ICRES 2018: International Conference on Robot
Ethics and Standards, Troy, NY, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.018

Virtue Ethics via Planning and Learning

N. S. GOVINDARAJULU, and S. BRINGSJORD and R. GHOSH

Rensselaer AI & Reasoning Lab, Rensselaer Polytechnic Institue, (RPI)
Troy, New York 12180, USA
[*] E-mail: govinn2@rpi.com
www.rpi.edu

We present our previous formalization of virtue ethics $\mathcal{V}_z^f$ based on Zagzebski Exemplarist Virtue Theory and evaluate how well it adheres to a set of conditions laid for virtue ethics out by Alfano.[*] Our formalization is based on planning and learning and is cast in a formal logic, a cognitive calculus (which subsumes a quantified first-order logic), that has been previously used to model robust ethical principles, in both the deontological and consequentialist traditions. Briefly, we find that the formalization largely adheres to Alfano's conditions, but a larger more detailed study is needed.

Keywords: virtue ethics, planning, learning

1. Introduction

While there has been extensive formal, computational, and mathematical work done in the two main camps of ethics, **deontological ethics** ($\mathcal{D}$) and **consequentialism** ($\mathcal{C}$) , there has been little such work done in formalizing and making rigorous **virtue ethics** ($\mathcal{V}$). If $\mathcal{V}$ is to be considered to be on equal footing with $\mathcal{D}$ and $\mathcal{C}$ for the purpose of building morally competent machines, we need to start with formalizing parts of virtue ethics.

What is virtue ethics? One quick way of summarizing virtue ethics is to contrast it with $\mathcal{C}$ and $\mathcal{D}$. In simple forms of $\mathcal{C}$, actions are evaluated based on their **total utlity** to everyone involved. The best action is the action that has the highest total utility. In $\mathcal{D}$, the emphasis is on **inviolable principles**, and reasoning from those principles to whether actions are obligatory, permissible, neutral, etc. In contrast to $\mathcal{D}$ and $\mathcal{C}$, some forms of virtue ethics can be summed up by saying the best action in a situation is the action that a **virtuous person** would do. A virtuous person is defined as a person that has learnt and internalized a diverse set of virtuous habits or traits. For a virtuous person, virtuous acts become second-nature, and hence are performed in many different situations. Note that unlike $\mathcal{D}$ and $\mathcal{C}$, it is not entirely straightforward how one could translate these notions into a form that is precise enough to be realized in machines.

One embroyonic project $\mathcal{V}_z^f$ based on learning has been laid out by us in [1]. The goal in this paper is to evaluate how well $\mathcal{V}_z^f$ adheres to the conditions laid out by Alfano in [2]. Alfano lays out a series of conditions that he considers to be the core of virtue ethics. The conditions are laid below:

Alfano's Hard Core of Virtue Ethics (from [2])

(1) **acquirability** It is possible for a non-virtuous person to acquire some of the virtues.

(2) **stability** If someone possesses a virtue at time t_1, then ceteris paribus she will possess that virtue at a later time t_2.

(3) **consistency** If someone possesses a virtue sensitive to reason r, then *ceteris paribus* she will respond to r in most contexts.

(4) **access** It is possible to determine what the virtues are.

[*] In $\mathcal{V}_z^f$, z stands for "Zagzebski" and f states that the account is formal in nature, and $\mathcal{V}_z$ is the same theory informally presented.

(5) **normativity** *Ceteris paribus*, it is better to possess a virtue than not, and better to possess more virtues than fewer.

(6) **real saints** There is a non-negligible cohort of saints in the human population.

(7) **explanatory power** if someone possesses a virtue, then reference to that virtue will sometimes help to explain her behavior.

(8) **predictive power** if someone possesses a high-fidelity virtue, then reference to that virtue will enable nearly certain predictions of her behavior; if someone possesses a low fidelity virtue, then reference to that virtue will enable weak predictions of her behavior.

(9) **egalitarianism** Almost anyone can reliably act in accordance with virtue.

2. A Quick Overview of $\mathcal{V}_z^f$

$\mathcal{V}_z^f$ is based on *exemplarist virtue theory* $\mathcal{V}_z$ and is cast in the **deontic cognitive event calculus** ($\mathcal{DCEC}$). We first give a brief overview of exemplarist virtue theory below before proceeding to give an encapsulated version of $\mathcal{V}_z^f$.

Exemplarist virtue theory ($\mathcal{V}_z$) builds on the **direct reference theory** (DRT) of semantics and has the emotion of **admiration** as a foundational object. In DRT, the meaning of a word is constructed by what the word points out. For example, to understand the meaning of *"water"*, a person need not understand and possess all knowledge about water. The person simply needs to understand that "water" points to something which is similar to *that* (with *that* pointing to water).

