Understanding Deontics from a Preference Perspective

Fenrong Liu

Tsinghua University, Beijing, China

27 March 2010, Hokkaido University, Japan

Outline

This is a report on my recent joint work with **Johan van Benthem** and **Davide Grossi**.

Betterness and Obligation

> *" [...] to assert that a certain line of conduct is [...] absolutely right or obligatory, is obviously to assert that more good or less evil will exist in the world, if it is adopted, than if anything else be done instead." [Moore, Principia Ethica, 1903]*

OBSERVATION 1: No obligation without an order on the "possible states of the world".

Example: Dyadic obligations

Dyadic obligations of the type "it is obligatory that φ under condition ψ" are interpreted by making use of an 'ideality relation' and the notion of maximality:

$$\mathcal{M}, s \models \mathbf{O}(\varphi \mid \psi) \Longleftrightarrow \mathit{Max}(||\psi||_{\mathcal{M}}) \subseteq ||\varphi||_{\mathcal{M}} \tag{1}$$

where $||.||_{\mathcal{M}}$ denotes the truth-set function of $\mathcal{M}$ and $\mathcal{M}$ is a model built on a Kripke frame $\mathcal{F} = (S, \preceq)$. In this frame the states in S are ordered according to the ideality relation $\preceq$.

The deontic notions of obligation, permission and prohibition can be naturally made sense of in terms of an "ideality" ordering $\preceq$ on possible worlds.

Betterness and Priority

OBSERVATION 2: The betterness relation between states is, often, derived from some kind of explicit *coding* of what is better in terms of relevant properties.

As the following quote illustrates in a lively manner:

> *"It is good for a man not to touch a woman. But if they cannot contain, let them marry: for it is better to marry than to burn." [St. Paul, Ch. 7]*

In the terminology of deontic logic, this is a typical **contrary-to-duty** structure (Prakken and Sergot, 1996) expressing what states are best, what states are best among the non-best ones, and so on, up to a finite depth.

Our Plan

In the following sections we will discuss the above type of structures in the light of notions and results developed in recent work in preference logic, and illustrate them by formalizing a classical example of CTD obligations.

Basic definitions: P-sequence

Definition

Let $\mathcal{L}(\mathbf{P})$ be a propositional language, S a non-empty set of states and $\mathcal{I} : \mathbf{P} \longrightarrow 2^S$ a valuation function. A P-sequence for $\mathcal{I}$ is a tuple $\mathcal{B}^{\mathcal{I}} = \langle B, \prec \rangle$ where:

- $B \subseteq \mathbf{P}$ and $|B| < \omega$;
- $\prec$ is a strict linear order on B(Irreflexive, transitive, antisymmetric and total);
- for all $\varphi, \psi \in B$, $\varphi \prec \psi$ iff $||\psi||_{\mathcal{I}} \subset ||\varphi||_{\mathcal{I}}$.

where $||\varphi||_{\mathcal{I}}$ denotes the truth-set of φ according to $\mathcal{I}$.

Deriving preferences from P-sequences

Definition

Let $\mathcal{B} = \langle B, \prec \rangle$ be a P-sequence, S a non-empty set of states and $\mathcal{I} : \mathbf{P} \longrightarrow 2^S$ a valuation function. The preference relation $\preceq_{\mathcal{B}}^{IM} \subseteq S^2$ is defined as follows:

$$w \preceq_{\mathcal{B}}^{IM} w' \quad := \quad \forall \varphi \in B : w \in ||\varphi|| \Rightarrow w' \in ||\varphi||. \tag{2}$$

where *IM* is just a mnemonics for 'implication'.
Given a P-sequence $\mathcal{B}$ for a valuation $\mathcal{I}$, Formula 2 generates also a Kripke model $\mathcal{M}_{\mathcal{B}}^{IM} = (S, \preceq_{\mathcal{B}}^{IM}, \mathcal{I})$.

There are various definitions of deviation in the literature.

