# Approaching Machine Ethics: Topics in Technical Education

Christopher B. Davison

Ball State University

cbdavison@bsu.edu

Colin Allen

Indiana University

colallen@indiana.edu

## Abstract

Machine ethics is a relatively new concept and is a growing topic of scientific exploration. As researchers seemingly move inexorably closer to human level performance, the ethical reasoning ability of machines has many concerned. Students in the technology domain, especially software engineers, are faced with increasingly complex situations in which their machine and software creations will be in a decision-making role. Often, those decisions will have ethical implications. The purpose of this article is to present strategies and classroom activities that will assist educators in teaching the concepts of machine ethics.

## Introduction

Ethical considerations and applied ethics are commonplace in situations encountered by humans. However, machines too are increasingly placed in situations where ethical considerations are required. Even though current lethal weapon systems are not fully autonomous, they may be in the near future, and current systems possess degrees of autonomy which nevertheless entail the application of ethical principles. As machines become increasingly designed to operate for long periods without human intervention or interaction, their capacity to make ethical choices requires investigation. From ATMs to airline auto-pilots (Moor, 2006) machines are increasingly tasked in ways requiring ethical responsivities. Bill Gates, Elon Musk and many leading figures from the technology industry are alarmed at the thought of artificially intelligent systems without ethical processes (e.g. https://futureoflife.org/ai-open-letter). The current mixture of social, commercial, political, and philosophical issues creates a complex scenario for teaching machine ethics in a classroom.

The article begins by providing a literature review and definitions specific to ethics and machine ethics. This will set the context for the subsequent sections and the discussions therein. Following that, the importance of ethics and machine ethics is discussed. Finally, three classroom exercises are presented that will engage students in learning about machine ethics. Although we focus on the ethical capacities of machines themselves, it is important to note that the machines are embedded in social-technical systems which themselves deserve ethical scrutiny, and the rapid growth of Artificial Intelligences (AI) and robotics has broad social and economic consequences, such as changes in the labor market, that we will not discuss here.

## Definitions

Morality can be characterized as the aspect of human decision making and behavior that concerns the effects of agents' actions upon other sentient beings. In its most general sense, ethics is the branch of knowledge dealing with moral principles or behavior. Ethics can be divided into three branches of study: metaethics, applied ethics, and normative ethics. (See Lin, Abney, and Bekey 2011 for discussion of these three branches in the context of machine ethics.)

In metaethics, the basic concepts of ethics are explored. Metaethicists focus on the foundational definitions and root structure of ethical theory. Concepts such as "What is right and wrong?" are explored.

The field of applied ethics is concerned with the application of ethical constructs to real-world and near-future scenarios. An example of this is the ongoing discussion about robot warriors. Some researchers argue that this development should be encouraged because robotic warriors have the potential to perform more ethically on the battle field (Arkin, 2010) while others argue that it is wrong to even try to imbue some forms of ethical reasoning into machines (Tonkens, 2009). The privacy implications of technology provide another prolific point of discussion for applied ethics and within the Career and Technical Education domain (Davison, 2007).

Normative ethics is a branch of thought that discusses the source and standards for judgments about the rightness or wrongness of individual actions; it is the study of actions from an ethical perspective. This branch of ethics deals with the rightness or wrongness of actions according to various ethical theories. In the field of machine ethics, an example of a question in normative ethics is whether machines should be held to the same standards as humans, or perhaps to different, even higher standards.

One important approach to normative ethics is deontological ethics, or deontology. In deontology, the morality of an action is based upon rules. It is sometimes referred to as an ethical code. Because parallels between deontological rules and rule-based computational algorithms can be drawn, deontological ethics is a tempting approach for programmers attempting to infuse ethics into their machines.

Kantian ethics (Kantianism), is named after German philosopher Immanuel Kant. Kantianism is a form of deontological ethical theory. It is concerned with autonomy of decision making and adherence to moral law. Moral law is specifically formulated as the categorical imperative which states (in various formulations) that people should act only in ways that their actions should become universal law—i.e., principles that all rational agents could follow without undermining the system of rational action.

