A Novel Approach For 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. It allows to introduce notions of conditional/unconditional obligation and permission that are interpreted relative to this semantics. We show how obligations and permissions can be represented compactly using existing preference frameworks from the artificial intelligence area.

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

1 Introduction

Artificial agents are used to automate task 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 [10]. 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. 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 [7, 15]. Deontic logic is a set of formal tools, usually based on modal logic [2, 5] which could be compositionally integrated with other logical formalism [3, 8]. In this work, we provide the semantics for a deontic logic based on the intuitive idea that obligations and permissions consist in preferences over worlds. 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. This approach allows to adopt well-known preference frameworks and algorithms to reason about them. For instance, given a set of obligations and permissions one

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

can compare them with an agent's preferences to understand how similar they are [9, 11, 12] and whether the agent is deviating from a desired behaviour. Our work is based on the idea of ceteris paribus preference originally introduced by Georg Henrik Von Wright [16, 17]. 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. 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. We call *preference model* 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 $\preceq$. Given $M = (W, \preceq)$ 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. 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. 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 $Atm \setminus X$. In this work we introduce how some deontic operators can be mapped to preference models and we focus on *CP-nets* [4]. They are a compact representation of conditional preferences over ceteris paribus semantics. Due to the lack of space, we refer to the literature [4, 1, 11] for more information.

2 Ceteris Paribus Deontinc Logic

The ceteris paribus deontic logic - CPDL$^+$ has the so-called *universal* modal operator which allows us to capture factual detachment of obligations.

Definition 1. *$\mathcal{L}_{\mathrm{CPDL}^+}(Atm)$ 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 $p \in \mathcal{L}_{\mathrm{CPDL}^+}$, if $\varphi, \psi \in \mathcal{L}_{\mathrm{CPDL}^+}$ then $\neg\varphi, \varphi \wedge \psi \in \mathcal{L}_{\mathrm{CPDL}^+}$, if $\varphi, \psi \in \mathcal{L}_{\mathrm{CPDL}^+}$ then $\mathsf{O}\varphi, \mathsf{P}\varphi, \mathsf{O}(\psi|\varphi), \mathsf{P}(\psi|\varphi)$, $\mathsf{U}\varphi \in \mathcal{L}_{\mathrm{CPDL}^+}$.*

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". 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 $\mathcal{L}_{\mathrm{CPDL}^+}(Atm)$ are defined as follows:

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

Definition 2 (Truth Conditions). *Let $M = (W, \preceq)$ be a preference model, let $w \in W$ and let $Atm_{-\varphi} = Atm \setminus Atm(\varphi)$ where $Atm(\varphi)$ is the set of atoms from Atm occurring in φ. Then:*

- $M, w \models p \iff p \in w$
- $M, w \models \neg\varphi \iff M, w \not\models \varphi$
- $M, w \models \varphi \wedge \psi \iff M, w \models \varphi$ *and* $M, w \models \psi$
- $M, w \models \mathsf{O}\varphi \iff \forall v, u \in W :$ *if* $M, v \models \varphi$ *and* $v \preceq_{Atm_{-\varphi}} u$ *then* $M, u \models \varphi$
- $M, w \models \mathsf{P}\varphi \iff \forall v, u \in W :$ *if* $M, v \models \varphi$ *and* $v \prec_{Atm_{-\varphi}} u$ *then* $M, u \models \varphi$
- $M, w \models \mathsf{U}\varphi \iff \forall v \in W : M, v \models \varphi$
- $M, w \models \mathsf{O}(\psi|\varphi) \iff \forall v, u \in ||\psi||_M :$ *if* $M, v \models \varphi$ *and* $v \preceq_{Atm_{-\varphi}} u$ *then* $M, u \models \varphi$
- $M, w \models \mathsf{P}(\psi|\varphi) \iff \forall v, u \in ||\psi||_M :$ *if* $M, v \models \varphi$ *and* $v \prec_{Atm_{-\varphi}} u$ *then* $M, u \models \varphi$

In other words, $\mathsf{O}\varphi$ means that, for every two possible worlds that are $Atm_{-\varphi}$-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 $Atm_{-\varphi}$-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 $\varphi \in \mathcal{L}_{\mathrm{CPDL}^+}(Atm)$ 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 $\varphi \in \mathcal{L}_{\mathrm{CPDL}^+}(Atm)$ is satisfiable relative to the class of preference models iff, there exists a preference model M and a world w in M, such that $M, w \models \varphi$.

The proposed model has several interesting properties that we list here:

- restricting our model to obligations and permissions that are stated only on atoms, then the induced preference model can be represented compactly by a CP-net;
- 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;
- if φ, ψ are conjunctive clauses and $Atm(\varphi) \cap Atm(\psi) = \emptyset$ then: $\models_{\mathcal{P}} (\mathsf{O}\varphi \wedge \mathsf{O}\psi) \rightarrow \mathsf{O}(\varphi \wedge \psi)$ and $\models_{\mathcal{P}} (\mathsf{P}\varphi \wedge \mathsf{P}\psi) \rightarrow \mathsf{P}(\varphi \wedge \psi)$;
- if φ is obligatory then it is also permitted;
- if the condition of a conditional obligation/permission is necessarily true then the obligation/permission is detached and becomes unconditional;
- CPDL$^+$ does not encounter Ross's paradox [14].

3 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 [15]), in CPDL$^+$ we capture this ceteris paribus aspect, by keeping fixed the truth values of the atoms not occurring in φ (i.e., $Atm_{-\varphi}$). 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 Rescher's idea that the concept of ceteris paribus should be defined in terms of a concept of independence between formulas [13] (see also [6]).

4 Conclusion and perspectives

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, named CPDL^+. 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., et al.: From preference logics to preference languages, and back. In: KR. pp. 9–13 (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. Gabbay, D., Horty, J., Parent, X., van der Meyden, R., van der Torre, L.: Handbook of deontic logic and normative systems (2013)
6. Girard, P.: Von wright's preference logic reconsidered (2006)
7. Grossi, D., Lorini, E., Schwarzentruber, F.: The ceteris paribus structure of logics of game forms. Journal of Artificial Intelligence Research **53**, 91–126 (2015)
8. Jones, A.J.: Deontic logic and legal knowledge representation. Ratio Juris **3**(2), 237–244 (1990)
9. 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)
10. Loreggia, A., Mattei, N., Rossi, F., Venable, K.B.: Preferences and ethical principles in decision making. In: Proc. 1st AIES (2018)
11. 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)
12. 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)
13. Rescher, N.: Semantic foundations for the logic of preference. The logic of decision and action pp. 37–62 (1967)
14. Ross, A.: Imperatives and logic. Philosophy of Science **11**(1), 30–46 (1944)

15. 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)
16. Von Wright, G.H.: The logic of preference (1963)
17. Von Wright, G.H.: The logic of preference reconsidered (1972)

Fig. 1: EKAW - Poster image.