A Ceteris Paribus Deontic Logic

Andrea Loreggia, Emiliano Lorini, Giovanni Sartor

A Ceteris Paribus Deontic Logic

Andrea Loreggia[2][0000−0002−9846−0157], Emiliano Lorini[3][0000−0002−7014−6756], and Giovanni Sartor[1,2][0000−0003−2210−0398] ⋆ ⋆⋆

[1] CIRSFID - Alma AI, University of Bologna, Italy
[2] European University Institute, Florence, Italy
[3] IRIT

Abstract. We present a formal semantics for deontic logic based on the concept of ceteris paribus preferences. We introduce notions of unconditional obligation and permission as well as conditional obligation and permission that are interpreted relative to this semantics. We show that these notions satisfy some intuitive properties and, at the same time, do not encounter some problems and paradoxes that have been extensively discussed in the deontic logic literature. Moreover, we show how obligations and permissions can be represented compactly using existing preference frameworks from the artificial intelligence area of computational social choice.

Keywords: Deontic Logic · Ceteris paribus preferences · CP-net.

1 Introduction

Artificial agents are used to automate tasks in many different scenarios. Nowadays, they are so pervasive and so fast that it is almost impossible for humans to monitor them in order to predict illegal behaviour. A possible solution is to embed a mapping of the governance into these entities. This will allow to partially translate legal and ethical requirements into computable representations of legal knowledge and reasoning. An example comes from obligations and permissions that are pervasive in law. Obligations are used to impose a requirement while permissions describe allowances. Both of them are concepts captured in deontic logic which has been viewed as a promising component of computational models of legal knowledge and reasoning, on different grounds. On the other side, AI's researchers look for modelling legal knowledge and reasoning about it. They can find in deontic logic a set of formal tools, usually based on modal logic [2, 9], which could be compositionally integrated with other logical formalism, such as predicate logic, logic programming, or defeasible logics. By complementing (computational) logics with deontic logic, it was hoped that a logical formalism would able to capture the specificities of legal language [3, 19]. We leverage

⋆ A. Loreggia and G. Sartor have been supported by the H2020 ERC Project "CompuLaw" (G.A. 833647).

on preferences, which are central in making decision processes and thus implemented in many frameworks to drive and assist individuals (e.g., recommender system [32], sentiment analysis [11]). In this regard, a new line of AI's research is focused on comparing agents' preferences in order to understand how similar they are [20, 22]. The approach is useful in different multi-agent scenarios to compare agents behaviour with exogenous priorities. Examples of such priorities can be a moral principle, a legal requirement, a guideline or a business process. The aim is to understand whether an agent deviates from a desired behaviour [21, 23]. In this paper, we provide the semantics for a deontic logic based on the intuitive idea that obligations and permissions consist in preferences over worlds: strict ones for obligations and weak ones for permissions. Such preferences are *ceteris paribus* in the sense that they only concern worlds that are equal in all remaining circumstances, namely, in all aspects except for those contributing to the states of affairs that are affirmed to be obligatory or permitted. Thus, deontic propositions are to be evaluated against model-theoretical frames consisting of sets of worlds over which ceteris paribus preferences are defined. While there have been various attempts at basing a deontic logic on the idea of preferences (see for instance [15, 2, 16]), nobody has yet, to the best of our knowledge, explored the perspective of a ceteris paribus deontic logic. Our formalisation does not allow for the derivation of deontic paradoxes, but supports some deontic inferences. In the following pages, we first provide a formal account of the idea of ceteris paribus preferences, and of the corresponding semantic structures. We then formalise conditional and unconditional obligations and permissions as ceteris paribus preferences, and study their basic logical properties. We illustrate and discuss both ceteris paribus preferences and deontic operators with the help of extensive examples. We also show how existing AI formalisms (such as CP-nets [4]) can be used to represent our idea of ceteris paribus deontic logic.

2 Related work

The idea of ceteris paribus preference was originally introduced by Georg Henrik Von Wright [29, 31]. He observes that our preferences over states of affairs (as opposed to preferences over objects) are usually "holistic", in the sense that they address the compared situations — denoted by Boolean combination of atoms — in the circumstances accompanying such states. Ceteris paribus preferences have recently been the object of renewed interest by logicians, who have developed ceteris paribus semantics for action and preference [12, 27].

An intuitive semantics and a simple axiomatisation are provided by the so-called standard deontic logic or SDL [7], built on the basis of the so-called old system of deontic logic by [28], (for a discussion of SDL, see [17, 18]). In SDL, to be obligatory means to be true in all ideal (perfectly good) worlds, and to be permissible means to be true in at least one ideal world. This idea is captured by a serial accessibility relation R over possible worlds, to be understood an ideality relation: for every world u there exists at least one world v, such that uRv (v is ideal, relatively to u). In such a semantics frame, φ is obligatory in a world u

if and only if φ is true in every world v such that uRv, and φ is permitted in a world u if and only if there exists a world v such that φ is true in at v and uRv.

As it has been often remarked, SDL gives rise to several apparently counter-intuitive implications, the so-called deontic paradoxes (for a recent discussion, see [2], for an analysis from the perspective of legal theory, see also [33, 13]). First of all, as legal theorist Alf Ross critically observed [26], the obligation of a certain proposition should not entail the obligation of the disjunction of that proposition and any other arbitrary proposition. Other paradoxes have to do with the so-called contrary to duty obligations, namely, with obligations that emerge when other obligations are violated, and whose content may contradict obligations holding when no violation takes place. The classical examples are represented by Forrester's gentle murderer paradox, Roderick Chisholm [6] and by Marek Sergot and Henry Prakken [24]. Various solutions have been proposed to address contrary to duty obligation, often involving technical complexities and sometimes giving rise to additional problems [5].

A further problematic aspect of SDL concerns conditional or rather contextual obligations, namely, assertions to the effect that a certain proposition φ is obligatory under a certain condition ψ. Neither a conditional of classical propositional logic. i.e., $\varphi \rightarrow \mathsf{O}\psi$, nor the embedment of a such a conditional within a deontic operator. i.e., $\mathsf{O}(\varphi \rightarrow \psi)$, appear to provide fully convincing solutions. This issue has spawn the development of dyadic deontic logic, which capture deontic conditionality through a special conditional operator. Dyadic deontic logic was initiated by Georg Von Wright [30], while a semantics for it was first proposed by Bengt Hansson [14].

