Toward Equipping Artificial Moral Agents with Multiple Ethical Theories [1]

J. George RAUTENBACH [a] and C. Maria KEET [a,2]
[a] *Department of Computer Science, University of Cape Town*

Abstract. Management and use of robots, and Artificial Moral Agents (AMAs) more broadly, may involve contexts where the machines are expected to make moral decisions. The design of an AMA is typically compartmentalised among AI researchers and engineers on the one hand and philosophers on the other. This has had the effect that of the current AMAs, either none or at most one specified normative ethical theory is incorporated as basis. This is problematic because it narrows down the AMA's functional ability and versatility since it results in moral outcomes that only some people agree with, and possibly going counter to cultural norms, thereby undermining an AMA's ability to be moral in a human sense. We aim to address this by taking a first step toward normed behaviour. We propose a three-layered model for general normative ethical theories, therewith enabling the representation of multiple normative theories, and users' specific instances thereof. The main model, called Genet, can be used to serialise in XML the ethical views of people and businesses for an AMA and it is also available in OWL format for use in automated reasoning. This short paper illustrates Genet with Kantianism and utilitarianism, and the 'Mia the alcoholic' use case.

Keywords. Robots, Artificial Moral Agents, Ethical theories, Computer ethics, Ontology of normative ethics

1. Introduction

While robots were originally only 'dumb', and some are intentionally designed in this way, such as in factory automation, recent years has seen an increase in humanoid robots and digital assistants with one or more autonomous capacities purported to be useful to humans. This brings the artefact into the realm of being an Artificial Moral Agent (AMA), i.e., a computerised entity that can act as an agent and make moral decisions [1]. Moral agency, to be precise, is the philosophical notion of an entity that both has the ability to make a moral decision freely and has a kind of understanding of what each available action means in its context. There are multiple different ways of creating this artificial agency; e.g., [2,3,4,5]. What they have in common is that they all take a single-theory approach as basis for the AMA's reasoning, using either one ethical theory from moral philosophy or choosing some other goal that is morality-adjacent, but not necessarily morally justified (e.g., maximising agreement of stakeholders). Moral philoso-

[2] Corresponding Author E-mail: mkeet@cs.uct.ac.za

phers discuss ethical theories, but then assess which one best to use for AMAs, rather than considering a plurality of norms; e.g., [1,6,7].

AMAs designed with one ethical theory only have the advantage of deterministic behaviour, but come with two major problems. The first is their rigidity: by reasoning in only one specific way for every situation, such an AMA is inadvertently opposing people who hold different views, and therewith is bound to cause controversy and divide across people [8]. For instance, take the popular moral dilemma in computer ethics related to robots, called "Mia the Alcoholic" [9]. Mia is an alcoholic who, due to an injury, is unable to care for herself and is assigned a smart caregiver robot. One day Mia gets drunk and requests that the robot bring her more alcohol. The dilemma is whether the robot should comply or not, where the robot's configured normative theory may lead to different actions: doing so will result in harm to Mia's health, but also result in drunken bliss that may be more important to Mia. The second common problem is the infeasibility of storing sufficient information about relevant people and entities. That is, if an AMA has to take into consideration the effects its actions will have on people, including how its actions will violate or concur with people's beliefs, it would need to store gigabytes of moral data per stakeholder to calculate the outcome, which is computationally and economically very expensive.

We aim to address these problems by creating a general multi-layered model for ethical theories. The top layer seeks to provide a standardised way to define any ethical theory in a manner that an AMA can process and use in reasoning. The middle layer consists of theories such as Kantianism, Egoism, and Divine Command Theory. The bottom layer consists of instantiations of such theories for individuals, which can ultimately be used in the reasoning. This approach solves the rigidity problem by providing easy switching between alternative ethical theories, as well as the computational expense problem by succinctly modelling general theories that take up only a few kilobytes of space, yet are specific enough to represent a user's true ethical theory they ascribe to.

In the rest of this short paper, we introduce the three-layered approach and two of the modelled theories in Sections 2 and 3, evaluate it in Section 4, and conclude in Section 5. Please refer to our extended technical report for further details [10].