In $\mathcal{V}_z$, moral terms are assumed to be understood similarly. Moral attributes are defined by direct reference when instantiated in exemplars (saints, sages, heroes) that one identifies through admiration. The emotions of admiration and contempt play a foundational role in this theory. Zagzebski posits a process very similar to scientific or empirical investigation, Exemplars are first identified and their traits are studied. Exemplars are then continously further studied to better understand their traits, qualities, etc. The status of an individual as an exemplar can change over time. Below is an informal version that we seek to formalize:

Informal Version $\mathcal{V}_z$

$\mathbf{I}_1$ Agent or person a perceives a person b perform an action α. This observation causes the emotion of admiration in a

$\mathbf{I}_2$ a then studies b and seeks to learn what traits (habits/dispositions) b has.

2.1. *The Background Calculus*

The computational logic we use is the **deontic cognitive event calculus** ($\mathcal{DCEC}$). This logic was used previously in [3,4] to automate versions of the doctrine of double effect $\mathcal{DDE}$, an ethical principle with deontological and consequentialist components. While describing the calculus is beyond the scope of this paper. Dialects of $\mathcal{DCEC}$ have also been used to formalize and automate highly intensional reasoning processes, such as the false-belief task [5] and *akrasia* (succumbing to temptation to violate moral principles).[6a] Arkoudas and Bringsjord[5] introduced the general family of **cognitive event calculi** to which $\mathcal{DCEC}$ belongs, by way of their formalization of the false-belief task. $\mathcal{DCEC}$ is a sorted (i.e. typed) quantified modal logic (also known as sorted first-order modal logic) that includes the event calculus, a first-order calculus used for commonsense reasoning. The calculus has a well-defined syntax and proof calculus; see Appendix A of [3]. The proof calculus is

[a] $\mathcal{DCEC}$ is both *intensional* and *intentional*. There is a difference between intensional and intentional systems. Broadly speaking, extensional systems are formal systems in which the references and meanings of terms are independent of any context. Intensional systems are formal systems in which meanings of terms are dependent on context such as cognitive states of agents, time etc. Modal logics used for modeling beliefs, desires and intentions are considered intensional systems. Please see the appendix in [3] for a more detailed discussion.

based on natural deduction [7], and includes all the introduction and elimination rules for first-order logic, as well as inference schemata for the modal operators and related structures.
While describing $\mathcal{V}_z^f$ and the background calculus $\mathcal{DCEC}$ in detail is beyond the scope of this paper, we briefly list out the two major components below:

Components of $\mathcal{V}_z^f$

$\mathbf{C}_1$ A formalization of emotions, particularly admiration. $\mathcal{V}_z^f$'s formalization of admiration in $\mathcal{DCEC}$ takes the following form: An agent a is said to admire another agent b's action α, if agent a believes the action is a good action.

$$holds(admires(a,b,\alpha,t),t') \leftrightarrow \left[\begin{array}{c} \Theta(a,t') \wedge \\ \mathbf{B}\left(a,t',\left[\begin{array}{l}(a \neq b) \wedge \bar{\mu}(action(\alpha,b),t) > 0 \ \wedge \\ \neg\exists f,t'.\begin{pmatrix} initiates(action(\alpha,b),f,t) \wedge \\ \mu(f,t') < 0 \end{pmatrix}\end{array}\right]\right)\end{array}\right]$$

$\mathbf{C}_2$ A notion of learning traits (and not just simple individual actions). If an agent a admires another agent b for action α in situations σ, then the agent a might learn a trait based on this action α (elaborated below).

2.2. *Learning Traits*

Note that when we look at humans learning virtues by observing others or by reading from texts or other sources, it is not entirely clear how models of learning that have been successful in perception and language processing (e.g. the recent successes of deep learning/differentiable learning/statistical learning) can be be applied. Learning in these situations is from one or few instances or in some cases through instruction and such learning may not be readily amenable to models of learning which require a large number of examples.

