An altruistic approach concerning Machine Ethics

A new approach concerning ethical principles for robot-robot interaction based on the mechanism of natural selection

Madelon Hulsebos

S1600621

An essay in the context of the course Sex&Death: from Darwinian Revolution to Philosophical Crisis

Pre-master, academic year 2014-2015

Faculty of Philosophy, Leiden University

Theme: Machine Ethics

Number of words: 3.173

Technology is a gift of God. After the gift of life it is perhaps the greatest of God's gifts. It is the mother of civilizations, of arts and of sciences. – FREEMAN DYSON

Abstract Christopher Langton highlighted in his conclusion of Artificial Life that scientists need to focus on doing everything in their ability to improve artificial life technologies with regard to our beneficial needs.

Based on this, Isaac Asimov has posed the Three Laws of Robotics that prescribe how robots should be programmed. These laws are determined from a human-centered perspective. This has implications for Artificial Intelligence concerning how these robots should be programmed regarding their behavior towards human beings. Ashrafian has proposed the first law concerning robot-robot interactions. The law as proposed by Ashrafian as well as several other laws that prescribe the Right behavior for robots to interact with other (artificial) agents assume that robots act out of self-interest. This paper considers this meta-physical puts this assumption to the test. Before I will assert whether the analogy of acting out of self-interest will hold when it comes to robots, I will analyze the concurrent laws of robotics. After reconsidering the assumption that involves ethical egoism, I will argue for a new approach of programming ethical rules into the specification of robots. This new approach will be based on altruism. The last paragraph provides a conclusion of the paper and provides recommendations for further research.

Keywords: Robot-robot interaction / Machine Ethics / Natural Selection / Altruism

The Framework of Machine Ethics

Developments in the past ten years within the fields of Artificial Intelligence, Computer Science and Cognitive studies have complex implications for the interaction between biological and/or non-biological entities. This is reflected in the following relations between entities:

- Artificial Agent → Human Beings

  This aspect involves the actions of artificial agents that affect the (individual) environment of human beings.

- Artificial Agent ← Human Beings

  This field involves the actions by human beings that affect the (individual) environment of artificial agents.

- Artificial Agent ↔ Artificial Agent

  This field applies to the ethical aspect of the interaction between artificial agents.

The ethical aspect of how artificial agents should act and be programmed has multiple sides as posed above. The first and second path of interaction concerns the field of Machine Ethics. The main goal for Machine Ethics is to determine a set of ethical rules and learning procedures from which robots can abstract ideal ethical principles to guide its own actions. One of the main problems for Machine Ethics right now is that ethical theories lack agreement and we must know which is the correct ethical theory on before we can program robots in a consistent way. The third path of interaction actually concerns the robot-robot interaction that will be taken into consideration. Many people are worried about the future regarding to the dominant role that autonomous robots might get in our society. Therefore many scientists and philosophers have been focusing on the ethical aspect of the situations in which human beings and autonomous agents interact. However questions as: "Is it necessary to pursue creating robots that make moral decisions in the same way as we make moral decisions?" and "What will guide situations that involve robot-robot interaction?", are not frequently initiated.

Foundations for Ethical Robot-robot Interaction

Asimov provided the first set of rules to regulate the behavior of robots towards human beings:

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. (Asimov, 1950)

From the moment that artificial agents are in the ability to make impactful decisions themselves they have to be programmed with procedures that in any case limit their behavior with respect to human beings. But I will focus on whether the ethical approaches of human beings (must) hold regarding to robot-robot interaction. To address this issue, I will take the Three Laws of Robotics as assumptions to monitor their sustainability. As I have outlined before the first law as posed by Asimov is, and should, not be refuted.

Susan Leigh Anderson has argued that the Three Laws of Robotics are an unsatisfactory basis for Machine Ethics, whether artificial agents have moral status or not. Her objection relates to the second law of robotics, that broadly prescribes a human-slave relation when it comes to the interaction between human beings and robots. The argument for her objection is based on the Kantian argument which in essence is the following:

> "Any ethical law that humans create must advocate the respectful treatment of even those beings/entities that lack moral standing themselves if there is any chance that humans' behavior towards other humans might be adversely affected otherwise. If humans are required to treat other entities respectfully, then they are more likely to treat each other respectfully." (Anderson, 2008)

In my opinion the Kantian argument as adjusted by Anderson rules out the second law when it comes to human-robot interaction. The second law as posed by Asimov prescribes robots to obey the orders given by human beings. This constitutes a human-slave relationship which seems unethical towards human beings and robots but could it hold in robot-robot interaction? Thus the first question concerning robot-robot interaction could be formulated as: How should robots behave with respect to orders that are given by other robots?