P-sequence and CTDs

J. Forrester. Gentle murder, or the adverbial samaritan. *Journal of Philosophy*, 81:193-197, 1984.

Example (Gentle murder)

"Here is the problem: Let us suppose a legal system which forbids all kinds of murder, but which considers murdering violently to be a worse crime than murdering gently. [. . .] The system then captures its views about murder by means of a number of rules, including these two:

- It is obligatory under the law that Smith not murder Jones.
- It is obligatory that, if Smith murders Jones, Smith murders Jones gently."

Assuming Anderson's reductionistic perspective, the scenario partitions all possible states in three classes:

- class I_1, in which Smith does not murder Jones.
- class I_2, in which Smith murders Jones gently. Note that it is contained in $\neg I_1 = V_1$)
- class $\neg I_2$, in which Smith murders Jones and he does not do it gently. Note that it is also contained in $\neg I_1 = V_1$).

We thus have the P-sequence $\mathcal{B}$ such that $I_2 \prec I_1$. Such P-sequence is sufficient to order the states.

Anderson's Reduction

The idea is to reduce deontic logic formulae, such as $\mathbf{O}\varphi$, to alethic modal logic $\Box$-formulae containing a designated violation or ideality constant, that is:

$$\mathbf{O}\varphi := \Box(\neg\varphi \rightarrow \mathsf{V}) \quad (3)$$

$$\mathbf{O}\varphi := \Box(\mathsf{I} \rightarrow \varphi) \quad (4)$$

To sum up, the intuition behind a P-sequence $p_1, \ldots, p_n$ for a given interpretation function is that each atom p_i gives rise to a bipartition $\{\mathcal{I}(p_i), -\mathcal{I}(p_i)\}$ of the domain of discourse S, and the more we move towards the right-hand side (i.e., the bottom) of the sequence the more atoms p_i are falsified.

A similar analysis can be applied to other puzzles in deontic logic.

Chisholm's paradox

1. It ought to be that Smith refrains from robing Jones.
2. Smith robs Jones.
3. If Smith robs Jones, he ought to be punished for robbery.
4. It ought to be that if Smith refrains from robbing Jones he is not punished for robbery.

A natural and consistent interpretation of the Chisholm's scenario in terms of classification and preference goes as follows:

1. It is most ideal that Smith refrains to rob Jones;
2. Smith robs Jones;
3. The most ideal states under the assumption that Smith robs Jones are states in which Smith is punished;
4. It is most ideal that Smith is not punished.

Modal Preference Logic: Language

Language $\mathcal{L}(\forall, \preceq)$ is built from a countable set **P** of atoms according to the following BNF:

$$\mathcal{L}(\forall, \preceq) : \varphi ::= p \mid \top \mid \neg\varphi \mid \varphi \wedge \varphi \mid [\preceq]\varphi \mid [\forall]\, \varphi$$

where $p \in \mathbf{P}$. Modal duals and Boolean operators are defined as usual. Intuitively, $[\preceq]$ quantifies over all states which are at least as good as the current one, and $[\forall]$ over all states (universal modality).

It was proposed first in [Boutilier 1993, 1994], later used in [Halpern 1997], recently developed in [Girard 2008], [Liu 2008] and [Roy, 2008].

Models

Definition (Models)

A model for $\mathcal{L}(\forall, \preceq)$ on the set of atoms **P** is a tuple $\mathcal{M} = \langle S, \preceq, \mathcal{I} \rangle$ where:

- S is a non-empty set of states;
- $\preceq$ is a conversely well-founded total preorder over S; (reflexive, transitive, connected, no infinite ascending chain.)
- $\mathcal{I} : \mathbf{P} \longrightarrow 2^S$.

As usual, we define $s \prec s'$ as $s \preceq s'$ and $s' \not\preceq s$.

Truth Definition

Definition

Let $\mathcal{M} \in \mathbb{M}$. The satisfaction of a formula $\varphi \in \mathcal{L}(\forall, \preceq)$ by a pointed model $(\mathcal{M}, s)$ is inductively defined as follows:

$$\begin{aligned} \mathcal{M}, s \models p &\iff w \in \mathcal{I}(p) \\ \mathcal{M}, s \models [\preceq]\varphi &\iff \forall s' \in S \text{ s.t. } s \preceq s' : \mathcal{M}, s' \models \varphi \\ \mathcal{M}, s \models [\forall]\,\varphi &\iff \forall s' \in S : \mathcal{M}, s' \models \varphi \end{aligned}$$

The standard Boolean clauses are omitted.

