AN ARETAIC APPROACH TO DEONTIC LOGIC

Joshua Yarrow [1] Dr. Sara L. Uckelman [2]

[1] Pembroke College, University of Cambridge
jmky2@cam.ac.uk

[2] Department of Philosophy, Durham University
s.l.uckelman@durham.ac.uk

Advances in Modal Logic – August 20, 2024

WHAT WOULD IT LOOK LIKE TO HAVE A DEONTIC LOGIC INSPIRED BY VIRTUE ETHICS?

WHAT WOULD IT LOOK LIKE TO HAVE A DEONTIC LOGIC INSPIRED BY VIRTUE ETHICS?

Specifically, given a multi-modal labelled transition system:

$$\mathfrak{M} = \langle W, R, V \rangle \text{ where } R \subseteq W \times A \times W$$

$$\alpha ::= a \mid \alpha \& \beta \mid \alpha + \beta$$

$$\varphi ::= \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid [\alpha]\varphi$$

what would truth conditions look like for these operators?

- $O(\alpha)$ for Obligation (You Must)
- $P(\alpha)$ for Permission (You May)
- $F(\alpha)$ for Forbiddance (You May Not)

WHAT IS VIRTUE ETHICS?

WHAT IS VIRTUE ETHICS?

Key Points:

- Human characteristics are the primary subjects of morality. (As opposed to rules, duties, consequences, etc.)
- ‘Virtues’ are morally good characteristics.
- ‘Vices’ are morally bad characteristics.
- Many interpretations, both ancient and contemporary.

WHAT IS VIRTUE ETHICS?

Definition (Golden Mean Theory)

In Aristotelian virtue ethics, each characteristic has a ‘vice of excess’, a ‘vice of deficiency’, and a ‘golden mean’ possessed by the ‘virtuous agent’.

Common example characteristics: generosity, courage, temperance, wit.

EXAMPLE: GENEROSITY

Balthazar instantiating excess generosity by gifting a Boeing 777 to the Baby Jesus. (Image created by OpenAI Dall·E 2)

STARTING DEFINITIONS

Definition (Characteristics)

Let C be an N-tuple of specified characteristics. Then $\mathrm{Char} : W \times A \to [-1, 1]^N$ maps world-action pairs to a deficiency-mean-excess continuum for each characteristic in C.

Definition (Criterion of Right Action)

Rosalind Hursthouse: “An action is *right* iff it is what a virtuous agent would characteristically (i.e., acting in character) do in the circumstances."
[1, p. 28]

CRA can help us define obligation if we take ‘right’ to mean ‘obligatory’.

TOWARDS DEFINING OBLIGATION

Surely the virtuous agent would only characteristically perform actions which minimise the norm of the $\mathtt{Char}$ vector?

Definition (Least Vicious Actions)

$$\begin{aligned}\mathtt{MinVice}(w) \quad = \quad & \{a \in A \mid \text{there is a } v \in W \text{ such that } wR_a v \\ & \text{and for all } b \in A \text{ where there is a } v' \in W \text{ such that } wR_b v', \\ & \|\mathtt{Char}(w,a)\| \leq \|\mathtt{Char}(w,b)\|\}\end{aligned}$$

But when this set has many elements, would the virtuous agent do all of them or just one?

FIRST ATTEMPTS AT OBLIGATION

This leads us to two initial alternatives.

Definition (Strong Obligation)

$$\mathfrak{M}, w \vDash O^S(a_1 \& \cdots \& a_n) \quad \text{iff} \quad \{a_1, \cdots, a_n\} \subseteq \texttt{MinVice}(w)$$
and there is a $v \in W$ such that $wR_{(a_1 \& \cdots \& a_n)}v$

Definition (Weak Obligation)

$$\mathfrak{M}, w \vDash O^W(a_1 + \cdots + a_n) \quad \text{iff} \quad \{a_1, \cdots, a_n\} = \texttt{MinVice}(w)$$

- Strong obligation is sometimes too strong and weak obligation is sometimes too weak.
- We have falsely assumed that `Char` assignments are independent of the performance of other atomic actions.

