F}d$ (or in general, any propositional combination of deontic atoms) then M_1 suffices and there is no need to compute the rest of MDHT-equilibrium models. □

We conclude this section focusing on the fragment of MDHT induced by total models. In the case of regular HT, total models are enforced by the excluded middle axiom $p \vee \neg p$, which makes the logic collapse to classical propositional logic. Moreover, the same happens with Equilibrium Logic: the addition of excluded middle removes non-monotonicity and also makes equilibrium models collapse to classical models. In the case of MDHT, we can extend the excluded middle axiom as follows:

$$EM(p) \wedge \mathbf{O}(EM(p)) \wedge \widehat{\mathbf{P}}(EM(p)) \qquad \textbf{(EM)}$$

for any atom $p \in At$, where $EM(p) := (p \vee \neg p) \wedge (\sim p \vee \neg\sim p)$.

Theorem 8. *A MDHT-interpretation M is total iff M is a model of* (**EM**).

As we can see, forcing the excluded middle makes the "two-component" behaviour of HT disappear, leading to total interpretations. This is analogous to the non-modal setting, where adding the excluded middle to HT, makes it collapse to classical propositional logic. Note that $\widehat{\mathbf{P}}(EM(p))$ in (**EM**) amounts to $\sim\widehat{\mathbf{O}}\sim(EM(p))$ and that, if we remove this operator $\widehat{\mathbf{O}}$ from the language, it is no longer possible to characterise the set of total interpretations as in Theorem 8. The reason for this is that, by removing $\widehat{\mathbf{O}}$, we lose existential quantification for context $\mathbf{o}$, and so, we lose universal quantification for context $\mathbf{f}$ over the worlds in the cluster. By examining the proof of Theorem 8 we can observe that any model $M = \langle \mathbb{W}, I \rangle$ of $EM(p)$ is forced to have a total actual world $I^h = I^t$, whereas the other part of the axiom $\mathbf{O}(EM(p)) \wedge \widehat{\mathbf{P}}(EM(p))$ enforces total clusters $\mathbb{W}^t = \mathbb{W}$. The fragment of total clusters can also be interesting: it removes non-monotonicity in deontic formulas, but keeps the logic programming behaviour for all the rest. To put an example, if we add the axiom $\mathbf{O}(EM(p)) \wedge \widehat{\mathbf{P}}(EM(p))$ to the formula $\mathbf{F}(d \wedge a)$ from Example 2, we obtain now 7 equilibrium models whose clusters are formed by all non-empty subsets of $[\{\sim\overline{a}, \sim\overline{d}\}\ \{\sim\overline{a}\}\ \{\sim\overline{d}\}]$.

5 Related Work and Conclusions

In this paper, we have introduced a modal variant of Deontic Equilibrium Logic with Explicit Negation [2] (DELX). This new variant, we called MDEL, allows a more cautious treatment of disjunctive obligations (resp. conjunctive prohibitions) avoiding the

equivalence $\mathbf{O}(a \vee b) \equiv \mathbf{O}a \vee \mathbf{O}b$ imposed by DELX. DELX and MDEL are based on model selection criteria imposed on the monotonic logics DHTX and MDHT, respectively, both of them proper extensions of the logic of Here-and-There [14] (HT). MDHT interpretations have the form of Kripke models consisting of a cluster of HT worlds (that captures the meaning of deontic formulas) plus an isolated HT world that stands for the actual world. We have proved that, for any formula φ, there exists a one-to-one correspondence between its DHTX-models and MDHT-models where the cluster is a singleton. We also proved that any DHTX-equilibrium model is also an MDHT-equilibrium model but not vice versa. However, when the formula is in a given syntactic fragment where deontic operators are only applied to explicit literals, then MDEL collapses to DELX, if we are only interested in entailment of formulas in that syntactic fragment.

Apart from the relation to DELX extensively studied in this paper, other deontic extensions of ASP that have a clear connection to MDEL are *Deontic Logic Programs* (DLP) [10,11] and *Deontic Temporal ASP* (DTASP) [9]. While MDEL provides a full logical characterisation for arbitrary formulas (with no restrictions on connective nesting), both DLP and DTASP are defined in terms of a syntactic transformation extending the ASP program reduct [8]. Besides, the two formalisms rely on SDL (modal logic KD) and do not allow paraconsistent situations where $\mathbf{O}a \wedge \mathbf{F}a$, something that disables a suitable treatment of contrary-to-duty (e.g. Example 3). Another related approach is *Defeasible Deontic Logic* (DDL) [12] (extending *Defeasible Logic* [19]) that is not a proper extension of ASP, but has a similar syntax to logic programming. DDL introduces multiple types of rule implications (strict, defeasible and defeaters) that, again, act as meta-logical connectives and cannot be arbitrarily nested as in MDEL.

As immediate future work, we plan to study additional formal properties of MDHT (especially in relation to HT with explicit negation [1]), and the complexity of reasoning tasks associated to MDHT and MDEL, having in mind a modal extension of the current reasoning tool `deolingo` [3] based on DELX. Finally, we also plan to study a formal relation to the aforementioned formalisms DLP, DTASP and DDL.

Acknowledgements We wish to thank Davide Soldà and Agata Ciabattoni for their useful comments on a previous version of this draft, and the anonymous reviewers for their suggestions. This research was partially supported by the Ministry of Science, Innovation and Universities (SPAIN) MICIU/AEI/10.13039/501100011033 grant PID2023-148531NB-I00. The CITIC center of UDC also receives funding from the Xunta of Galicia and the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01).

