Oliver Bendel

Towards a Machine Ethics

Bendel, Oliver. Towards a Machine Ethics. In: Technology Assessment and Policy Areas of Great Transitions: Book of Abstracts. 1st PACITA Project Conference, March 13 – 15, 2013. Prague, Czech Republic. pp. 229 – 230. Via http://pacita.strast.cz/en/conference/documents.

There is an increasing use of autonomous machines such as agents, chatbots, algorithmic trading computers, robots of different stripes and unmanned ground or air vehicles. They populate the modern world like legendary figures and artificial creatures in Greek mythology – with the main difference being that they are real in the narrow sense of the word. Some are only partially autonomous (acting under human command) while others are completely autonomous within their area of action. A genuinely autonomous machine should be able to act in a moral way, able to make decisions that are good for humans, animals and the environment. But what does it mean for machines to behave morally? Should they learn moral rules? Should they evaluate the consequences of their acts? Or should they become a virtuous character, following Aristotle? How is it possible to implement the classical normative models of ethics and is there a need for new ones?

In this paper the young field of machine ethics is explored. The main question is if it is possible to implement morality into autonomous machines. The answer is based on literature analysis and personal considerations and derivatives. Firstly, the concept and the classification of machine ethics are clarified with respect to the circumstance that it is not an established discipline. On the one hand, machine ethics can be subsumed under information ethics (including computer and net ethics) and technical ethics. On the other hand, it may be seen as a counterpart of human ethics, in that the autonomous system is a subject of morality. Secondly, new literature on the field is reviewed, focusing on a book about machine ethics, which was edited by Michael and Susan Leigh Anderson, two leading experts from the United States. Thirdly, the main topics of machine ethics are described; it is distinguished between different kinds of systems and situations in which they act, and present strategies of the industry are outlined. Fourthly, the paper tries to answer the question if and how it is possible to implement the classical normative models of ethics and which models should be preferred. Seven important normative approaches are described and estimated relating to their suitability for machine processing. Then the focus shifts to duty-based ethics, ethics of responsibility and virtue ethics that seem to be serious candidates. With a short technical analysis it can be shown that they fit to machine processing, apart from some limitations. The most promising approach may be the combination of the selected normative models. It is not only similar in the “normal” human ethics, but also an opportunity to balance out weaknesses of the autonomous machines and to allow them alternatives. In addition, other methods like orientation on reference persons and social media evaluation could be used.

The research field is, despite contributions of robot ethics since the 90s, full of challenges and difficulties. In this respect, the paper is work in progress and merely a small piece in the big puzzle of machine ethics. It is aimed at ethicists and experts of technology assessment as well as at KI experts and computer scientists. The author is sceptical about the possibility of implementing a moral code in a machine in a satisfactory manner. Moreover, the requirements of machine processing could be different from system to system (and even from situation to situation). But there will be a substantial interest from industry and military that would like to bring their solutions in the market respectively in the areas of conflict, and, in a different sense, of philosophy to solve some of the central questions. To say it from the philosophical point of view: Machine ethics will be the touchstone of ethics in general.