TOWARDS DEFINING OBLIGATION (AGAIN)

We need to level up our semantics to work on joint action complexes:

$$\mathtt{Char} : W \times \mathcal{P}(A) \to [-1, 1]^N$$

Which means more auxiliary definitions (yay!):

Definition (Set of Executable Action Complexes)

$$\begin{aligned} \mathtt{ActComplex}(w) \quad = \quad & \{\{a_1, \cdots, a_n\} \in \mathcal{P}(A) \mid \text{there is a } v \in W \\ & \text{such that } wR_{(a_1 \& \cdots \& a_n)}v\} \end{aligned}$$

Definition (Minimal Vice Complexes)

$$\begin{aligned} \mathtt{MinViceComplex}(w) \quad = \quad & \{\mathbf{a} \in \mathtt{ActComplex}(w) \mid \text{for all } \mathbf{b} \in \mathtt{ActComplex}(w), \\ & ||\mathtt{Char}(w, \mathbf{a})|| \leq ||\mathtt{Char}(w, \mathbf{b})||\} \end{aligned}$$

DEFINING OBLIGATION

- Now we can define obligation to combine the benefits of the strong and weak definitions.
- We use a 'Choice Normal Form' that can represent any action expression as a top-level choice between joint action complexes.

Definition (Composite Obligation)

Let $\alpha := ((a_1^1 \& \cdots \& a_{n_1}^1) + \cdots + (a_1^m \& \cdots \& a_{n_m}^m))$ be in Choice Normal Form. Then:

$$\mathfrak{M}, w \vDash O(\alpha) \quad \text{iff} \quad \{\{a_1^1, \cdots, a_{n_1}^1\}, \cdots, \{a_1^m, \cdots, a_{n_m}^m\}\} = \texttt{MinViceComplex}(w)$$

- Essentially weak obligation defined on joint action complexes.
- Free choice is interpreted as strictly exclusive such that $O(\alpha + \beta)$ means 'You ought to do either *just* α or *just* β'.

WHAT ABOUT PERMISSION?

Traditionally permission is the dual of obligation. This cannot work here as we do not have action negation. Looking at it differently:

1. Virtue is acquired gradually by habituation (this is an Aristotelian view)
2. Habituating non-perfect characteristics that are closer to virtue than the agent currently is will still help them to acquire virtue up to a point (assumption)
3. The development of virtue should be morally encouraged
4. Therefore any act whose performance would nudge the agent's characteristic profile towards virtue is permissible

DEFINING PERMISSION

If we define an agential characteristic profile $\texttt{AgentChar} \in [-1,1]^N$ then we can capture this notion of permissibility:

Definition (Improvement-based Permission)

$$\mathfrak{M}, w \vDash P(\alpha) \quad \text{iff} \quad \begin{array}{l} ((a_1^1 \& \cdots \& a_{n_1}^1) + \cdots + (a_1^m \& \cdots \& a_{n_m}^m)) \\ \text{is the Choice Normal Form of } \alpha \\ \text{and for all } \mathbf{a} \in \{\{a_1^1, \cdots, a_{n_1}^1\}, \cdots, \{a_1^m, \cdots, a_{n_m}^m\}\} \\ \text{and all } i \in [\![1, N]\!], |\texttt{Char}(w, \mathbf{a})_i| \leq |\texttt{AgentChar}_i| \end{array}$$

Corollaries:

- The virtuous agent is only allowed to be perfectly virtuous
- $P(\alpha + \beta) \leftrightarrow P(\alpha) \wedge P(\beta)$ is valid

FORBIDDANCE AND THE IMPERMISSIBLE OBLIGATORY

- Unusually, obligation and permission are now independent.
- It is possible to create models where the least vicious, i.e. obligatory, action is still impermissible.
- We believe this may represent a salient moral category – when one must perform an act which will 'tarnish' their moral character.
- Therefore we define the forbidden acts as all *other* impermissible acts.

Definition (Forbiddance)

$$F(\alpha) := \neg P(\alpha) \wedge \neg O(\alpha)$$

Dynamic Virtues

Definition (Exponential Update)

$$\mathfrak{M} \uparrow_W^{\{a_1,\cdots,a_n\}} = \langle W, R, V, \mathtt{Char}, \mathtt{AgentChar}', \tau \rangle$$

where $\tau \in (0,1]$ and
$\mathtt{AgentChar}' := ((1-\tau) \times \mathtt{AgentChar}) + (\tau \times \mathtt{Char}(w, \{a_1,\cdots,a_n\}))$

Conditional Obligation

$O(\alpha \mid \Box\varphi, \Diamond\psi, +\{c_1,\cdots,c_{l^+}\}, -\{d_1,\cdots,d_{l^-}\})$

- Box and Diamond arguments as a propositional outcome filter
- Commitment to specified atomic actions
- Refusal to perform specified atomic actions

FUTURE POSSIBILITIES

Three interesting directions this work could go down:

1. A logic of virtue epistemology [2]
2. Modelling / recognising virtue with neural networks to populate the Char function (reminiscent of RLHF but for each characteristic individually)
3. Multi-agent models

REFERENCES

[1] Rosalind Hursthouse. *On Virtues Ethics*. Oxford University Press, 2000.

[2] Linda Trinkaus Zagzebski. *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge: Cambridge University Press, 1996.