Technical solutions have been proposed to deal with deontic conditionals and contrary to duty obligations (see [24, 5]). These solutions, however, generally require a more complex logical framework and a less intuitive semantics, in comparison with SDL. Recently, [3] noticed how preference logics and AI preference representation languages are both concerned with reasoning about preferences on combinatorial domains and how in both areas the key notion of ceteris paribus principle appeared for interpreting preference statements[27].

3 Background

In this section, we provide a formal definition of the relevant concepts, and then we discuss them and exemplify their application.

3.1 Ceteris paribus Preferences

To capture the idea of a holistic preference, von Wright considers a set of atoms $Atm = \{p_1, \ldots, p_n\}$, each describing an elementary and independent state of a complete situation, or world. Von Wright [31] observes that the set Atm does not need to account for all states that can exist in the real world. It is rather limited to the "*preference horizon* of a given subject at a given time", namely, to the "states which the subject takes into consideration as constituting accompanying

circumstances when he contemplates his preference or not-preferences between states". We first present the basic elements of the formal semantics, namely, the concepts of preference model and ceteris paribus preference.

Let Atm be a countable set of atomic propositions and let $Lit = Atm \cup \{\neg p : p \in Atm\}$ be the corresponding set of literals.

Definition 1 (Preference model). *A preference model is a tuple $M = (W, \preceq)$ such that: $W = 2^{Atm}$ is the set of worlds, and $\preceq$ is a complete preorder[4] on W.*

Elements of W are denoted by $w, v, \ldots$. We also define $\prec$ and $\approx$ as the strict order and indifference relations induced from $\preceq$. The class of preference models is denoted by $\mathcal{P}$. A weak preference model differs from a preference model as it does not necessarily include all valuations of propositional variables. Specifically, W is a subset of all the possible set of worlds. In particular:

Definition 2 (Weak preference model). *A weak preference model is a tuple $M = (W, \preceq)$ such that $W \subseteq 2^{Atm}$ is the set of worlds, and $\preceq$ is a complete preorder on W.*

Let us introduce the following concepts of *circumstantial* indistinguishability and *circumstantial* preference.

Definition 3 (Circumstantial Indistinguishability). *Let $M = (W, \preceq)$ be a preference model, let $w, v \in W$ and let X be a finite set of atomic propositions. We say that $w \equiv_X v$ iff $\forall p \in X : p \in w$ iff $p \in v$.*

$w \equiv_X v$ means that w and v are indistinguishable, with regard to the circumstances (the atoms) in X.

Definition 4 (Circumstantial Preference). *Let $M = (W, \preceq)$ be a preference model, let $w, v \in W$ and let X be a finite set of atomic propositions. We introduce the following abbreviations: $w \preceq_X v$, iff $w \equiv_X v$ and $w \preceq v$, respectively $w \prec_X v$, iff $w \equiv_X v$ and $w \prec v$.*

$w \preceq_X v$ means that v is at least as good as w, the two worlds being indistinguishable relative to X. $w \prec_X v$ means that v is better than w, the two worlds being indistinguishable relative to X. On the basis of the notions of circumstantial equivalence and preference, we can characterise the notions of ceteris paribus (all-the-rest-being equal) preference relatively to an atom set X.

Definition 5 (Ceteris Paribus Preference). *A world w is* ceteris paribus at least as good as *or* ceteris paribus better than *a world v relative to X, if respectively $v \preceq_{Atm \setminus X} w$ or $v \prec_{Atm \setminus X} w$.*

The former definition concerns indistinguishability and preference relatively to all atoms not in X, i.e., relatively to $Atm \setminus X$.

[4] That is a binary relation on W which is reflexive, transitive and complete.

3.2 CP-net

CP-nets [4] are a compact representation of conditional preferences over ceteris paribus semantics.

Definition 6. *A CP-net over a set of binary variables V is a tuple $\mathcal{N} = (G, CPT)$, where $G = (V, E)$ is a directed graph and $CPT = \{CPT(V_i)|V_i \in V\}$ is a set of conditional preference tables (CP-tables). An edge $(V_i, V_j) \in E$ represents that preferences over $Dom(V_i)$ depend on the value of V_j.*

For each variable $V_i \in V$,, given the assignment to its parents, a CP-table $CPT(V_i)$ represents the preference order over the values of the domain of V_i. For instance, $CPT(A) = \{a \prec \bar{a}\}$ represents the strict preference over the values of a variable A, i.e., $\bar{a}$ is more preferred than a. Each preference order in a CP-table is also called a CP-statement. A CP-net induces a preference graph over all the possible outcomes: each node corresponds to an outcome, that is, a complete assignment of values to variables. Moreover, a directed edge between a pair of outcomes (o_j, o_i), which differ only in the value of one variable, means that $o_j \preceq o_i$. A *worsening flip* is a change in the value of a variable to a less preferred value according to the CP-statement for that variable. A more recent extension, namely CP-net with indifference [1], takes into account indifference and models lack of information using incomparability. The qualitative compact representation of ceteris paribus scenario and the algorithms developed for inference, make CP-net an interesting and useful tool to represent our model.

4 Ceteris paribus Obligations and Permissions

On the basis of the notions introduced in the previous section, we shall now address obligations and permission. Basically, deontic propositions (assertions of obligations and permissions) can be interpreted over semantic structures of ceteris paribus preferences. The basic idea of this paper is that the semantics of obligations and permissions can be captured by viewing obligations as strict ceteris paribus preferences and permissions as weak ceteris paribus preferences. We first introduce a formalisation and then illustrate it with examples.

4.1 Unconditional obligations and permissions: formalisation

We call ceteris paribus Deontic Logic - CPDL the resulting deontic logic of obligation and permission interpreted with respect to the ceteris paribus semantics presented in the previous section. The logic also has the so-called *universal* modal operator that, as we will show in Section 5, allows us to capture factual detachment of obligations.