References

1. Aguado, F., Cabalar, P., Fandinno, J., Pearce, D., Pérez, G., Vidal, C.: Revisiting explicit negation in answer set programming. Theory and Practice of Logic Programming **19**(5-6), 908–924 (2019). https://doi.org/10.1017/S1471068419000267
2. Cabalar, P., Ciabattoni, A., van der Torre, L.: Deontic equilibrium logic with explicit negation. In: Gaggl, S.A., Martinez, M.V., Ortiz, M. (eds.) Logics in Artificial Intelligence - 18th

European Conference, JELIA 2023, Dresden, Germany, September 20-22, 2023, Proceedings. Lecture Notes in Computer Science, vol. 14281, pp. 498–514. Springer (2023)
3. Cabalar, P., Manteiga, O.: deolingo: Extending answer set programming with deontic reasoning. In: van Berkel, K., Ciabattoni, A., Horty, J.F. (eds.) Deontic Logic and Normative Systems - 17th International Conference, DEON 2025, Vienna, Austria, July 1-3, 2025. pp. 61–77. College Publications (2025)
4. da Costa Jr, N., Carnielli, W.: On paraconsistent deontic logic. Philosophia **16**, 293–305 (12 1986). https://doi.org/10.1007/BF02379748
5. Gabbay, D., Horty, J., Parent, X., van der Mayden, R., van der Torre, L. (eds.): Handbook of Deontic Logic and Normative Systems, Volume 2. College Publications (2021)
6. Gabbay, D., Horty, J., Parent, X., van der Meyden, R., van der Torre, L. (eds.): Handbook of Deontic Logic and Normative Systems. College Publications (2013)
7. Gebser, M., Kaminski, R., Kaufmann, B., Ostrowski, M., Schaub, T., Wanko, P.: Theory solving made easy with clingo 5. In: Carro, M., King, A., Saeedloei, N., Vos, M.D. (eds.) Technical Communications of the 32nd International Conference on Logic Programming, ICLP 2016 TCs. OASIcs, vol. 52, pp. 2:1–2:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2016). https://doi.org/10.4230/OASIcs.ICLP.2016.2, https://doi.org/10.4230/OASIcs.ICLP.2016.2
8. Gelfond, M., Lifschitz, V.: The stable model semantics for logic programming. In: Proc. of the 5th Intl. Conference on Logic Programming (ICLP'88). pp. 1070–1080 (1988)
9. Giordano, L., Martelli, A., Dupré, D.T.: Temporal deontic action logic for the verification of compliance to norms in ASP. In: Francesconi, E., Verheij, B. (eds.) International Conference on Artificial Intelligence and Law, ICAIL 2013. pp. 53–62. ACM (2013)
10. Gonçalves, R., Alferes, J.J.: An embedding of input-output logic in deontic logic programs. In: Ågotnes, T., Broersen, J.M., Elgesem, D. (eds.) DEON 2012, Proceedings. LNCS, vol. 7393, pp. 61–75. Springer (2012)
11. Gonçalves, R., Alferes, J.J.: Deontic logic programs. In: Gini, M.L., Shehory, O., Ito, T., Jonker, C.M. (eds.) International conference on Autonomous Agents and Multi-Agent Systems. pp. 1333–1334. IFAAMAS (2013)
12. Governatori, G., Olivieri, F., Rotolo, A., Scannapieco, S.: Computing strong and weak permissions in defeasible logic. Journal of Phil. Logic **42**(6), 799–829 (2013). https://doi.org/10.1007/s10992-013-9295-1
13. Governatori, G., Tosatto, S.C., Rotolo, A.: A defeasible deontic logic for pragmatic oddity. In: Liu, F., Marra, A., Portner, P., Putte, F.V.D. (eds.) Deontic Logic and Normative Systems - 15th International Conference, DEON 2020/21. pp. 239–254. College publications (2021)
14. Heyting, A.: Die formalen Regeln der intuitionistischen Logik. Sitzungsber. Preuss. Akad. Wiss., Phys.-Math. Kl. pp. 42–56 (1930)
15. Horty, J.F.: Deontic logic as founded on nonmonotonic logic. Ann. Math. Artif. Intell. **9**(1-2), 69–91 (1993). https://doi.org/10.1007/BF01531262
16. Horty, J.F.: Moral dilemmas and nonmonotonic logic. Journal of Philosophical Logic **23**(1), 69–91 (1994)
17. Lifschitz, V.: Action Languages, Answer Sets, and Planning, pp. 357–373. Springer Berlin Heidelberg, Berlin, Heidelberg (1999)
18. Nute, D. (ed.): Defeasible Deontic Logic. Kluwer, Dordrecht (1997)
19. Nute, D.: Defeasible logic. In: Handbook of Logic in Artificial Intelligence and Logic Programming. Oxford University Press (1993)
20. Pearce, D.: A new logical characterisation of stable models and answer sets. In: NMELP. LNCS, vol. 1216, pp. 57–70. Springer (1997)
21. Prakken, H., Sergot, M.: Dyadic deontic logic and contrary-to-duty obligations. In: Nute, D. (ed.) Defeasible Deontic Logic. pp. 223–262. Dordrecht (1997). https://doi.org/10.1007/978-94-015-8851-5_10

22. von Wright, G.H.: Deontic logic. Mind **60**(237), 1–15 (1951)