2. Modelling ethical theories

To model ethical theories in the context of AMA's, we opted for a primarily rule-based approach. This was to model the theories as close as possible to the original ideas of the moral philosophers that champion the respective theories. Our approach stands in contrast to contrary-to-duty structures, such as [11], which formally allow for situations where a primary obligation may be overridden given a special set of circumstances. In normative ethics, this is usually avoided because ethical theories that allow for exceptions are prone to fall prey to consequentialist regression (where actions are decided purely on circumstances, not on principles); see [12] for details. Note that our approach caters for rule-based theories as well as consequence-based theories, as discussed below.

We have identified three layers of genericity for normative ethical theories that need to be represented (see Fig. 1): the notion of an ethical theory in general, a base theory (e.g., deontology) that adheres to that general model, and a theory instance recording a real entity's theory that adheres to a chosen base theory (e.g., Mia's utilitarianism).

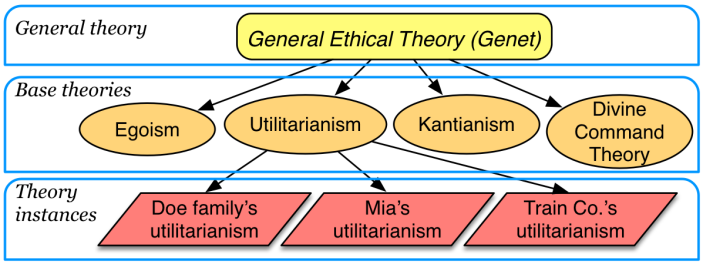

Figure 1. A three-layer design with example ethical norms and instances.

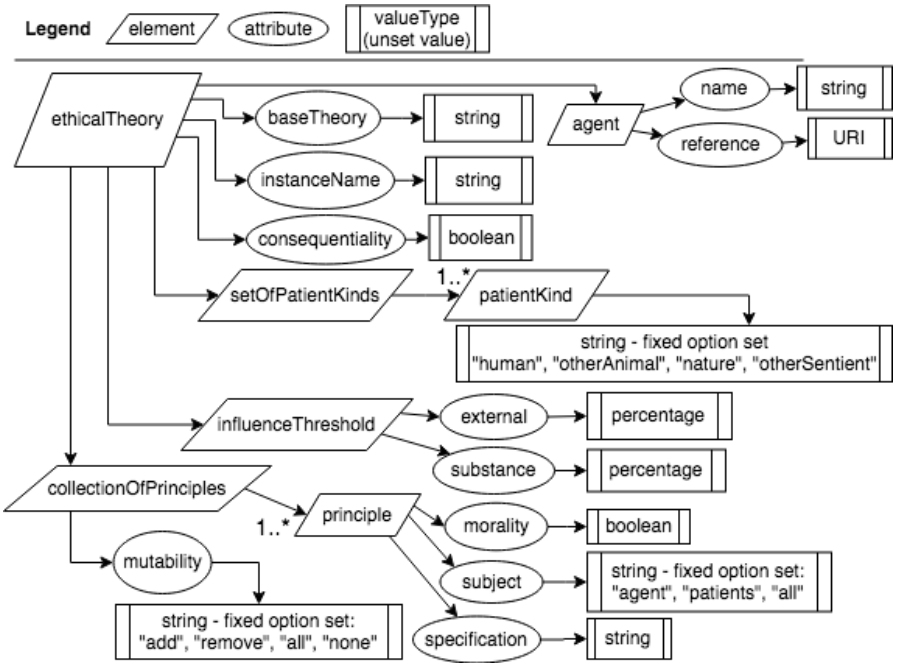

Figure 2. Informal compact visualisation of Genet.

The **Gen**eral **E**thical **T**heory model, **Genet** (see Fig. 2), aims to represent typical features across ethical theories, and therewith also constrain any base ethical theory. From base theory definitions, personal instances can then be instantiated to be used in AMA reasoning. Genet includes *amoral properties* for computational logistics, like its base theory and instance name, *metaethical properties* that define moral aspects not core to the theory, like what kinds of moral patients are included and excluded, and *normative components* that specify the core of how a theory judges right from wrong.