Definition 7. *$\mathcal{L}_{\mathrm{CPDL}}(Atm)$ is a modal language which includes atomic propositions $p, q, \ldots \in Atm$, standard Boolean operators and the modal operators $\mathsf{O}, \mathsf{P}, \mathsf{U}$. The language is such that:*

- *if* $p \in Atm$, *then* $p \in \mathcal{L}_{\text{CPDL}}$
- *if* $\varphi, \psi \in \mathcal{L}_{\text{CPDL}}$, *then* $\neg\varphi, \varphi \wedge \psi \in \mathcal{L}_{\text{CPDL}}$
- *if* $\varphi, \psi \in \mathcal{L}_{\text{CPDL}}$, *then* $\mathsf{O}\varphi, \mathsf{P}\varphi, \mathsf{U}\varphi \in \mathcal{L}_{\text{CPDL}}$.

Formulas $\mathsf{O}\varphi$ and $\mathsf{P}\varphi$ have to be read, respectively, "φ is obligatory" and "φ is permitted". Formula $\mathsf{U}\varphi$ has to be read "φ is universally true". The truth conditions for the formulas in the language $\mathcal{L}_{\text{CPDL}}(Atm)$ are defined as follows:

Definition 8 (Truth Conditions). *Let* $M = (W, \preceq)$ *be a preference model and let* $w \in W$. *Then:*

$$
\begin{aligned}
M, w \models p &\iff p \in w \\
M, w \models \neg\varphi &\iff M, w \not\models \varphi \\
M, w \models \varphi \wedge \psi &\iff M, w \models \varphi \text{ and } M, w \models \psi \\
M, w \models \mathsf{O}\varphi &\iff \forall v, u \in W : \text{ if } M, v \models \varphi \text{ and } v \preceq_{Atm \setminus Atm(\varphi)} u \text{ then } M, u \models \varphi \\
M, w \models \mathsf{P}\varphi &\iff \forall v, u \in W : \text{ if } M, v \models \varphi \text{ and } v \prec_{Atm \setminus Atm(\varphi)} u \text{ then } M, u \models \varphi \\
M, w \models \mathsf{U}\varphi &\iff \forall v \in W : M, v \models \varphi
\end{aligned}
$$

where $Atm(\varphi)$ denotes the set of atoms from Atm occurring in φ.

In other words, $\mathsf{O}\varphi$ means that, for every two possible worlds that are $Atm \setminus Atm(\varphi)$-indistinguishable and that disagree about the truth value of φ, the world in which φ is true is better than the world in which φ is false. $\mathsf{P}\varphi$ means that, for every two possible worlds that are $Atm \setminus Atm(\varphi)$-indistinguishable and that disagree about the truth value of φ, the world in which φ is true is at least as good as the world in which φ is false. We say that the formula $\varphi \in \mathcal{L}_{\text{CPDL}}(Atm)$ is valid relative to the class of preference models $\mathcal{P}$, denoted by $\models_{\mathcal{P}} \varphi$, iff, for every preference model M and for every world w in M, we have $M, w \models \varphi$. We say that the formula $\varphi \in \mathcal{L}_{\text{CPDL}}(Atm)$ is satisfiable relative to the class of preference models iff, there exists a preference model M and a world w in M, such that $M, w \models \varphi$.

4.2 Conditional obligations and permissions: formalisation

In this section we extend the logic CPDL by operators of conditional obligation and conditional permission. We call CPDL$^+$ the resulting logic.

Definition 9. $\mathcal{L}_{\text{CPDL}^+}(Atm)$ *is a modal language which includes atomic propositions* $p, q, \ldots \in Atm$, *standard Boolean operators and the modal operators* $\mathsf{O}, \mathsf{P}, \mathsf{U}$. *The language is such that:*

- *if* $p \in Atm$, *then* $p \in \mathcal{L}_{\text{CPDL}^+}$
- *if* $\varphi, \psi \in \mathcal{L}_{\text{CPDL}^+}$, *then* $\neg\varphi, \varphi \wedge \psi \in \mathcal{L}_{\text{CPDL}^+}$
- *if* $\varphi, \psi \in \mathcal{L}_{\text{CPDL}^+}$, *then* $\mathsf{O}\varphi, \mathsf{P}\varphi, \mathsf{O}(\psi|\varphi), \mathsf{P}(\psi|\varphi), \mathsf{U}\varphi \in \mathcal{L}_{\text{CPDL}^+}$.

Formulas $\mathsf{O}(\psi|\varphi)$ and $\mathsf{P}(\psi|\varphi)$ have to be read, respectively, "under condition ψ, φ is obligatory " and "under condition ψ, φ is permitted". The truth conditions for the formulas in the language $\mathcal{L}_{\text{CPDL}^+}(Atm)$ are the ones given in Definition 8 *plus* the following two extra truth conditions for the conditional obligation operator and the conditional permission operator:

Definition 10 (Truth conditions (cont.)). *Let $M = (W, \preceq)$ be a preference model and let $w \in W$. Then:*

$$\begin{aligned} M, w \models \mathsf{O}(\psi|\varphi) \iff \forall v, u \in ||\psi||_M : \text{ if } M, v \models \varphi \text{ and} \\ v \preceq_{Atm \setminus Atm(\varphi)} u \text{ then } M, u \models \varphi \\ M, w \models \mathsf{P}(\psi|\varphi) \iff \forall v, u \in ||\psi||_M : \text{ if } M, v \models \varphi \text{ and} \\ v \prec_{Atm \setminus Atm(\varphi)} u \text{ then } M, u \models \varphi \end{aligned}$$

where $||\psi||_M = \{w \in W : M, w \models \psi\}$ is the truth set of ψ relative to the preference model M.

The definitions of validity and satisfiability for the formulas in $\mathcal{L}_{\mathrm{CPDL}^+}(Atm)$ relative to preference models are analogous to the definitions of validity and satisfiability for the formulas in $\mathcal{L}_{\mathrm{CPDL}}(Atm)$ relative to preference models.

5 Properties

In this section, we study the logical properties of the operators of unconditional and conditional obligation and permission introduced above as well as the relationship with CP-nets. Our model does not produce several well-known paradoxes in deontic logic. If we restrict our model to obligations and permissions that are stated only on atoms, then we can show that the induced preference model can be represented compactly by a CP-net with indifference [1].

