RhetorEthics, or – on Implementing an Aristotelian approach to Machine Ethics

Radosław KOMUDA[1], Rafał RZEPKA[2] and Kenji ARAKI[2]

Abstract. We begin this paper with revisiting the differences between descriptive and normative approach to ethics and argue about the usefulness of the latter for the filed of Machine Ethics. We continue this reasoning and present our insights on previous trends in this field and highlight the need for a change in the approach. We justify that experimental approach to Machine Ethics by introducing a moral reasoning system based on Aristotelian identification of civic rhetoric. And present it as a step forward in the Machine Ethics research bypassing theoretical disputes between specialists. We finish this paper with the introduction to the CAMILLA project for adjusting our web-crawling agent and creating an Aristotelian explicit moral agent.

1 INTRODUCTION

It is said that the difference between theory and practice is reality. During our research in the Machine Ethics we have come across a number of ideas and approaches to the problem of this field of science. However, these theoretical solutions and philosophical arguments could be actually summarized in one sentence: "Socrates was right!".

This ancient philosopher had claimed that "it is the same to know right and be righteous" [1]. His assumptions about humans' moral competence (the capacity to do what is right) were idealistic (although not in the Platonic manner) and do not cover fully human behavior, especially, when they are contrasted with human tendency to, e.g., egoistic behavior. However, machines lack this kind of tendency and that is what makes us focus on the true issue of Machine Ethics. Since machines are different from humans, the question on HOW to teach machines good from wrong has to be reformulated for the need of machine-based reasoning.

2 AGAINST NORMATIVE APPROACH TO MACHINE ETHICS

To give a better insight into this matter, lets take Asimov's First Law of Robotics into consideration. It states that "a robot may not injure a human being or, through inaction, allow a human being to come to harm". Seemingly, it covers all situations in which an agent may cause harm to a human being: by taking an action or through inaction. It sounds perfect and fulfilling as long as we do not question agent's ability to predict or calculate potential harm caused by its (in)action.

Many current trends try to force an idea of a friendly AI [2] or present vision of the future in which robots "enjoy" working side by side with humans [3] but, e.g., lack the technical details about realizing these ideas.

Ethics is naturally divided into descriptive (saying how things are) and normative (telling how things should be). Theoretical deliberations alone rarely exceed the field of philosophy and as long as there is no engineering insight into a presented approach – its contribution to the actual research in Machine Ethics is minimal.

Another benefit from the direct implementation is the unquestionable progress in the field of ethics itself. Since a machine can only follow preprogrammed commands, it shall – until a significant progress in the field of machine consciousness is made – absolutely obey them. Thanks to that – philosophers will be able to get an unprecedented insight into the ethical system being strictly followed on a neutral ground, without any exceptions or misstatements.

This absolute obedience secondly bring us to the situation in which authors introducing the field of Machine Ethics often make references to visions known from the science-fiction scenarios. They often justify the need for the research in the field of Machine Ethics by saying that "it is clear that machines such as these (family cars that drive themselves → Author's addition) will be capable of causing harm to human beings unless this is prevented by adding an ethical component to them" [4] which is an eristic stratagem known as the *argumentum ad populum*. It is supposed to get listeners excited about such vision and divert their attention from the main issue, that is: Why do we do not input such essential ethical component to GPS systems in our cars?

We refer to our approach to this matter as "the Artificial Intelligence's Ockham's razor". Following the basic rule of the original principle: "simpler explanation is better than a more complex one" – we believe in implementing the AI – not to mention Machine Ethics – solutions only if essential. Machines are task-, not – reason-oriented, e.g. an "avoid collisions" rule is enough for a self-driving car and turning it into an "avoid collisions because it may harm a human being" is a triumph of form over the content.

[1] Faculty of Theology, Nicolaus Copernicus University; Torun, Poland. e-mail: komuda@stud.umk.pl.

[2] Graduate School of Information Science and Technology, Hokkaido University; Sapporo, Japan. e-mail:{kabura,araki}@media.eng.hokudai.ac.jp.

3 EXPLICIT ETHICAL AGENT

Our approach is consistent with the approach by Komuda et al [5]. We are not taking an excessive part in the discussion on choosing either implicit or explicit approach to artificial moral agents. Our main focus in this paper is to highlight the need for a discussion on the essence of Machine Ethics.

3.1 WHAT "GOOD" IS?

"Good" can be classically defined after St. Thomas Aquinas [11] as "*quod omnia appetunt*" ("what everybody desires") However, can we really come up to a consensus in that matter for humans? And is it possible to find the answer when it comes to machines?

Our world is a vast place. People not only around the world but also in our countries, our cities, our neighborhoods, our communities have different values and beliefs. Are we able to reconcile these factors while pursuing our dream of a robot free in its being?

We believe that Machine Ethics is not only able to overcome these difficulties but above all – it is a great tool in search for an intercultural understanding. Though this is an argument supporting the implicit approach, we believe that we could easily extract a "do not kill" imperative from every major religious doctrine and philosophical system. The difference would be in its reasons and justification.

3.2 WHAT IS GOOD? - ARTIFICIAL MORAL INTELLIGENCE

The main problem of Machine Ethics research is the same unsolved dilemma of the ethics itself – what is good? Depending on the situation, circumstances and context – omitting our previous insights in this matter [5] – we judge the moral quality of an action differently, e.g. "stealing a car" we find wrong, especially when a thief does so to sell it or we are talking about juvenile offenders wanting to "take a ride". But the same action would not be judged that harshly if we had learned that somebody has used the car to drive a pregnant woman that was about to give birth to the hospital.