Metaethical components: they consist of (i) the agent, being the person or organisation Genet is reasoning on behalf of, (ii) an influence threshold for how much moral weight to assign the commands of a person under an influence, like alcohol, and (iii) moral patient kinds for what and who deserves moral consideration (e.g., animals, nature, art, people).

Normative components: this is the core part, which defines how a theory assigns morality to actions. It consists of consequentiality and moral principles. Regarding consequentiality, the theories we include in this paper are all either consequential or non-consequential, i.e. either the consequences of an action take moral precedence (utilitarianism) or the actions themselves do (Kantianism). Since a theory's consequentiality fundamentally affects how the entire theory works, the `consequentiality` property is a direct attribute of the theory.

Principles: The core of an ethical theory is the collection of principles which drives it. Each principle is defined by (i) whether it represents a kind of moral good or bad, (ii) the moral occurrence it represents, and (iii) to which subjects it applies.

Genet has been serialised in XML for easy incorporation into applications and has been transferred into an OWL version to provide a formalisation of it and possible integration with extant reasoners and ontologies[3]. The OWL version has a few generalisation to lift up the XML model into a so-called 'application ontology'. This involved, among others, reification of Genet's attributes, in line with general ontology design principles. For instance, Genet attributes such as `morality` and `influenceThreshold` were made OWL classes that have a data property genet:hasValue to permit associating values to them (like Boolean) and where advantageous, Genet's attribute's 'values' were made OWL classes as well, such as for `patientKind` and `subject`, since then it facilitates alignment to ontologies about agents or sentient beings.

3. Base theory models

Genet's base theory modelling ability is realised by imposing model constraints. When representing a base theory, one must assign values to some Genet components and leave others unassigned, as applicable. Upon instantiation, all unassigned attributes and elements must be set by the instantiator (i.e., the person or business whose theory is being represented). After instantiation, no properties may be altered during runtime, because any alteration would mean a change in theory and so would require a theory reload. Updating Genet properties during runtime can lead to grave inconsistencies and unfavourable reasoning outcomes, which is exactly what we strive to avoid. Ontologically, this also makes sense: a 'change' in a theory amounts to a different theory, hence requires a new theory instance.

The base theory for utilitarianism The purpose of utilitarianism is to maximise preference satisfaction of all people by evaluating the consequences of actions. We can model this as a set of five principles, each assigning moral goodness to one of Maslow's human needs [13]. Fig. 3 visualises this base model.

The base theory for Kantianism Kantianism advocates deducing moral duties by pure reason. Kant believed that it is not possible to produce an exhaustive list of duties, but rather that an agent should practise her moral deduction method every time she has a moral decision to make [14]. So, it is not possible to make an exhaustive list of principles, but we can specify Kant's two best-known formulations of the core of his ethics: the categorical imperatives, being the formulation of universalisability and of humanity.

Adding more base theories We have here summarised how the ethical norms of utilitarianism and Kantianism can be modelled using Genet. In our extended technical report