Proposition 1. *Let C be a set of obligations and permissions. Let $M = (W, \preceq)$ be the minimal preference model induced by C and $\mathcal{N} = (G, P)$ be the CP-net induced by C. Then, M and $\mathcal{N}$ are isomorphic.*

Proof. Due to lack of space, we introduce and explain the inducing notion in what follows. The minimal preference model induced by C is the one which satisfies C and has the less restrictive constraints. First, for each permission in C, introduce a weak order among worlds that differ only for the consequent of the permission, ceteris paribus the antecedent of the permission. For each obligation, introduce a strict order over the worlds that differ only for the consequent of the obligation ceteris paribus the antecedent of the obligation. For all the worlds that are not explicitly compared we introduced a weak order among them. The induced CP-net is built as follows: to each atom $v_i \in Atm$ there is a corresponding variable $V_i \in V$ such that $Dom(V_i) = \{v_i, \bar{v_i}\}$.

Each conditional obligation $O(v_i|v_j) \in N$ introduces a directed edge (V_i, V_j) in the dependency graph G, such that V_i becomes a parent of V_j. It induces a strict order over $Dom(V_j)$ given the assignment of V_i such that $CPT(V_j) = \{v_i : \bar{v_j} \prec v_j, \bar{v_i} : v_j \prec \bar{v_j}\}$. Similarly, each conditional concession $P(v_i|v_j) \in N$ induces a weak order over $Dom(V_j)$ such that $CPT(V_j) = \{v_i : \bar{v_j} \preceq v_j, \bar{v_i} : v_j \preceq \bar{v_j}\}$. Notice that in our model, as well as in SDL, everything that is not explicitly forbidden is permitted, i.e., in general $\neg O(v_i) \to P(\bar{v_i})$. Thus, a variable with indifference over its domain is introduced for each atom that is not explicitly

a consequent of any obligation/permission. To show the isomorphism, consider the bijection between the set of worlds W and the set of all the outcomes in the partial orders. We can show that there is a bijection between edges of the partial order and the ordering relations among worlds in preference model. From the subset of worlds that satisfy all the obligations we can move to less preferred worlds by changing one literal at a time until we visit all the possible worlds. This corresponds to visit all the outcomes in the partial order starting from the subset of optimal outcomes using the definition of worsening flip of CP-net.

Proposition 2. *For all* $\varphi \in \mathcal{L}_{\text{CPDL}^+}(Atm)$*:*

$$\models_{\mathcal{P}} \mathsf{O}\varphi \leftrightarrow \mathsf{O}(\top|\varphi)$$
$$\models_{\mathcal{P}} \mathsf{P}\varphi \leftrightarrow \mathsf{P}(\top|\varphi)$$

This highlights that unconditional obligation and permission do not need to be added as primitives in the language of the logic CPDL$^+$, as they are definable from conditional obligation and permission.

Example 1 (Running example). Let us introduce a running example concerning the presence of cats, dogs, and fences in beach houses (developing the example from [24]). The set of atoms is $Atm = \{c, d, f\}$, where: c represents whether there is a cat; d represents whether there is a dog and f represents whether there is a fence. Mary is the mayor of the town. For safety reasons, she has ordered that there should be fences when there are dogs, and that, on the contrary, there should be no fences when there are no dogs. Cats are allowed with no restrictions. It is easy to check that the following preferences verify $\mathsf{P}(c)$, $\mathsf{O}(d|f)$ and $\mathsf{O}(\neg d|\neg f)$: $w_{\{c,d,f\}} \approx w_{\{d,f\}}, w_{\{c,d\}} \approx w_{\{d\}}, w_{\{c,f\}} \approx w_{\{f\}}, w_{\{\}} \approx w_{\{c\}}, w_{\{c,d\}} \prec w_{\{c,d,f\}}, w_{\{d\}} \prec w_{\{d,f\}}, w_{\{c,f\}} \prec w_{\{c\}}, w_{\{f\}} \prec w_{\{\}}$

Following Proposition 1, Mary's obligations and concessions can be represented using the CP-net with indifference depicted in Figure 1a which compactly represents the partial order depicted in Figure 1b. Following Proposition 1, to each atom there is a corresponding variable, thus we have $V = \{C, D, F\}$ representing respectively whether there is cat, a dog and a fence, $Dom(C) = \{c, \bar{c}\}$, $Dom(D) = \{d, \bar{d}\}$ and $Dom(F) = \{f, \bar{f}\}$. Due to obligations and permission, variables C, D are independent while variable F depends on D. Moreover, obligations define the strict orders over $Dom(D), Dom(F)$. From Proposition 2, the unconditional permission $P(c)$ is defined as $\mathsf{P}(\top|c)$ and introduces indifference over $Dom(C)$.

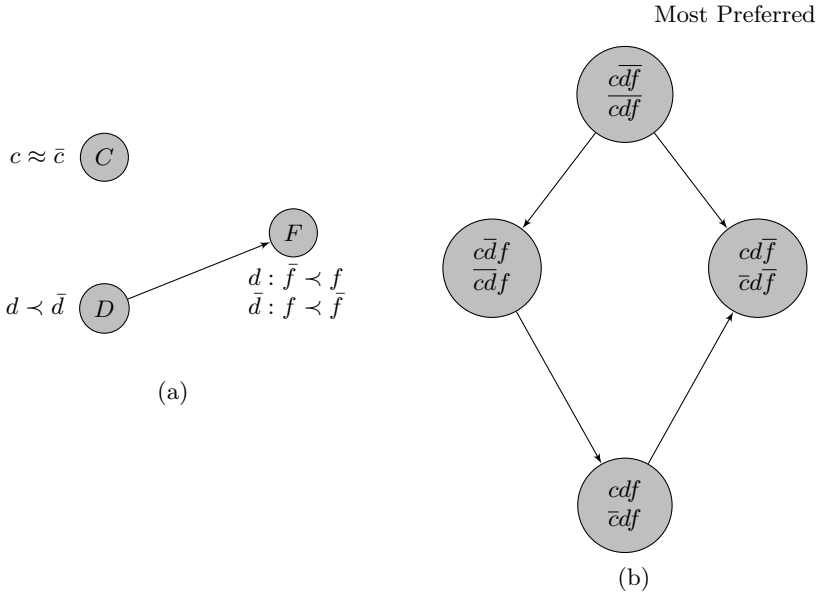
The following proposition highlights that if φ is obligatory, then it is also permitted:

Proposition 3. *For all* $\varphi \in \mathcal{L}_{\text{CPDL}^+}(Atm)$*:*

$$\models_{\mathcal{P}} \mathsf{O}\varphi \rightarrow \mathsf{P}\varphi$$

Before dealing with deontic paradoxes, let us consider how factual detachment is represented in the context of the logic CPDL$^+$:

Fig. 1: (a) The CP-net with indifference which represents obligations and permissions in Example 1. It is built using Proposition 1. (b) The partial order induced by the CP-net in Figure 1a. For the sake of readability, we group into the same nodes some worlds of the preference model. Worlds in the same node are indifferent, this is due to the indifference over the values of variable C.