The abstract learning method that we will use is **generalization**. If we have a set of set of formulae $\{\Gamma_1,\ldots,\Gamma_n\}$, the generalization of $\{\Gamma_1,\ldots,\Gamma_n\}$, denoted by $g\big(\{\Gamma_1,\ldots,\Gamma_n\}\big)$ is a Γ such that $\Gamma \vdash \wedge\Gamma_i$. See one simple example below:

Example 1

$$\frac{\begin{array}{r}\Gamma_1 = \{talkingWith(jack) \rightarrow Honesty\} \\ \Gamma_2 = \{talkingWith(jill) \rightarrow Honesty\}\end{array}}{\textbf{generalization}\ \Gamma = \{\forall x.talkingWith(x) \rightarrow Honesty\}}$$

One particularly efficient and well-studied mechanism to realise generalization is **anti-unification**. Anti-unification that has been applied successfully in learning programs from few examples.[b] In anti-unification, we are given a set of expressions $\{f_1,\ldots,f_n\}$ and we need to compute an expression g that when substituted with an appropriate term θ_i gives us f_i. E.g. if we are given $hungry(jack)$ and $hungry(jill)$, the anti-unification of those terms would be $hungry(x)$.

Example 2

$$\frac{\begin{array}{r}likes(jill,jack) \\ likes(jill,jim)\end{array}}{\textbf{anti-unification}\ \ likes(jill,x)}$$

In higher-order anti-unification, we can substitute function symbols and predicate symbols. Here P is a higher-order variable.

[b]This discipline known as inductive programming seeks to build precise computer programs from examples.[8]

Example 3

$$\begin{array}{rl} & likes(jill, jack) \\ & loves(jill, jim) \\ \hline \textbf{anti-unification} & P(jill, x) \end{array}$$

2.3. *Defining Traits*

We need agents to learn traits and not just single actions. We define below what it means for an agent to have a trait. First, a situation $\sigma(t)$ is simply a set of formulae that describes what fluents hold at a time t along with other event calculus constraints and descriptions. An action type α is said to consistent in a situation $\sigma(t)$ for an agent a if:

$$\sigma(t) \cup \{happens\big(action(\alpha, a), t\big)\} \nvdash \bot$$

Trait

An agent a is said to have an action type α as a trait if there are at least m situations $\{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ in which there are unique alternatives $\{\alpha_1, \ldots, \alpha_m\}$ available but *instantiations* of α is performed in a large fraction $\gamma \gg 1$of these situations.

2.4. *Learning from Exemplars and Not Just From Examples*

We start with a learning agent l. An agent e is identified as an exemplar by l *iff* the corresponding emotion of admiration is triggered n times or more. A learnt trait is defined below:

Learnt Trait

A learnt trait is simply a situation $\sigma(t)$ and an action type α: $\langle \sigma(t), \alpha \rangle$

Once e is identified, the learner then identifies one or more traits of e by observing e over an extended period of time. Let $\{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ be the set of situations in which instantiations $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ of a particular trait α are triggered. The learner then simply associates the action type α with the generalization of the situations $g(\{\sigma_1, \sigma_2, \ldots, \sigma_n\})$. That is the agent has incorporated this learnt trait:

$$\Big\langle g\big(\{\sigma_1, \sigma_2, \ldots, \sigma_n\}\big), \alpha \Big\rangle$$

For instance, if the trait is *"being truthful"* and is triggered in situations: *"talking with alice,"*, *"talking with bob"*, *"talking with charlie"*; then the association learnt is that *"talking with an agent"* should trigger the *"being truthful"* action type.

2.5. *Example*

We present a simple example. Assume that we have a market place where things that are old or new can be bought and sold. A seller can either honestly state the condition of the item $\{old, new\}$ or not correctly report the state of the item. For an honest seller, we have the following two situations that can be observed (not for easy readability, we omit the the specific item under consideration):

Situation 1

$$\begin{aligned} \sigma_1 &\equiv holds(old, t) \\ \alpha &\equiv happens(utter(old), t) \end{aligned}$$

Situation 2

$$\sigma_2 \equiv holds(new, t)$$
$$\alpha \equiv happens(utter(new), t)$$

The learnt trait is then given below. The trait says that one should always correctly utter the state of the item.

$$\Big\langle holds(x, t), happens(utter(x), t) \Big\rangle$$

3. Evaluation wrt to Alfano's Conditions

How well does the formal system above adhere to Alfano's Conditions? We outline a quick evaluation below (with our explanations emphasized):

Alfano's Hard Core of Virtue Ethics (from [2])

(1) **acquirability** It is possible for a non-virtuous person to acquire some of the virtues. *Learning is central to* $\mathcal{V}_z^f$.