The third law as prescribed by Asimov suggests that a robot must protect its own existence as long as such protection does not conflict with the other two laws. Enfin, I will take "A robot must protect its own existence (..)" into further consideration. The way as it is formulated it suggests that robots should behave according to the principles of ethical egoism when it comes to their own existence. The following example will illustrate how ethical egoism is involved: Imagine a situation where it comes to the existence of Robot 1 and Robot 2. According to the third law of Robotics, Robot 1 and Robot 2 should protect their own existence unless their actions to protect their own existence contrast with law one or two. If neither of their actions will contrast these laws, they will encounter a conflict where Robot 1 and Robot 2 both will pursue their self-preservation. Ethical egoism holds that individuals ought to do what is in their self-interest. This results in the following question: Does a robot necessarily have to protect its own existence if protecting it does not contrast the first two laws? Or rephrased: Can they possibly lead an altruistic existence?

Previous Law for Robot-robot Interaction

Hutan Ashrafian, a lecturer and surgeon at the Imperial College London, is the only person that has proposed an "AIonAI law". He has formulated the law as the following:

> "All robots endowed with comparable human reason and conscience should act towards one another in a spirit of brotherhood." (Ashrafian, 2014)

As he is suggesting himself, it seems possible that it will not hold since we might stand at the beginning of a new era when it comes to artificial intelligence. In my opinion the formulation of this law does not hold because it is ungrounded in the first place, and second because "brotherhood" could possibly mean something divergent from what it means to human beings. Ashrafian's ground for formulating the law in this way is:

> "As a global civilization, we have already considered the benefits of mutual respect and an adherence to a common principle of inherent rights formally listed in the Universal Declaration of Human Righs (UDHR). The adoption of such principles for AIonAI interactions would seem reasonable, rational, utilitarian and workable in cases where artificial

intelligences or robots do not contradict other fundamental robotic laws such as the prevention of harm to humanity or humans."(Ashrafian, 2014)

The predicate 'brotherhood' means that individuals must show mutual respect towards one another, to treat each other in such a way as proper to the relation of a brother. Though in the way as Ashrafian proposes, I suspect that he considers that the proposition "show mutual respect" involves the same actions that attest of mutual respect concerning robot-robot interaction as it does concerning human-human interaction. But the way as human beings treat each other as proper to the relation of a brother, does not have to be the same way as robots treat each other as proper to the relation of a brother. Ashrafian declares thereafter that Article 30 of UDHR, which states: "Non-permissible to perform any destruction of rights and freedoms", applies to robot-robot interaction and human-human interaction. Therefore Ashrafian is suggesting that a robot is not permitted to destruct itself in benefit of the fitness of the entire robotic population because this would involve robot-robot interaction as well. The analysis of the AIonAI law as proposed by Ashrafian has resulted in the question: Which ethical approach is suitable to guide robot-robot interaction relating to the guidelines to act towards one another in a spirit of brotherhood?

An Approach to Address the Three Questions

Three questions have emerged from the analysis of The Three Laws of Robotics by Asimov and the AIonAI law by Ashrafian within the framework that involves robot-robot interaction , which were:

- How should robots behave with regard to orders that are given by other robots?
- Can robots possibly lead an altruistic existence?
- Which ethical approach is suitable to guide robot-robot interaction relating to the guidelines to act towards one another in a spirit of brotherhood?

The last question gives a broader perspective on ethical rule-based robot behavior. To determine the ideal set with ethical principles that will guide moral decisions concerning robot-robot interaction, it is required to seriously consider ethical theories that originate from a human-centered perspective (Wallach & Allen, 2008). Several approaches to determine the fundamental ethical principles have leaded to contrasting views on what is the Right behavior. Currently this forms an issue within the

field of ethics. There is no agreement yet regarding which ethical theory comes to the Right behavior. Besides this issue many philosophers do not agree on the meta-ethical level as well concerning whether ethical theories even can be objective and universal or not. These conflicts have multiple complications for the implementation of rule-based ethical behavior relating to robot-human interaction, but are these issues also involved when it comes to robot-robot interaction? In the book Formal Ethics from 1996, Gensler proposed a set of formal ethical rules that seems to include neutral rules with respect to meta-ethical and normative issues. This set of formal ethical rules is based on a set of four axioms, which approximate the neutral fundamental principles for all ethical approaches, which include the following:

1. P (Prescriptivity) — "Practice what you preach"
2. U (Universalizability) — "Make similar evaluations about similar cases"
3. R (Rationality) — "Be consistent"
4. E (Ends-Means) — "To achieve an end, do the necessary means". (Gensler, 1996)

This set of axioms forms a reasonable starting point to determine universal principles for ethical rule-based robot-robot interaction. The first axiom “Practice what you preach” implies that for each individual applies that it logically entails to do A if the individual ought to do A. From this it is derivable that it completely depends on what a robot is ought to do. Until now two different approaches have been taken into consideration concerning the ultimate goal for robots. The first side of the coin seen from a human-centered perspective considers the ultimate goal for robots to obey the orders given by human beings and protect their existence if this does not conflict obeying the orders from human beings (Asimov, 1950). The other side of the coin seen from a human-centered perspective considers the ultimate goal for robots maximize its own performance measure (Russell & Norvig, 2003).

The Two Approaches of Specifying the Ultimate Goal

Asimov has posed a law that suggests that robots in the first place should obey the orders given by humans. If they do not have orders to obey, they are ought to remain acting in purpose of their self-preservation. In a situation where two robots interact this approach indicates that each robot is ought to

protect it-self. The second approach as indicated by Russell and Norvig concerns the principle of Maximum Expected Utility (MEU). MEU states that a rational agent should choose an action that maximizes the agent's expected utility. In chapter 2 they define the term rational agent as:

> "A rational agent is one that does the right thing – conceptually speaking, every entry in the table for the agent function is filled out correctly. Obviously, doing the right thing is better than doing the wrong thing, but what does it mean to do the right thing? As a fist approximation, we will say that the right action is the one that will cause the agent to be most successful."

It can be deduced from this, that a robot should maximize its own success. To merge these two conclusions, where the first part of the first approach is outdated regarding the priority to obey the orders given by humans, the general approach to program the behavior of the robots concerns a set of ethical rules in the scope of ethical egoism. But I would like to introduce the third side of the coin: Do not exist for yourself and do not obey the orders from another entity. What if we would apply the mechanism of natural selection to robot-robot interactions, is altruism a possible approach to optimize the expected utility of the robotic society?

The Foundations for an Altruistic Approach for Robot-robot Interaction

It still remains dubious whether biological moral agents, human beings for instance, can be altruistic or not, though an environment in which robots interact could be an altruistic environment whether robots are sentient and conscious or not. As long as robots are not biological moral agents, natural selection on a large scope could be considered as preferable. This means that a robot will be ought to do whatever comes with the Maximum Expected Utility regarding to what is best for the fitness of the robotic society. To frame the Maximum Expected Utility in terms of the fitness of the robotic society I would like to make an analogy that is based on the mechanism of natural selection. For each organism acting in a particular environment there are traits that are favored considering their environment because these traits increase the fitness of the organism. This mechanism results in the increasing chance for the organism to survive in its environment (Sterelny & Griffiths, 1999). This phenomenon is conceptualized as adaptation which occurs at the genetic level. Natural selection on the genetic level

in this analogy is reflected in the properties of each robot that will get selected based on its fitness so that the fittest robot will survive in its environment. When it comes to robot-robot interaction altruism and sacrifice is preferred in order to increase the fitness of the robotic society even when they become moral agents. The goal of selection, whether it concerns humans or robots, regarding to nature is to improve the fitness of the agent and thereby to improve the fitness of the species to which that agent belongs. By means of this selection mechanism this will lead to a fitness society, nature or environment. In my opinion this is not an impossible goal for robots therefore altruism and self-sacrifice should be the behavior that each robot prescribes what it is ought to do concerning robot-robot interaction in order to reach the Maximum Expected Utility regarding to what is best for the fitness of the robotic society.

The Implementation of the Altruistic Specification

Let's go back to the issue as I have posed whether robots should always protect their own existence.