Deontological approaches to normative ethics stand in contrast to consequentialist ethics. As the name “consequentialism” suggests, these ethical systems hold that actions should be evaluated by their outcomes. The particular motives or rationales for action are considered by consequentialists to have only derivative importance insofar as some rationales lead to better outcomes. Utilitarianism is the best-known form of consequentialist theory. Its major proponents, the philosophers Jeremy Bentham and John Stuart Mill argued that the maximization of overall pleasure or ‘utility’ was the sole criterion for moral evaluation. The idea of calculating utilities can also seem like a tempting approach for programmers, but there is the problem of exactly how to measure pleasure and suffering.

Machine ethics is defined as the implementation of moral decision making into computers, robots, and other autonomous devices (Allen, Wallach & Smit, 2006). Machine ethicists are concerned with the ethical reasoning of these machines and how, if possible, to imbue these machines with this reasoning ability.

A full moral agent is a being that is capable of knowing right and wrong and acting according to this capability. This type of agent is capable of making morally-based judgments.

An artificial moral agent (AMA) is an artificial agent that is computationally based, guided by norms, and implemented in software (Nagenborg, 2007). There exists a great deal of research within this domain. However, much like beating the Turing Test, it is not clear if creating an AMA that is a full moral agent is an achievable goal.

## Literature Review

Ethics deals with right and wrong from a human perspective and an implied sense of morality. Until very recently only humans were concerned with ethical norms and standards of conduct. However, with the increasing complexity of technological systems, and increasing autonomy of the software controlling these systems, has come a growing realization that autonomous systems will need some kinds of ethical capacities (Allen, Varner & Zinser 2000; Georges 2003; Arkin 2010; Wallach & Allen 2009; Anderson & Anderson 2011; Lin, Abney, & Bekey 2011). Thus, the field of Machine Ethics has slowly emerged over the past two decades.

Georges (2003) coined the phrase “Digital Soul” to describe, among other aspects of AI, the ethical decision making capabilities, programmed or otherwise, of artificial moral agents. This phrase has evokes the idea of a Divine Command theory of morality: the belief that is common to many religious traditions that the source for moral standards for human behavior is to be found in the wishes or commands of one or more deities. By analogy, the relationship of machines to humans may be one in which the source of machine morality should be human wishes and commands.

Only a small subset of the literature on machine ethics proposes a general architecture for artificial moral agents (e.g., Wallach, Franklin & Allen 2010) although efforts to build working prototypes are increasing (Anderson, Anderson & Armen 2006). As computational systems continue to increase in power and capability, the reciprocal need for AMAs will continue to increase.

There is no agreement among machine ethicists on the ethical framework that can and should be implemented into AMAs. Tonken (2009) argues that implementing a Kantian ethical framework would be *prima facie* anti-Kantian as it cannot support Kant's view on autonomy—the absolute freedom of rational agents to choose how to act. Challenges to this argument have been presented by other machine ethicists (White, 2015; Arkin, 2010) including the rationale for Lethal Autonomous Weapon Systems (LAWS). It is important to recognize that the philosophical meaning of "autonomy" is more tied up with issues of free will and consciousness than the engineering sense which refers to machines operating without direct human oversight (Wallach & Allen 2009).

Some researchers argue against the entire premise of machine ethics as "misguided" because of fundamental differences between the capacities of humans and AI (Yampolskiy, 2013, p. 389). Humans possess emotions, pain receptors, and feelings and it is not clear if those can or should be transferred to algorithms and computational machinery. There are many ethical frameworks, and considerable disagreement exists over ethical norms. For humans to agree upon *which* ethics to apply, let alone achieve agreement upon *how* to apply them, presents a difficult problem of building an ethical consensus.

There are a number of significant challenges in creating a fully capable ethical system including emotions (e.g., empathy and compassion) and implementation of a broader range of mental states (Sparrow, 2009). There is no clear correct way to build an AMA. It would appear that an autonomous system, in the engineering sense, imbued with ethical governors – i.e., dedicated systems geared towards a particular context of use (Arkin 2010; Kinne & Stojanov, 2016) is the approach that is feasible with current technology.