(2) **stability** If someone possesses a virtue at time t_1, then *ceteris paribus* she will possess that virtue at a later time t_2. *Once a trait is learnt, it cannot be lost.*

(3) **consistency** If someone possesses a virtue sensitive to reason r, then *ceteris paribus* she will respond to r in most contexts. *Definition of a trait.*

(4) **access** It is possible to determine what the virtues are. *By examing when the emotion of admiration is consistently triggered, one can isolated virtuous traits.*

(5) **normativity** *Ceteris paribus*, it is better to possess a virtue than not, and better to possess more virtues than fewer. *Admiration is triggered only for actions that are beneficial.*

(6) **real saints** There is a non-negligible cohort of saints in the human population. *Not relevant for machine ethics.*

(7) **explanatory power** If someone possesses a virtue, then reference to that virtue will sometimes help to explain her behavior. *Definition of a trait.*

(8) **predictive power** If someone possesses a high-fidelity virtue, then reference to that virtue will enable nearly certain predictions of her behavior; if someone possesses a low fidelity virtue, then reference to that virtue will enable weak predictions of her behavior. *Definition of a trait.*

(9) **egalitarianism** Almost anyone can reliably act in accordance with virtue. *Definition of a traits and learning of traits.*

4. Discussion

Objection 1 *While the presented approach checks out in terms of formal logic, naturally, it cannot quite escape the anthropocentric nature of virtues on a meta-ethical level.*
Response We agree with this statement. Virtue ethics talks about virtues of *persons*, and because of this a non-person centric version of virtue ethics might not be possible. This is not an issue with our approach but with virtue ethics in general.
Objection 2 *Why were Alfano's conditions chosen?*
Response Our goal is not to formalize or espouse one ethical theory. Our goal is to build mathematical and computational tools for implementing a wide range of ethical theories. For example, given an ethical theory $\mathcal{E}$, this task is made easier when there is a rigorous but still informal version $\mathcal{E}_r$ that we can formalize as $\mathcal{E}_r^f$. On the other hand, we need to have independent ways of evaluating whether formalizations $\mathcal{E}_r^f$ capture $\mathcal{E}$. Alfano's conditions are the most rigorous set of *empirically grounded* claims that we have come across for virtue ethics that serves the evaluation purpose .
Objection 3 *I don't see how the statement 'if someone possesses a virtue sensitive to reason r, then ceteris paribus she will respond to r in most contexts' is captured by the definition of a trait.*
Response By the way a trait is defined, actions performed by an agent in situations that triggers a trait τ will be instances of an action type α, these instances will be largely similar (as they are instantiated from *one* action type).

5. Conclusion

We have presented an initial formalization of a virtue ethics theory in a calculus that has been used in automating other ethical principles in deontological and consequentialist ethics. Many important questions have to be addressed in future research. Among them, are questions about the nature and source of the utility functions that are used in the definitions of emotions. We also need to apply this model to realistic examples and case studies. The lack of such formal examples and case studies is a bottleneck here.

References

1. N. S. Govindarajulu, S. Bringsjord and R. Ghosh, One Formalization of Virtue Ethics via Learning, To be Presented at the 2018 International Association for Computing and Philosophy (IACAP) - Annual Meeting, June 21-23, 2018, Warsaw, Poland, (2018).
2. M. Alfano, *Journal of Philosophical Research* **38**, 233 (2013).
3. N. S. Govindarajulu and S. Bringsjord, On Automating the Doctrine of Double Effect, in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, ed. C. Sierra (Melbourne, Australia, 2017). Preprint available at this url: https://arxiv.org/abs/1703.08922.
4. N. S. Govindarajulu, S. Bringsjord, R. Ghosh and M. Peveler, Beyond the doctrine of double effect: A formal model of true self-sacrifice, International Conference on Robot Ethics and Safety Standards, (2017).
5. K. Arkoudas and S. Bringsjord, Toward Formalizing Common-Sense Psychology: An Analysis of the False-Belief Task, in *Proceedings of the Tenth Pacific Rim International Conference on Artificial Intelligence (PRICAI 2008)*, eds. T.-B. Ho and Z.-H. ZhouLecture Notes in Artificial Intelligence (LNAI)(5351) (Springer-Verlag, 2008).
6. S. Bringsjord, N. S. Govindarajulu, D. Thero and M. Si, Akratic Robots and the Computational Logic Thereof, in *Proceedings of ETHICS • 2014 (2014 IEEE Symposium on Ethics in Engineering, Science, and Technology)*, (Chicago, IL, 2014). IEEE Catalog Number: CFP14ETI-POD.
7. G. Gentzen, Investigations into Logical Deduction, in *The Collected Papers of Gerhard Gentzen*, ed. M. E. Szabo (North-Holland, Amsterday, The Netherlands, 1935) pp. 68–131. This is an English version of the well-known 1935 German version.
8. S.-H. Nienhuys-Cheng and R. De Wolf, *Foundations of Inductive Logic Programming* (Springer Science & Business Media, 1997).