To implement altruistic behavior into the specification of a robot, I would like to make a simple suggestion. The rule that should be programmed should determine the following factors in the environment in which the robot interacts:

- What are my intrinsic specification properties that improve the fitness of the overall robotic society
- What are the intrinsic specification properties that improve the fitness of the overall robotic society of the robot that I am interacting with
- What are the intrinsic specification properties that increases the fitness of the overall robotic society

When it has monitored the factors as posed above, the robot should do the following based on a simple 'if-then' rule:

1. *If* (my intrinsic specification properties that improve the fitness of the overall robotic society > the intrinsic specification properties that improve the fitness of the overall robotic society) *Then* (*set value* Protect-My-Own-Existence=1)
2. *If* (my intrinsic specification properties that improve the fitness of the overall robotic society

< the intrinsic specification properties that improve the fitness of the overall robotic society)

Then (*set value* Protect-My-Own-Existence=0)

As suggested above, the variable Protect-My-Own-Existence should not be considered as to fight for its own existence. It suggests a variable on which the two robots can calculate which specifications of the two robots has the highest value relating to the fitness of the robot.

A Kantian Review of the Altruistic Approach Considering

It is not to doubt on whether the Kantian approach of evaluating normative ethical theories is universal or not, because not one approach is universal. Though in my opinion the Kantian approach enables me to evaluate the ethical rule-based behavior, as I have proposed, from a human-centered perspective. It would also be consistent to evaluate the approach by Kantian reasoning because I have used this perspective to review the Three Laws of Robotics by Asimov as well. Kant's categorical imperative relating to the principle of universalizability states the following:

> "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law." (Kant & Patton, 2005)

The categorical imperative is an approach from the work "Grounding for the Metaphysics for *Morals"* by Kant, this implies that the categorical imperative is only applicable from a human centered view on this moment since robots do not have moral status (yet). The evaluation will be as following: If it comes to robot-robot interaction, can it become a universal law to let robots destruct or themselves to improve the fitness of their society? The answer to this question is 'yes'. Because we are

If robots actually will be accepted as moral agents in the future, the evaluation by means of the Kantian categorical imperative should be from a robot centered view. The evaluation will then be as following: If it comes to robot-robot interaction, can a robot want that it will become a universal law to destruct itself to improve the fitness of the society? The answer to this question is 'yes' as well, from a robot-centered perspective it is acceptable and preferable to enforce altruistic behavior regarding robot-robot interaction because as I have argued before, the main aim for the robot is to

improve the fitness of its society and this only works if every robot would act according the same ethical principles.

Conclusions and Recommendations for Further Research

Two out of the three laws of robotics as posed by Asimov do not hold regarding human-robot interaction. Furthermore I have argued that scientists and philosophers should focus more on the ethical aspect and possibilities within an environment where robots interact with each other. It has appeared that all ethical approaches are determined from a human-centered view. As appears, all recent philosophical and scientific approaches to determine ethical rule-based robotic behavior are based on the assumption that every robot ought to protect itself from harm apart from the situation as it occurs. This has resulted into reconsideration of these assumptions which lead to the conclusion that it is a possibility to specify altruistic robotic behavior regarding robot-robot interaction. I strongly recommend to do more research on the assumptions that are made in the past for ethical programming concerning robot-robot interaction. In my opinion, this recommendation involves the field of Machine Ethics.

Acknowledgements

Artificial Intelligence and Artificial Life have become the fields of my interest as a result of this particular course given by August Martin. Therefore, I would like to offer my sincerest gratitude to August Martin for providing such a fruitful source of inspiration. I ascribe my enthusiasm while writing this paper to his encouragement, effort and support.

References

Anderson, S. L. (2008). Asimov's "Three Laws of Robotics" and machine metaethics. *AI and Society, 22*, 477–493. doi:10.1007/s00146-007-0094-5

Ashrafian, H. (2014). AIonAI: A Humanitarian Law of Artificial Intelligence and Robotics. *Science and Engineering Ethics,* 21 (1), 29-40. doi: 10.1007/s11948-013-9513-9

Asimov, I. (1950). The evitable conflict. *Asounding Science Fiction,* 29(1), 48-68.

Kant, I., & Patton, H. J. (2005). Groundwork for the Metaphysic of Morals (1785). *Practical Philosophy*, 37–108. Retrieved from http://books.google.com/books?id=0hCsbUjFiBwC&pgis=1

Russell, S. J., & Norvig, P. (2003). Artificial Intelligence: A Modern Approach. *Second Edition, Prentice Hall,Upper Saddle River, New Jersey*.

Sterelny, K., Griffiths, P. E. (1999). Sex and Death: An Introduction to Philosophy of Biology. *The University of Chicago Press, Chicago.*

Wallach, W., Allen, C. (2008). Moral Machines: Teaching Robots Right from Wrong. *USA: Oxford University Press, New York.*