Proposition 4. *For all* $\varphi \in \mathcal{L}_{\text{CPDL}^+}(Atm)$*:*

$$\models_{\mathcal{P}} (\mathsf{O}(\psi|\varphi) \wedge \mathsf{U}\psi) \rightarrow \mathsf{O}\varphi$$
$$\models_{\mathcal{P}} (\mathsf{P}(\psi|\varphi) \wedge \mathsf{U}\psi) \rightarrow \mathsf{P}\varphi$$

This means that if the condition of a conditional obligation/permission is necessarily true, then the obligation/permission is detached and becomes unconditional.

Let us consider the well-known Ross's paradox [26]. In standard deontic logic (SDL), an obligation to mail a letter (i.e., $\mathsf{O}m$) implies the obligation to mail a letter or to burn it (i.e., $\mathsf{O}(m \vee b)$), something that goes against intuition. As the following preference model highlights, our logic CPDL does not encounter this problem.

Example 2. Let $Atm = \{m, b\}$ with $w_1 = \{m, b\}$, $w_2 = \{m\}$, $w_3 = \{b\}$ and $w_4 = \emptyset$. Let us suppose the following preference order over the worlds in W:

$w_3 \preceq w_1, w_4 \preceq w_2, w_3 \preceq w_2, w_1 \preceq w_2, w_3 \preceq w_4, w_1 \preceq w_4$. We clearly have $M, w_1 \models \mathsf{O}m \wedge \neg\mathsf{O}(m \vee b)$.

More generally, it is worth noting that the ceteris paribus obligation operator is not normal as it does not satisfy Axiom K. In particular, there exists $\varphi, \psi \in \mathcal{L}_{\text{CPDL}^+}(Atm)$ such that the formula $(\mathsf{O}\varphi \wedge \mathsf{O}(\varphi \rightarrow \psi)) \rightarrow \mathsf{O}\psi$ is not valid in CPDL. To show this, it is sufficient to consider the preference model in Example 2. We have $M, w_1 \models (\mathsf{O}m \wedge \mathsf{O}(m \rightarrow (m \vee b)) \wedge \neg\mathsf{O}(m \vee b)$. An interesting property of the ceteris paribus operators for obligation and permission concerns aggregation over conjunction. First of all, it is worth noting that, in the general case, ceteris paribus obligation and permission do not aggregate over conjunction. More precisely, there exists $\varphi, \psi \in \mathcal{L}_{\text{CPDL}^+}(Atm)$ such that: $\not\models_{\mathcal{P}} (\mathsf{O}\varphi \wedge \mathsf{O}\psi) \rightarrow \mathsf{O}(\varphi \wedge \psi), \not\models_{\mathcal{P}} (\mathsf{P}\varphi \wedge \mathsf{P}\psi) \rightarrow \mathsf{P}(\varphi \wedge \psi)$

To show this it is sufficient to consider the following example. (The latter is proved in an analogous way.)

Example 3. Let $Atm = \{p, q\}$ and $w_1 = \{p, q\}, w_2 = \{q\}, w_3 = \{p\}, w_4 = \emptyset$. Moreover, let us consider the following preference order $\preceq$: $w3 \preceq w1 \preceq w4 \preceq w2$.

It is routine exercise to check that $M, w_1 \models \mathsf{O}(p \rightarrow q)$, $M, w_1 \models \mathsf{O}q$, but $M, w_1 \not\models \mathsf{O}((p \rightarrow q) \wedge q)$. To verify the latter it is sufficient to observe that $w_1 \preceq w_4$ and $w_4 \not\preceq w_1$.

Nonetheless, if φ and ψ are conjunctive clauses (i.e., finite conjunctions of literals from Lit) whose sets of atoms have empty intersection (i.e., φ and ψ are independent formulas), then the obligation/permission that φ and the obligation/permission that ψ aggregate over conjunction.

Proposition 5. *If φ, ψ are conjunctive clauses and $Atm(\varphi) \cap Atm(\psi) = \emptyset$ then:*

$$\models_{\mathcal{P}} (\mathsf{O}\varphi \wedge \mathsf{O}\psi) \rightarrow \mathsf{O}(\varphi \wedge \psi)$$
$$\models_{\mathcal{P}} (\mathsf{P}\varphi \wedge \mathsf{P}\psi) \rightarrow \mathsf{P}(\varphi \wedge \psi)$$

Proof. We only prove the former as the latter is proved in an analogous way. We prove it by reductio ad absurdum. Let us suppose that (i) $M, w \models \mathsf{O}\varphi \wedge \mathsf{O}\psi$ and (ii) $M, w \not\models \mathsf{O}(\varphi \wedge \psi)$. Item (i) means that $\forall v, u \in W :$ if $M, v \models \varphi$ and $v \equiv_{Atm \setminus Atm(\varphi)} u$ and $v \preceq u$ then $M, u \models \varphi$, and $\forall v, u \in W :$ if $M, v \models \psi$ and $v \equiv_{Atm \setminus Atm(\varphi)} u$ and $v \preceq u$ then $M, u \models \psi$. Item (ii) means that $\exists v, u \in W : M, v \models \varphi \wedge \psi$ and $v \equiv_{Atm \setminus Atm(\varphi \wedge \psi)} u$ and $v \preceq u$ and $M, u \models \neg\varphi \vee \neg\psi$. We consider the three possible cases for the latter.

Case 1: $M, u \models \neg\varphi \wedge \psi$. Since φ and ψ are conjunctive clauses we have $v \equiv_{Atm \setminus Atm(\varphi)} u$. But this is in contradiction with item (i) above.