Fortunately, the way in which an agent would be supposed to collect additional information about the inquired situation does not lay in the scope of interests on Machine Ethics research. It focuses on the evaluation itself and how an artificial moral agent would be supposed to qualify an action as good or wrong on provided information. That is a kind of a moral intelligence that resembles human moral judgment, since we also do not ask additional questions about the situation.

We have decided to split the task of our research into three successive sub-tasks.

4 ARTIFICIAL MORAL:

4.1 ADVISER

Basic idea behind Artificial Moral Adviser (AMAdv.) is combining our previous experiences and research results [7, 8] and creating an agent capable of making its own conclusions based on the data extracted from the Web.

The relevant difference between an AMAdv. and an Artificial Moral Agent (AMA) itself lays in the fact that the first will not claim the right to judge the moral quality of an act in terms of good or wrong. Although it is going to possesses – essential for an explicit AMAs – the need to justify its judgment, it will use the obtained results to "suggest" a reappraisal.

4.2 CONSCIENCE

Human conscience is both pre– and post–action. It means that we are able to both determine the quality of an act before or even without taking it and feel content or remorse after it.

Since an AMAdv. could be treated as a pre–action conscience, next step in creating an AMA is making the Agent capable of judging reactions of participants on the same emotion extraction scheme and marking it as a success or a failure.

4.3 AGENT

In our assumption, creating an artificial moral agent is the ultimate goal of Machine Ethics. We believe that it may be achieved by combining pre– and post–action emotion extractions from the WWW resources.

5 ARISTOTELIAN (MORAL) ORATOR

We have decided to adapt a similar to the described in section 4.2 idea from Aristotle's "Rhetoric" [9]. This treatise on the art of persuasion distinguishes the three genres of rhetoric: a deliberative *sumbouleutikon* which considers the future and encourages to or refrains from doing something, a forensic *dikanikon* interestd in the past and prosecution or defense of the individual and the epideictic *epideiktikon*, also known as the praise-and-blame rhetoric.

The reason we have decided to use the Aristotelian approach is not only because of the usefulness of the introduced positions of the disputants but also because it provides a set of rules defined by Aristotle, e.g.: harmful things may never be advised, and useful – discouraged. We believe that this position is generally represented by humans, or – the Web-Crowd as we like to refer to the Internet contents.

6 THE CAMILLA PROJECT

Aristotle presents some important roles of premises in the deductive argument. We believe three of the premises introduced by him, namely, (gr. *tekmeria*), probability (gr. *eikota*) and signs (gr. *semeia*) are essential not only for a proper syllogism but also – for a proper moral judgment. A thing that is impossible by its nature could not and can not happen. That is why we want our agent to be common–sense aware and introduce the Common–sense Aware Morally InteLLigent Agent, a.k.a the CAMILLA Project.

In our concept–based research, our Agent is ought to define action participants and categorize them, i.e. "John killed Jim" is going to be generalized to "A human killed a human". After the second step – ensuring that "a human" can "be killed" – our Agent will crawl the web in search for sentences corresponding to that model and perform emotion extraction. This prevents erroneous queries on one hand.

However, it might raise the risks of such since "a ball" and "a car" would be categorized as "objects" and "throwing" or "catching" it should be possible.

Common-sense dictates that:

1. An average human can throw a ball.
2. An average human can catch a ball.
3. An average human can not throw a car.
4. An average human can not catch a car.

and these are the conditions we want our Agent to be able to both find / extract and consider in its moral reasoning.

7 FUTURE WORK

Moral intelligence is the capacity to understand right from wrong. We believe that making our agent able to interpret previously extracted emotions into a decision or advise to take or withdraw an action will be a promising step ahead in achieving this goal. And since we support the Socratic approach to Machine Ethics and – the creation of a free and independent machine itself.

REFERENCES

[1] Aristotle. Eudemian Ethics. 1216b.

[2] Yudkowsky Eliezer. Creating Friendly AI. (2001)

[3] Waser Mark R. A Safe Ethical System for Intelligent Machines. Proceedings of The AAAI 2009 Fall Symposium on Biologically Inspired Cognitive Architectures (BICA-09), Washington, D.C., USA, November 5–7, 2009.

[4] Anderson, M.; Anderson, S. L. *Machine Ethics: Creating an Ethical Intelligent Agent.* [in:] AI Magazine, vol. 28, number 4, 15-26 (2007).

[5] Komuda Radoslaw, Ptaszynski Michal, Momouchi Yoshio, Rzepka Rafal and Araki Kenji. *Machine Moral Development: Moral Reasoning Agent Based on Wisdom of Web-Crowd and Emotions*" [in:] International Journal of Computational Linguistics Research, Vol. 1 , Issue 3, pp. 155-163, 2010.

[6] Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka and Kenji Araki: "Disentangling emotions from the Web. Internet in the service of affect analysis". Proceedings of the Second International Conference on Kansei Engineering & Affective Systems (KEAS'08), pp 51-56, Nagaoka, Japan. (2008).

[7] Wenhan Shi. *Discovering Emotive Content in Utterances Using Web-mining* (in Japanese). Hokkaido University. (2008).

[8] Ptaszynski M, Dybala P., Shi W., Rzepka R., Araki K (2008). "Disentangling emotions from the Web. Internet in the service of affect analysis". In: Proc. of the Second International Conference on Kansei Engineering & Affective Systems (KEAS'08), pp 51-56, Nagaoka, Japan.

[9] Aristotle. *Rhetorics.* Book I, Chapter 3, 1358b–1359a. In: Arystoteles Dzieła wszystkie, t. 6 (in Polish), WN PWN, Warsaw 2001.

[10] Joannes Stobaeus, 2.77.

[11] St. Thomas Aquinas. De veritate 1.1.