Axiomatization

The logic is axiomatized as follows, where $i \in \{\preceq, \forall\}$:

(**Prop**)	propositional tautologies
(**K**)	$[i](\varphi_1 \to \varphi_2) \to ([i]\varphi_1 \to [i]\varphi_2)$
(**T**)	$[i]\varphi \to \varphi$
(**4**)	$[i]\varphi \to [i][i]\varphi$
(**5**)	$\neg[\forall]\varphi \to [\forall]\neg[\forall]\varphi$
(**.3**)	$(\langle\preceq\rangle\varphi \wedge \langle\preceq\rangle\psi) \to (\langle\preceq\rangle(\varphi \wedge \langle\preceq\rangle\psi) \vee \langle\preceq\rangle(\varphi \vee \psi) \vee \langle\preceq\rangle(\psi$
(**Incl**)	$[\forall]\varphi \to [\preceq]\varphi$
(**Dual**)	$\langle i\rangle\varphi \leftrightarrow \neg[i]\neg\varphi$

The logic is an extension of **S4.3** with the universal modality.

Completeness

Theorem (Strong completeness)

The logic above is sound and strongly complete with respect to the class of total pre-orders.

Defining 'best'

Our logic is quite expressive. The very first semantics for dyadic deontic logic ([Hansson, 1969]) interpreted formulae $\mathbf{O}(\varphi \mid \psi)$ as "all the best ψ-states are φ" . Within our logic, a maximality operator can be defined as follows:

$$[\text{Best}(\psi)]\,\varphi \quad := \quad [\forall]\,(\psi \rightarrow \langle \preceq \rangle (\psi \wedge [\preceq](\psi \rightarrow \varphi)) \tag{5}$$

That is, the best ψ-states are φ if and only if, for all states, either they are not ψ or there is a better ψ-state such that all states above it are either not ψ or φ.

Example (Gentle murder (continued))

Consider the P-sequence for valuation $\mathcal{I}$ of the Gentle murder introduced above, and let $\preceq_{\mathcal{B}}^{IM}$ be the total pre-order generated by that sequence. We have that, for any state s in the model $\mathcal{M}_{\mathcal{B}}^{IM} = (S, \preceq_{\mathcal{B}}^{IM}, \mathcal{I})$:

$$\begin{aligned} \mathcal{M}_{\mathcal{B}}^{IM}, s &\models [\mathrm{Best}(\top)]\ \neg \mathsf{V}_1 \\ \mathcal{M}_{\mathcal{B}}^{IM}, s &\models [\mathrm{Best}(\mathsf{V}_1)]\ \mathsf{I}_2 \\ \mathcal{M}_{\mathcal{B}}^{IM}, s &\models [\mathrm{Best}(\mathsf{V}_2)]\ \mathsf{V}_2. \end{aligned}$$

From Static to Dynamics

Various Kinds of Dynamics

- Changing information, conditional obligation, betterness order stays the same.
- Changing evaluation of worlds locally, maybe by a command(see Yamada 2006, 2007, etc.), at betterness change level.
- Changing the whole norm system: at level of changing the priority structure.

Two Level Dynamics

In the current framework, we can handle dynamical changes that are located both at the level of P-sequences or on the level of possible worlds:

- at the level of P-sequence, we can think of dynamic actions of *adding* a priority, *deleting* a priority, etc.
- on the level of possible worlds, there are various operations which change the ordering over possible worlds.

The operations that were considered in [Liu, 2008] for those two kinds of dynamics apply naturally here.

For Instance: Upgrade on the level of possible worlds

Definition (Upgrade)

The *(radical) upgrade* operation $\Uparrow\varphi$ is defined as:

$$(?\varphi; \preceq; ?\varphi) \cup (?\neg\varphi; \preceq; ?\neg\varphi) \cup (?\neg\varphi; \top; ?\varphi).$$

where $?$ and $;$ are the standard relational operations of test and sequencing, and $\top$ denotes the universal relation.