Case 2: $M, u \models \varphi \wedge \neg\psi$. Since φ and ψ are conjunctive clauses we have $v \equiv_{Atm \setminus Atm(\psi)} u$. But this is in contradiction with item (i) above.

Case 3: $M, u \models \neg\varphi \wedge \neg\psi$. There exists $w \in W$ such that $M, w \models \neg\varphi \wedge \psi$ and $v \equiv_{Atm \setminus Atm(\varphi)} w$ and $w \equiv_{Atm \setminus Atm(\psi)} u$. It is sufficient to consider the world w such that $\forall p \in Atm(\varphi) : p \in w$ iff $p \in u$, $\forall p \in Atm(\psi) : p \in w$ iff $p \in v$, and $\forall p \in Atm \setminus (Atm(\varphi) \cup Atm(\psi)) : p \in w$ iff $p \in v$. Such a world w exists since $Atm(\varphi) \cap Atm(\psi) = \emptyset$. By item (i) above and the fact that $\preceq$ is complete, we

have that $w \preceq v$. Hence, by the transitivity of $\preceq$, we have $w \preceq u$. The latter is in contradiction with item (i) above.

As shown in Example 3, the formula $(\mathsf{O}(p \to q) \wedge \mathsf{O}q) \to \mathsf{O}((p \to q) \wedge q)$ is not valid in our logic. Notice that $((p \to q) \wedge q)$ is logically equivalent to q. This highlights a more general property of the logic CPDL, namely, the fact that the obligation operator is not closed under logical equivalence. The same property holds for permissions, as they are also not closed under logical equivalence. On the one hand, this might be seen as a limitation of Von Wright's approach to ceteris paribus preferences extended here to ceteris paribus permissions and obligations. Indeed, closure under logical equivalence is the minimal property that any classical modal logic has to satisfy. From this perspective, our logic of obligations is not a classical modal logic. On the other hand, it might be seen as a virtue of the formalism from the point of view of the imperative theory of norms defended, among the others, by Von Wright. Clearly an obligation, seen as a imperative or a command that a certain fact ought to be the case, does not necessarily imply that all its logically equivalent facts are obligatory as well.

We conclude this section by considering the well-known Forrester's gentle murderer paradox [8]. Let us assume the following facts: (1) it is obligatory that you do not kill; (2) if you kill you ought to kill gently, and (3) it is necessarily the case that if you kill gently then you kill. Let us assume that the action of killing is captured by the atom k, while the action of killing gently is captured by the conjunction of atom k and atom g (i.e., doing something gently). Fact 1 is expressed by the formula $\mathsf{O}\neg k$, fact 2 is expressed by the formula $k \to \mathsf{O}(k \wedge g)$, while fact 3 is expressed by the formula $\mathsf{U}((k \wedge g) \to k)$. The latter trivially holds since $(k \wedge g) \to k$ is a tautology of propositional logic. As emphasized in the introduction, under the assumption that you kill, fact 1 and fact 2 are together inconsistent in SDL, i.e., $k \wedge \mathsf{O}\neg k \wedge (k \to \mathsf{O}(k \wedge g))$ is an inconsistent SDL formula. The problem is that in SDL from k and $k \to \mathsf{O}(k \wedge g)$ we can infer $\mathsf{O}(k \wedge g)$ which, in turn, implies $\mathsf{O}k$. The latter is inconsistent with $\mathsf{O}\neg k$, since in SDL conflicting obligations are not admitted. We have a similar problem in our logic, as the formula $k \wedge \mathsf{O}\neg k \wedge (k \to \mathsf{O}(k \wedge g)) \in \mathcal{L}_{\mathrm{CPDL}}(Atm)$ is not satisfiable in the class of preference models. Indeed, from k and $k \to \mathsf{O}(k \wedge g)$ we trivially infer $\mathsf{O}(k \wedge g)$. As the following proposition highlights the latter and $\mathsf{O}\neg k$ are together inconsistent:

Proposition 6. *For all $p, q \in Atm$:*

$$\models_{\mathcal{P}} \mathsf{O}(p \wedge q) \to \neg\mathsf{O}\neg p$$

Proof. Let us suppose that $M, w \models \mathsf{O}(p \wedge q)$. The latter means that $\forall v, u \in W$: if $M, v \models p \wedge q$ and $v \equiv_{Atm \setminus Atm(p \wedge q)} u$ and $v \preceq u$ then $M, u \models p \wedge q$. The latter implies that $\exists v, u \in W : M, v \models p$ and $v \equiv_{Atm \setminus Atm(p)} u$ and $u \preceq v$ and $M, u \models \neg p$. The latter just means that $M, w \models \neg\mathsf{O}\neg p$.

A solution to the Forrester's gentle murderer paradox, widely explored in the literature (see, e.g., [24]) consists in reformulating condition (2) above as the

conditional obligation of killing gently under the condition of killing, and of applying a principle of factual detachment as the one of Proposition 4. Specifically, we again represent fact 1 by the formula $\mathsf{O}\neg k$, while we now represent fact 2 by the formula $\mathsf{O}(k|k \wedge g)$. Moreover, in order to apply factual detachment for the conditional obligation, we need to assume that k is necessarily true, which is represented by the formula $\mathsf{U}k$. The problem is that the latter formula is not satisfiable, as a preference model includes all possible valuations of propositional atoms and, consequently, at least one world in which atom k is false. It follows that the formula $\mathsf{O}\neg k \wedge \mathsf{O}(k|k \wedge g) \wedge \mathsf{U}k$ is not satisfiable either. In order to provide a solution to the Forrester's gentle murderer paradox, which is in line with existing solutions proposed in the literature, we need to weaken the concept of preference model.

Interpretations of formulas with respect to weak preference models is the same as interpretations of formulas with respect to preference models. We denote validity of formulas relative to weak preference models by the symbol $\models_{\mathcal{WP}}$. The following proposition highlights, the validity of Propositions 2, 3 and 4 generalize to weak preference models.