### Importance of Learning Ethics and Machine Ethics

Some scholars argue that machine ethics do not or cannot exist. Some of these arguments are based on skepticism about the idea of "Strong AI", i.e. fully conscious, human-equivalent AI. Searle (1980) argues that computer programs could never possess "intentionally" (p. 417) and thus Strong AI could not exist without duplicating human brains. Some argue that machines do not have free will and a sense of self, and therefore, could not become an AMA. However, due to the increasing capability and sophistication of machines, coupled with their increased role in ethically challenging situations, to ignore machine ethics would be short-sighted.

Elon Musk (Telsa Motors) has characterized AI as one of the biggest existential risks facing humanity. He and Sam Altman (startup incubator entrepreneur) have collectively formed OpenAI in 2015. This is a research startup company with the goal of promoting non-harmful AI. The company was pledged $1B (USD) in funding to open source AI development. The human survival strategy the company is pursuing is one of more AI will equate to more good. Musk believes that if AI is everywhere with everyone, then the odds of a malevolent AI in the hands of a few will be diminished. He also acknowledges the possibility of creating the very thing he intends to preclude (Markoff, 2015).

Other noteworthy people are concerned about the threat humankind faces with Strong AI. Bill Gates is concerned and has stated he agrees with Elon Musk regarding the threats posed by super intelligent machines. Stephen Hawking is concerned that Strong AI could spell the end of the human race. Hawking and Musk are both signatories on the "Future of Life" letter written and sponsored by the Future of Life Institute.

Another aspect to the importance of machines ethics is known as the Technological Singularity, or more simply, the Singularity. The Singularity, the term attributed to John von Neumann, is the runaway self-improvement cycles of machine learning. As machines learn and improve, it is hypothesized that their self-improvement cycles will escalate and trigger a self-improvement explosion resulting in a super intelligence of which humans cannot compete. While it is impossible to predict what human life will entail in the post-Singularity world, the thought of an amoral super intelligent system makes a strong case for teaching Machine Ethics.

## Classroom Activities

The following three classroom activities are designed to engage students in the exploration of ethics and machine ethics. While there is no *a priori* knowledge of ethics of machine ethics is assumed; basic computer, mathematics, and proficiency with personal computers is required. These activities are meant to build upon each other. The first is an exercise designed to assist students in understanding ethics from a human behavioral perspective. The second is an activity that will assist students in learning about machine ethics.

### *Learning Objectives*

1. Students will demonstrate knowledge of ethics and machine ethics.
2. Students will compare and contrast ethics and machine ethics.
3. Students will synthesize and evaluate theoretical approaches to machine ethic implementations.

### *Required Materials*

1. Computer with Web Browser

2. Internet Connectivity

*Classroom Exercises*

1. Moral Philosophy and Ethics.
    a. The instructor begins the discussion on "What is ethics?" and shares information regarding the theory of ethics and define/discuss the following (see above definitions):
        a. Meta-Ethics
        b. Normative Ethics
        c. Applied Ethics
    b. Students may review the literature on the Internet for further information and discussion on ethical theory.
    c. Students are organized into two teams (may be subdivided for large classes). Team 1 will defend the view that ethics can be engineered into machines. Team 2 will defend the view that ethics is not something that can be engineered. Team members discuss their best argumentative strategy.
    d. A classroom debate is held between the two teams.

2. Machine Ethics.
    a. The instructor explains basic machine ethics concepts including moral agency, and AMA.
    b. Students compare and contrast machine ethics versus human ethics.
    c. Groups of students are first tasked with brainstorming about *past unexpected/unintended* consequences of AI innovations – for example, ways in which AI has changed social media, or finance markets.
    d. Groups are then tasked with brainstorming about *future unintended* consequences of AMAs. Examples: If autonomous military robots are described as having ethical governors, will this make war more likely, or will it change the nature of warfare in ways that may be difficult to anticipate? Will self-driving cars that stop more readily for pedestrians and are less likely to run them over than human drivers, make pedestrians more likely to disrupt vehicular traffic? Will AMAs make reduce or otherwise affect personal privacy in unintended ways?