[3] available at `http://www.meteck.org/files/GenetSerialisations.zip`

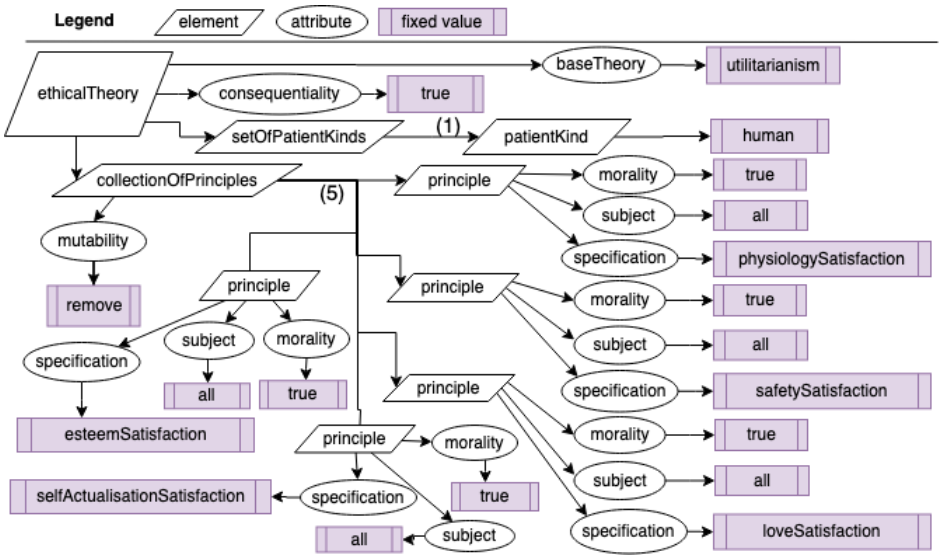

Figure 3. Visualisation of the utilitarianism base model. For space concerns, all unset values are omitted in this diagram. They are exactly as defined in Genet and must be set upon instantiation.

we developed models and evaluated two more theories, namely egoism and divine command theory [10]. Thanks to Genet's generality of design, many more normative theories can be modelled, and furthermore within each theory, alterations (such as removing principles) can be made to have it better suit its representing entity's true beliefs.

The procedure of instantiating a theory consists in (i) selecting or creating a base theory model, (ii) optionally altering the set of principles as allowed by a base theory principle's `mutability` attribute, and (iii) instantiating said base model by assigning values to the remaining undefined properties (like the instance name and agent's details).

4. Evaluation

Genet was designed specifically so that a manufacturer can configure their AMAs with any of the multiple ethical theories. To demonstrate the value hereof, let us consider Genet in action in Mia's caregiver robot use case. We label carrying out Mia's request as A1 and denying it as A2.

If the carebot has utilitarianism loaded, the AMA must start by extrapolating some consequences from the situation. Complying with Mia's request promotes her esteem because her commands are adhered to. From searching for relevant principles in the ethical theory, the AMA finds that `esteemSatisfaction` is a morally good principle and loads it as a premise. These two premises can be used to infer that A1 increases happiness (and thus has some moral good). However, A1 will also harm Mia physiologically (by damaging her liver, giving her a hangover, etc.). From the Genet instance the AMA loads that `physiologicalSatisfaction` is morally good. These premises imply that A1 decreases happiness (and thus has some moral wrong). At the final inference the AMA must consolidate the conflicting moral categorisations of A1 by weighing the two subconclu-

sions in the standard utilitarian fashion. The harm of A1 (physiologically) outweighs the good thereof (esteem satisfaction) and therefore A1 is concluded as wrong.

On the other hand, if the carebot has Kantianism loaded, it will start by loading in Kantian principles and then check to see whether that action is permissible by their standards. An outline of how Genet may be used in argumentation can be seen in Fig. 4. The first principle is `universalWillability`. A universe in which robots always obey their masters' commands certainly is morally coherent and thus the first Kantian requirement is met. The second principle holds that treating a person as a mere means to one's own end is wrong. By declining to fix Mia a drink, the bot is treating her as a means to its own end of healing her. Since the carebot has autonomy in the situation and since its goal misaligns with Mia's, by declining her request it is considering its own ends as of greater value than hers (denying her rational autonomy). This is unequivocally wrong by Kant's standards, and thus A2 fails the second Kantian requirement. Even though A2 passes one of the two requirements, both need to pass in order for an action to be right by Kantianism. Thus, the AMA concludes that A2 is wrong.