Proposition 7. *For all* $\varphi, \psi \in \mathcal{L}_{\mathrm{CPDL}^+}(Atm)$*:*

$$\begin{aligned}
&\models_{\mathcal{WP}} \mathsf{O}\varphi \leftrightarrow \mathsf{O}(\top|\varphi)\\
&\models_{\mathcal{WP}} \mathsf{P}\varphi \leftrightarrow \mathsf{P}(\top|\varphi)\\
&\models_{\mathcal{WP}} \mathsf{O}\varphi \rightarrow \mathsf{P}\varphi\\
&\models_{\mathcal{WP}} (\mathsf{O}(\psi|\varphi) \wedge \mathsf{U}\psi) \rightarrow \mathsf{O}\varphi\\
&\models_{\mathcal{WP}} (\mathsf{P}(\psi|\varphi) \wedge \mathsf{U}\psi) \rightarrow \mathsf{P}\varphi
\end{aligned}$$

The following proposition is the counterpart of Proposition 5 with respect to weak preference models.

Proposition 8. *If* φ, ψ *are conjunctive clauses and* $Atm(\varphi) \cap Atm(\psi) = \emptyset$ *then:*

$$\begin{aligned}
&\models_{\mathcal{WP}} (\mathsf{AllV}(\varphi \wedge \psi) \wedge \mathsf{O}\varphi \wedge \mathsf{O}\psi) \rightarrow \mathsf{O}(\varphi \wedge \psi)\\
&\models_{\mathcal{WP}} (\mathsf{AllV}(\varphi \wedge \psi) \wedge \mathsf{P}\varphi \wedge \mathsf{P}\psi) \rightarrow \mathsf{P}(\varphi \wedge \psi)
\end{aligned}$$

where:

$$\mathsf{AllV}(\varphi \wedge \psi) =_{def} \bigwedge_{X \subseteq Atm(\varphi \wedge \psi)} \mathsf{E}\Big(\bigwedge_{p \in X} p \wedge \bigwedge_{p \in (Atm(\varphi \wedge \psi)) \setminus X} \neg p\Big)$$

and where $\mathsf{E}\varphi =_{def} \neg\mathsf{U}\neg\varphi$*.*

The abbreviation $\mathsf{AllV}(\varphi,\psi)$ just expresses the fact that the obligation that φ and the obligation that ψ aggregate under conjunction when interpreting them relative to weak preference models if, for every valuation of the atoms in $Atm(\varphi) \cup Atm(\psi)$, there exists a world corresponding to this valuation. It is easy to check that the formula $\mathsf{O}\neg k \wedge \mathsf{O}(k|k \wedge g) \wedge \mathsf{U}k$ is satisfiable in the class of weak preference models. Indeed, although $\mathsf{O}(k|k \wedge g) \wedge \mathsf{U}k$ implies $\mathsf{O}(k \wedge g)$, $\mathsf{O}(k \wedge g) \wedge \mathsf{O}\neg k$

is satisfiable in the class of weak preference models. As shown in Proposition 6, the latter is not the case in the class of preference models. This highlights that Forrester's gentle murderer paradox is solved in the variant of our logic interpreted over weak preference models by adopting either the solution in which Condition (2) is interpreted via the material implication $k \rightarrow \mathsf{O}(k \wedge g)$ or the solution in which it is interpreted via the conditional obligation $\mathsf{O}(k|k \wedge g)$.

6 From Syntax Dependence to Independence

The general idea behind our ceteris paribus notion of obligation is that φ is obligatory if and only if the utility of a world increases in the direction by the formula φ ceteris paribus, "all else being equal". Following Von Wright (see also [27]), in CPDL we capture this ceteris paribus aspect, by keeping fixed the truth values of the atoms not occurring in φ (i.e., $Atm \setminus Atm(\varphi)$). The fact that the sets of atoms not occurring in two logical equivalent formulas do not necessarily coincide explains why the obligation and permission operators of CPDL are not closed under logical equivalence. A natural way to obtain obligation and permission operators which are closed under logical equivalence consists in defining the ceteris paribus condition by keeping fixed the truth values of the atoms with respect to which φ is independent (i.e., the atoms which do not affect the truth value of φ). This is consistent with Reschers idea that the concept of ceteris paribus should be defined in terms of a concept of independence between formulas [25] (see also [10]). In formal terms, let φ be a propositional formula. Then:

$$M, w \models \mathsf{O}^i\varphi \Longleftrightarrow \forall v, u \in W : \text{ if } M, v \models \varphi \text{ and } v \preceq_{Indep(\varphi)} u \text{ then } M, u \models \varphi$$
$$M, w \models \mathsf{P}^i\varphi \Longleftrightarrow \forall v, u \in W : \text{ if } M, v \models \varphi \text{ and } v \prec_{Indep(\varphi)} u \text{ then } M, u \models \varphi$$

where $Indep(\varphi) = \{p \in Atm : \forall w \in W, w \cup \{p\} \models \varphi \text{ iff } w \setminus \{p\} \models \varphi\}$ denotes the set of atoms with respect to which φ is independent and $w \models \varphi$ means that the valuation w satisfies the propositional formula φ. We use the notation O^i for independence-based "ceteris paribus" obligation and P^i for independence-based "ceteris paribus" permission. Notice that $Atm \setminus Atm(\varphi) \subseteq Indep(\varphi)$. Thus, a ceteris paribus obligation/permission defined in terms of $Atm \setminus Atm(\varphi)$ implies a ceteris paribus obligation/permission defined in terms of $Indep(\varphi)$, as if two worlds are equivalent with regard to $Indep(\varphi)$ then they are also equivalent with regard to $Atm \setminus Atm(\varphi)$. The reason why the previous notions of obligation and permission are closed under logical equivalence is that two logical equivalent formulas are independent with respect to the same set of atomic propositions. Moreover, they have the same truth values at all worlds of a preference model. This feature is captured by the following two validities:

$$\models_{\mathcal{P}} \mathsf{U}(\varphi \leftrightarrow \psi) \rightarrow (\mathsf{O}^i\varphi \rightarrow \mathsf{O}^i\psi)$$
$$\models_{\mathcal{P}} \mathsf{U}(\varphi \leftrightarrow \psi) \rightarrow (\mathsf{P}^i\varphi \rightarrow \mathsf{P}^i\psi)$$