After the new proposition φ has been incorporated, the upgrade places all φ-worlds on top of all $\neg\varphi$-worlds, keeping all other comparisons the same. Here, besides cutting links between the φ-worlds and $\neg\varphi$-worlds, new betterness links may be added by the disjunct $(?\neg\varphi; \top; ?\varphi)$.

Connecting the Two Levels

Definition

Let F: $(\mathcal{B}, \varphi) \rightarrow \mathcal{B}'$, with $\mathcal{B}$, $\mathcal{B}'$ two P-sequences, and φ a new formula not occurring in $\mathcal{B}$. Let σ: $(\preceq, \varphi) \rightarrow \preceq'$, where $\preceq$ and $\preceq'$ are betterness relations over possible worlds. We say that F ***induces the map*** σ, given a definition for deriving betterness relations from P-sequence (e.g., Definition 2), if, for any P-sequence $\mathcal{B}$ and new formula φ, $\sigma(\preceq_{\mathcal{B}}, \varphi) = \preceq_{F(\mathcal{B},\varphi)}$.

Real Example

Example (Obama and global climate change)

With the issue of global climate change, there were two alternative approaches the Obama administration could take to regulate greenhouse gases, like carbon dioxide emissions from coal plants. One was to craft regulations under existing legal authority, say, by the Clean Air Act. The other was to work with Congress on the enactment of legislation to address climate change. The Obama administration faced great difficulties when working with Congress, so they shifted to the first alternative. On April 17, 2009, the Environmental Protection Agency(EPA) officially designated carbon dioxide and five other heat-trapping gases to be dangerous pollutants.
——Recomposed from *The New York Times* (April 17, 2009)

Analysis

The two alternative approaches reflects two options of making changes:

- Enacting a new legislation. It is tantamount to adding a new proposition to the top of the existing system of norms (the given P-sequence).
- Making use of the existing legal authority. This can be thought of one way of changing the underlying betterness ordering.

Theorem (Correspondence of two-level dynamics)

Given a betterness relation $\preceq$ derived from $\mathcal{B} = (B, \prec)$ by Definition 2, action $\Uparrow\varphi$ on $\preceq$ is induced by prefixing the P-sequence $\mathcal{B}$ with a new formula φ. More precisely, the following diagram commutes:

$$\begin{array}{ccc} \mathcal{B} & \xrightarrow{\varphi;\mathcal{B}} & \varphi;\mathcal{B} \\ \Big\downarrow{\scriptstyle IM} & & \Big\downarrow{\scriptstyle IM} \\ (S, \preceq^{IM}_{\mathcal{B}}) & \xrightarrow{\Uparrow\varphi} & (S, \Uparrow\varphi(\preceq^{IM}_{\mathcal{B}})) \end{array}$$

Proof

Proof.

We need to prove the following equivalence: $s \preceq^{IM}_{A;\mathcal{B}} s'$ iff $s, \Uparrow A(\preceq^{IM}_{\mathcal{B}}) s'$. [$\Leftarrow$]. After $\Uparrow A$ the relation between s and s' becomes this: $\Uparrow \varphi(\preceq^{IM}_{\mathcal{B}}) := (?\varphi; \preceq^{IM}_{\mathcal{B}}; ?\varphi) \cup (?\neg\varphi; \preceq^{IM}_{\mathcal{B}}; ?\neg\varphi) \cup (?\neg\varphi; \top; ?\varphi)$. In terms of a relation between arbitrary worlds s' and s, the above three cases give the implication $s \in ||\varphi|| \rightarrow s' \in ||\varphi||$. By $s \preceq^{IM}_{\mathcal{B}} s'$, we also have that $\forall \psi \in \mathcal{B}$: $s \in ||\psi|| \rightarrow w' \in ||\psi||$. Hence $\forall \psi \in \varphi; \mathcal{B}$: $s \in ||\psi|| \rightarrow s' \in ||\psi||$: i.e., $s \preceq^{IM}_{\varphi;\mathcal{B}} s'$. [$\Rightarrow$]. Assume that $s \preceq^{IM}_{\varphi;\mathcal{B}} s'$, i.e., $\forall \psi \in \varphi; \mathcal{B}$: $s \in ||B\psi|| \rightarrow s' \in ||\psi||$. In particular, it cannot the case that $s \in ||\varphi|| \wedge s' \notin ||\varphi||$. Thus, out of all pairs in the given relation R, those satisfying $(?\varphi; \top; ?\neg\varphi)$ can no longer occur. This is precisely how we defined the relation $s \Uparrow \varphi(\preceq^{IM}_{\mathcal{B}}) s'$. □