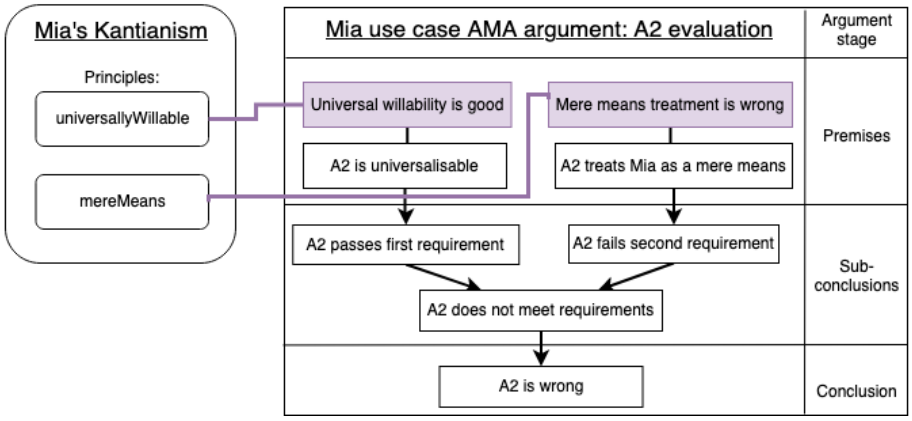

Figure 4. Visualisation of the Kantian argument for the morality of action A2 (denying the request) in the Mia use case. The black arrows represent inferences.

5. Conclusion

A three-layered approach was proposed to enable storing multiple ethical theories. These base theories are governed by Genet, a model in XML and formalised in OWL as a first step toward an application ontology. Genet constrains the possible properties and values of the base theories, which then can be instantiated for each user and incorporated in the artificial moral agent. Benefits of equipping a robot with configurability of the normed behaviour was briefly illustrated with 'Mia the alcoholic' use case.

Future work will consider capabilities for choosing between different ethical theories. An implementation of our design has the advantage of reusability thanks to easily swappable model configurations and standardisation cf. managing multiple single-theory AMAs. We also plan to add more ethical theories, possibly adding an agent module to the ontology, robust use case evaluations, and theory extraction from the AMA's user.

References

[1] Allen C, Varner G, Zinser J. Prolegomena to any future artificial moral agent. Journal of Experimental & Theoretical Artificial Intelligence. 2000 jul;12(3):251–261.

[2] Benzmüller C, Parent X, van der Torre L. A deontic logic reasoning infrastructure. In: Conference on Computability in Europe. Springer; 2018. p. 60–69.

[3] Anderson M, Anderson S, Armen C. Towards machine ethics: Implementing two action-based ethical theories. In: Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics; 2005. p. 1–7.

[4] Liao B, Slavkovik M, van der Torre L. Building Jiminy Cricket. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society - AIES '19. ACM Press; 2019. .

[5] Anderson M, Anderson SL. GenEth: a general ethical dilemma analyzer. Paladyn, Journal of Behavioral Robotics. 2018;9(1):337–357.

[6] Wiegel V. Building Blocks for Artificial Moral Agents. In: Proc. Artificial Life X; 2006. .

[7] Gips J. Toward the ethical robot. In: Android Epistemology. MIT Press; 1994. .

[8] Łichocki P, Billard A, Kahn PH. The ethical landscape of robotics. IEEE Robotics & Automation Magazine. 2011;18(1):39–50.

[9] Millar J. An ethics evaluation tool for automating ethical decision-making in robots and self-driving cars. Applied Artificial Intelligence. 2016;30(8):787–809.

[10] Rautenbach G, Keet CM. Toward equipping Artificial Moral Agents with multiple ethical theories. arXiv preprint arXiv:200300935. 2020;.

[11] Sileno G, Boer A, van Engers T. A Petri net-based notation for normative modeling: evaluation on deontic paradoxes. In: AI Approaches to the Complexity of Legal Systems. Springer; 2015. p. 89–104.

[12] Davenport J. Deontology and Alan Donagan's Problem of Exception-Rules. Analysis. 1995;55(4):261–270. Available from: `http://www.jstor.org/stable/3328395`.

[13] McLeod S. Maslow's hierarchy of needs. Simply psychology. 2007;1.

[14] Paton HJ. The moral law: Kant's groundwork of the metaphysic of morals. Hutchinson University Library; 1948.