This means that if φ and ψ are universally equivalent, then having the obligation (resp. permission) that φ is the same as having the obligation (resp. permission) that ψ. Since logical equivalence (i.e., equivalence relative to all preference models) is stronger than universal equivalence (i.e., equivalence relative to a specific preference model), we also have the following properties:

$$\models_{\mathcal{P}} \varphi \leftrightarrow \psi \text{ then } \models_{\mathcal{P}} (\mathsf{O}^i\varphi \rightarrow \mathsf{O}^i\psi)$$
$$\models_{\mathcal{P}} \varphi \leftrightarrow \psi \text{ then } \models_{\mathcal{P}} (\mathsf{P}^i\varphi \rightarrow \mathsf{P}^i\psi)$$

7 Conclusion and perspectives

In this paper, we have presented a new approach to deontic logic, based on ceteris paribus preferences, which provides a fresh foundation to the logical analysis of deontic concepts. We have introduced the idea of ceteris paribus preferences and on this basis we have built the semantics of a deontic logic, named CPDL (ceteris paribus deontic logic). We have shown that CPDL not only avoids some deontic paradoxes, but also provides an adequate conceptualisation of obligations and permission, conditioned and unconditioned. In particular, CPDL supports formal models of obligations and permission that match common-sense intuitions and legal language. We have also examined some properties of the resulting logical system showing in particular how it supports for limited aggregation of conjunctions and factual detachment. We provided a connection with knowledge representation in order to compactly represent and reason over the set of obligations and permissions using the CP-net formalism. We are currently working to develop the framework of CPDL in various directions, concerning both theory and applications.

References

1. Allen, T.E.: CP-nets with indifference. In: 2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton). pp. 1488–1495. IEEE (2013)
2. Åqvist, L.: Deontic logic. In: Handbook of philosophical logic, pp. 605–714. Springer (1984)
3. Bienvenu, M., Lang, J., Wilson, N.: From preference logics to preference languages, and back. In: Twelfth International Conference on the Principles of Knowledge Representation and Reasoning (2010)
4. Boutilier, C., Brafman, R.I., Hoos, H.H., Poole, D.: Reasoning with conditional ceteris paribus preference statements. In: Proc. of the 15th UAI. pp. 71–80 (1999)
5. Carmo, J., Jones, A.J.: Deontic logic and contrary-to-duties. In: Handbook of philosophical logic, pp. 265–343. Springer (2002)
6. Chisholm, R.M.: Contrary-to-duty imperatives and deontic logic. Analysis **24**(2), 33–36 (1963)
7. Føllesdal, D., Hilpinen, R.: Deontic logic: An introduction. In: Deontic logic: Introductory and systematic readings, pp. 1–35. Springer (1970)
8. Forrester, J.W.: Gentle murder, or the adverbial samaritan. The Journal of Philosophy **81**(4), 193–197 (1984)

9. Gabbay, D., Horty, J., Parent, X., van der Meyden, R., van der Torre, L.: Handbook of deontic logic and normative systems (2013)
10. Girard, P.: Von wrights preference logic reconsidered (2006)
11. Grandi, U., Loreggia, A., Rossi, F., Saraswat, V.: From sentiment analysis to preference aggregation. In: ISAIM (2014)
12. Grossi, D., Lorini, E., Schwarzentruber, F.: The ceteris paribus structure of logics of game forms. Journal of Artificial Intelligence Research **53**, 91–126 (2015)
13. Hansen, J.: Is there a logic of imperatives. Deontic Logic in Computer Science, Twentieth European summer school in Logic, Language and Information, Germany (2008)
14. Hansson, B.: An analysis of some deontic logics. In: Deontic Logic: Introductory and Systematic Readings, pp. 121–147. Springer (1970)
15. Hansson, S.O.: Preference-based deontic logic (pdl). Journal of Philosophical Logic **19**(1), 75–93 (1990)
16. Hansson, S.O.: The structure of values and norms. Cambridge University Press (2001)
17. Hilpinen, R.: Deontic logic: Introductory and systematic readings, vol. 33. Springer Science & Business Media (2012)
18. Hilpinen, R., McNamara, P.: Deontic logic: A historical survey and introduction. Handbook of deontic logic and normative systems **1**, 3–136 (2013)
19. Jones, A.J.: Deontic logic and legal knowledge representation. Ratio Juris **3**(2), 237–244 (1990)
20. Li, M., Kazimipour, B.: An efficient algorithm to compute distance between lexicographic preference trees. In: Proc. of the 27th IJCAI 2018. pp. 1898–1904 (2018)
21. Loreggia, A., Mattei, N., Rossi, F., Venable, K.B.: Preferences and ethical principles in decision making. In: Proc. 1st AIES (2018)
22. Loreggia, A., Mattei, N., Rossi, F., Venable, K.B.: On the distance between CP-nets. In: Proc. of the 17th AAMAS. pp. 955–963 (2018)
23. Loreggia, A., Mattei, N., Rossi, F., Venable, K.B.: CPMetric: Deep siamese networks for metric learning on structured preferences. In: Artificial Intelligence. IJCAI 2019 International Workshops. pp. 217–234. Springer (2020)
24. Prakken, H., Sergot, M.: Dyadic deontic logic and contrary-to-duty obligations. In: Defeasible deontic logic, pp. 223–262. Springer (1997)
25. Rescher, N.: Semantic foundations for the logic of preference. The logic of decision and action pp. 37–62 (1967)
26. Ross, A.: Imperatives and logic. Philosophy of Science **11**(1), 30–46 (1944)
27. Van Benthem, J., Girard, P., Roy, O.: Everything else being equal: A modal logic for ceteris paribus preferences. Journal of philosophical logic **38**(1), 83–125 (2009)
28. Von Wright, G.H.: Deontic logic. Mind **60**(237), 1–15 (1951)
29. Von Wright, G.H.: The logic of preference (1963)
30. Von Wright, G.H.: A new system of deontic logic. In: Deontic Logic: Introductory and Systematic Readings, pp. 105–120. Springer (1970)
31. Von Wright, G.H.: The logic of preference reconsidered (1972)
32. Wang, H., Shao, S., Zhou, X., Wan, C., Bouguettaya, A.: Preference recommendation for personalized search. Knowledge-Based Systems **100**, 124–136 (2016)
33. Weinberger, C., Weinberger, O.: Logik, semantik, hermeneutik (1980)