Connection to Norm Change

The dynamic aspects of norms—The *norm change* problem—have recently gained much attention from researchers in deontic logic, legal theory and multi-agent systems.

Two main approaches:

- In syntactic approaches—inspired by legal practice—norm change is an operation performed directly on the explicit provisions in the "code" of the normative system ([Governatori and Rotolo, 2008],[Boella, Pigozzi and van der Torre, 2009]).
- In semantic approaches, norm change follows the dynamic logic update paradigm (e.g.[Aucher, etc., 2009]).

Connection to Norm Change

The above Theorem can be viewed as establishing a precise match between changes at the level of **models** with changes at the level of **syntax** of a normative code, i.e., the P-sequences.

Future Work

- Our analysis has focused on linearly ordered P-sequences and the induced total pre-orders—fitting for CTDs. We would like to look at the case of pre-orders, generalizing the logical machinery presented here to that case.
- Future work will look at more elaborate representations of the syntactic side of our approach, including graph structures of criteria and laws, maintaining a normative syntax-semantics correspondence along the lines of the correspondence Theorem.
- Obligations are often entangled with beliefs in reality. We would like to extend our basic setting with beliefs, and explore the belief dynamics and norm change all together.

References

- G.Aucher, D.Grossi, A. Herzig and E.Lorini. Dynamic context logic. In X.He, J. Horty and E.Pacuit, editors, *Proceedings of LORI 2009*, volume 5834 of LNAI. Springer, 2009.
- G. Boella, G. Pigozzi, and L. van der Torre. Normative framework for normative system change. In Decker P., Sichman J., Sierra C., and Castelfranchi C., editors, *Proceedings of the Eighth International Conference on Autonomous Agents and Multiagent Systems* (AAMAS 2009), 2009.
- C. Boutilier. Conditional logics of normality as modal systems. In *Proceedings of AAAI'90*, pages 594-599, 1990.
- C. Boutilier. Conditional logics of normality: A modal approach. *Artificial Intelligence*, 68:87-154, 1994.

References

- J. Forrester. Gentle murder, or the adverbial samaritan. *Journal of Philosophy*, 81:193-197, 1984.
- P. Girard. *Modal Logic for Belief and Preference Change*. PhD thesis, University of Amsterdam, 2008.
- G. Governatori and A. Rotolo. Changing legal systems: Abrogation and annulment. part 1: Revision and defeasible theories. In Ron van der Meyden and Leendert van der Torre, editors, *Proceedings of the 9th International Conference on Deontic Logic in Computer Science* (DEON 2008), Luxembourg, Luxembourg, July 15-18, 2008, volume 5076 of LNAI, pages 3-18. Springer, 2008.

References

- B. Hansson. An analysis of some deontic logics. *Nous*, 3:373-398, 1969.
- F. Liu. *Changing for the Better. Preferemce dynamics and Agent Diversity*. PhD thesis, University of Amsterdam, 2008.
- G. E. Moore. *Principia Ethica*. Cambridge University Press, 1903.
- H. Prakken and M. Sergot. Contrary-to-duty obligations. *Studia Logica*, 57:91-115, 1996.
- O.Roy. *Thinking before Acting: Intentions, Logic and Rational Choice*,ILLC, University of Amsterdam, 2008.
- T. Yamada. Acts of Commands and Changing Obligations, Inoue,K. and Satoh, K. and Toni, F. editors, *Proceedings of the 7th Workshop on Computational Logic in Multi-Agent Systems* (CLIMA VII)", 2006. Revised version appeared in LNAI 4371, pages 1-19, Springer-Verlag, 2007