Kantian one day, Consequentialist the next: Moral emotions as mediators between ethical frameworks for robots

EXTENDED ABSTRACT

The field of machine ethics attempts to design and develop the computational underpinnings necessary for a robot to make ethical decisions in real-world environments. One of the main issues faced by machine ethics researchers, and scholars in many other realms, is the apparent lack of agreement as to the existence and nature of a correct moral theory. This problem is central to machine ethics research because it appears difficult or impossible to create a morally competent robot if there is no agreed upon moral theory to inform one's design. Philosophers have, over the centuries, developed numerous ethical frameworks which could be used to guide machine ethics research. Some philosophical theories of ethics (e.g., Utilitarianism) more easily lend themselves to software encoding and robot action selection than others. The mere ease with which a particular ethical theory can be programmed into a robot should not necessarily be the decisive factor in theory selection. Moreover, ethical frameworks can lead to conflicting recommendations, raising the issue of how to adjudicate between them. Obviously, this is a complex and nuanced subject matter but for simplicity sake, we will use an example to highlight this point. Kantians and Utilitarians continue to debate many aspects of ethical decision-making, including what the fundamental goal of ethics is. While Kantians focus on one's ethical duties, including the obligation to respect rational agents, Utilitarians seek to generate good consequences for society.

Since a lack of consensus persists regarding which particular ethical framework should be embraced, our research seeks to generate action recommendations for a robot grounded on the insights from several ethical frameworks. This added flexibility may allow the system to be more adaptive when confronting a situation that it has not faced in the past (and arguably more in line with how humans actually reason through an ethical decision). Moreover, the system may be able to use its experience including the ethical dilemmas it has faced, and the solutions it has applied, to individualize its ethical reasoning in a nuanced and unique manner. This paper will focus on the philosophical justification for the selection of which ethical frameworks are incorporated into this architecture. We intend to explore how and why different ethical theories might be integrated and arbitrated by moral emotions as well as what the expected robot behavior might be. We will also look at situations and robot roles that could demand a different mix of theoretical underpinnings. For example, postulating how Humanism, Collectivism, and Moral Relativism might be used to create an ethical healthcare robot.

Many ethical frameworks could, in principle, be used as a candidate option for guiding a robot's decision-making. Our initial work will seek to encode at least three ethical frameworks into a robot's design matrix: (1) Utilitarianism, (2) Kant's ethical theory, and (3) W.D. Ross's duty-based view. Utilitarianism, probably the most well-known form of consequentialism, frames moral goodness in terms of what will promote "the greatest good" for society. While many different types of Utilitarianism have been formulated over the years (Driver 2014), the general overarching notion that it advocates is to assess the potential benefits and harms that different entities may experience with regard to a proposed course of action.

Deontology is a collection of views that define ethical rightness and wrongness independently of an act's consequences; rather, it focuses on whether ethical obligations have been upheld (Alexander and Moore 2016). For example, Immanuel Kant, through his various formulations of the Categorical Imperative, sought to provide a process that guides decision-making towards absolute and universal ethical maxims.

Our third framework is derived from the philosopher W.D. Ross who sought to combine insights from Utilitarianism and Kantianism and formulate a view that captures the ethical duties that humans have to one another (Ross 1930). According to Ross, humans have to adhere to a collection of *prima facie* duties. Roughly stated, a *prima facie* duty is something we are obligated to uphold (e.g., promoting good) unless superseded by a more important duty (e.g., harm avoidance) in a particular case (Skelton 2012).

The relatively young machine ethics community has focused largely to date on developmental ethics, i.e., how an agent would develop its own sense of right and wrong in situ. In general, these efforts largely ignore moral emotions as a scientific matter worthy of consideration. There is, however, strong evidence that moral emotions guide human beings in making ethical judgments. For example, according to Gazzaniga (2005), there are three neuroscientific aspects of moral cognition: (1) moral emotions; (2) theory of mind; and (3) abstract moral reasoning. Regarding moral emotions, they arguably can bias decision-making in a way that supports positive moral behavior. Thus, we will seek to incorporate moral emotions in a robot's architecture, perhaps informed by Haidt's taxonomy (2003): Other-Condemning (e.g., Anger); Self-conscious (e.g., Guilt); Other-Suffering (e.g., Compassion); and Other-Praising (e.g., Gratitude).

For our research, moral emotions will be used as a basis for guiding a robot's decision-making when there is divergence regarding what different ethical frameworks would recommend. In other words, moral emotions will help a robot to navigate situations when two or more ethical frameworks would disagree about what the appropriate course of action is. At least initially, we will leave aside circumstances where the disagreement resides within one framework (e.g., between Utilitarians). Briefly stated, the ethical decisions that our robot architecture make will be mediated by the emotional, or simulated emotional, state of the agent (human or robot). These decisions will be compared to the selections of humans in similar situations.

References

Alexander, Larry and Moore, Michael, "Deontological Ethics", The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/>.

Driver, Julia, "The History of Utilitarianism", *The Stanford Encyclopedia of Philosophy* (Winter 2014 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2014/entries/utilitarianism-history/>.

Gazzaniga, M. *The Ethical Brain*, Dana Press, 2005.

Haidt, J. "The Moral Emotions", in *Handbook of Affective Sciences*, Oxford Press, 2003.

Ross, David. *The Right and the Good.* Oxford: Oxford University Press, 1930.

Skelton, Anthony, "William David Ross", *The Stanford Encyclopedia of Philosophy* (Summer 2012 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2012/entries/william-david-ross/>.

- Provide a short abstract of 150-250 words suitable for inclusion in a program.

SHORT ABSTRACT

The field of machine ethics in the process of designing and developing the computational underpinnings necessary for a robot to make ethical decisions in real-world environments. Yet a key issue faced by machine ethics researchers is the apparent lack of consensus as to the existence and nature of a correct moral theory. Our research seeks to grapple with, and perhaps sidestep, this age-old and ongoing philosophical problem by creating a robot architecture that does not strictly rely on one particular ethical theory. Rather, it would be informed by the insights gleaned from multiple ethical frameworks, perhaps including Kantianism, Utilitarianism, and Ross's duty-based ethical theory, and by moral emotions. Arguably, moral emotions are an integral part of a human's ethical decision-making process and thus need to be accounted for if robots are to make decisions that roughly approximate how humans navigate through ethically complex circumstances. The aim of this presentation is to discuss the philosophical aspects of our approach.