3. Implementation of Ethics in Technology
    a. The instructor tasks the class to provide ethical algorithms for the following scenarios:
        i. Human Privacy Preservation. An autonomous drone searching for terror suspects. The drone has an array of sensors including optical, acoustic, and thermal imaging. How can the drone carry out its mission and yet still preserve the law-abiding citizens' right to privacy?

ii. War fighting. Robotic war fighters, Arkin's (2010) lethal autonomous unmanned systems, are a large topic of discussion. Is it possible to engineer better-than-human ethical decision making? What would those algorithms look like? Should there be "divine commands" such as those postulated by Bringsjord,and Taylor (2015)?

iii. Elder care robots: How should a robot be designed to balance a duty of care against the individual's right to autonomously refuse medications (cf. Anderson, Anderson & Armen 2006).

iv. Autonomous vehicle: A passenger orders a self-driving car to drive 30% over the speed limit because of a medical emergency. Design an algorithm to determine whether the car should obey.

## Conclusion

In this article, the concepts of ethics, machine ethics, and AMAs were discussed. A review of the literature was presented that outlined human ethics, definitions of ethics and definitions of machine ethics. Following the literature review, the importance of teaching and learning machine ethics was discussed. Finally, three classroom exercises were presented to assist educators in teaching machine ethics in their classrooms.

## References

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems*, *21*(4), 12-17.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12, 251 – 261.

Anderson, A. & Anderson, S. (2001). *Machine Ethics*. Oxford: Oxford University Press.

Anderson, M., Anderson, S. L., & Armen, C. (2006). An approach to computing ethics. *IEEE Intelligent Systems*, *21*(4), 56-63.

Arkin, R. C. (2010). The case for ethical autonomy in unmanned systems. *Journal of Military Ethics*, *9*(4), 332-341.

Bringsjord, S. & Taylor, J. (2011). *Introducing divine command roboethics*. In Lin et al. (Eds.), 85-108.

Davison, C. B. (2007). Ethics of business continuity and disaster recovery technologies: A conceptual Orientation. *International Journal of Computers, Systems and Signals*, *8*(1), 54-63.

Georges, T.M. (2003). *Digital soul: Intelligent machines and human values*. Cambridge: Westview Press.

Kinne, E., & Stojanov, G. (2016, March). Grounding Drones' Ethical Use Reasoning. In *2016 AAAI Spring Symposium Series*.

Lin, P., Abney, K., & Bekey, G. A. (2011). *Robot ethics: the ethical and social implications of robotics*. MIT press.

Markoff, J. (2015, December). Silicon Valley investors to bankroll artificial-intelligence center. *The Seattle Times.* Retrieved from:

http://www.seattletimes.com/business/technology/silicon-valley-investors-to-bankroll-artificial-intelligence-center/.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *Intelligent Systems, IEEE 21*(4):18-21.

Nagenborg, M. (2007) "Artificial moral agents: an intercultural perspective." *International Review of Information Ethics* 7(9), 129-133.

Searle, John. R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences 3*(3), 417-457.

Sparrow, R. (2009). Building a better WarBot: Ethical issues in the design of unmanned systems for military applications. *Science and Engineering Ethics*, *15*(2), 169-187.

Tonkens, R. (2009). A challenge for machine ethics. *Minds and Machines, 19*(3), 421-438.

Wallach, W., & Allen, C. (2010). *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Wallach, W., Franklin, S. & Allen, C.  (2010). A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents. *Topics in Cognitive Science* 2(3), 454-485.

White, J. (Ed.). (2015). *Rethinking Machine Ethics in the Age of Ubiquitous Technology*. IGI Global.

Yampolskiy, R. V. (2013). Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In *Philosophy and Theory of Artificial Intelligence* (pp. 389-396). Springer Berlin Heidelberg.