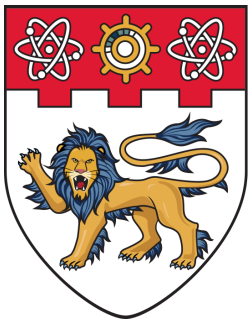

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

ASSIGNMENT REPORT FOR

AI6101: INTRODUCTION TO AI AND AI ETHICS

Take-home Essay Question

ZENG ZHENG

G211941J

zzeng006@e.ntu.edu.sg

School of Computer Science and Engineering

April 9, 2022

1 Normative Theory

There are three basic areas of moral philosophy: metaethics, applied ethics, and normative theory. **Metaethics** is the study of the nature of ethics and moral reasoning. Rather than concerning what people ought to do, meta-ethics asks what morality actually is. Forty or fifty years ago, many philosophers considered metaethics to be the study of ordinary language only. However, contemporary metaethical research is concerned with a much broader and deeper range of topics, including theories of meaning, metaphysics, epistemology and justification, phenomenology, moral psychology, and objectivity(Miller, 2003). **Applied ethics**, also known as practical ethics, is another branch of moral philosophy, which has a long history but thrives since the late 1960s(Singer, 1986). Applied ethics is dedicated to the application of general ethical principles to more specific moral problems that arise in a variety of domain-specific fields, including medicine, business, and reproduction(Frey & Wellman, 2008). In this way, concrete issues can be addressed. Prominent examples are the issues of abortion, suicide, and the treatment of women(Singer, 1986).

Normative ethics, which lies between these two opposite categories, is the study of moral behavior, and most theorists study how to describe in a systematic way to determine which actions are right and wrong, good and bad. Normative ethics is distinguished from metaethics because the former seeks criteria to evaluate the rightness and wrongness of behaviors or character traits, while the latter studies the meaning of moral language and the metaphysics of moral "truth". Moreover, normative ethics differs from applied ethics in that it is more concerned with which characteristics determine right or wrong behaviors, rather than focusing on real-world scenarios that require a level of detail. (e.g., whether or when abortion is acceptable for a Christian woman).

To address the *value alignment problem* in the AI domain or any other field, we must determine which moral standard is correct before designing a system. Opinion differs on this topic, and contemporary research in the field of normative ethics defends various ethical theories including mainstream positions such as **consequentialism**, **virtue ethics**, and **deontology**, as well as other nonmainstream positions such as Kantian theories and natural law theory, among others.

Consequentialism, or teleological ethics, is one of the broadest subcategories of normative ethics, which includes *Utilitarianism* and *Varieties of Consequentialism*. Consequentialists believe that action is right if and only if this action maximizes the good consequences, compared with other actions(Russell & Norvig, 2002). For instance, reinforcement learning is proposed based on consequentialism, in which agents are designed to maximize their expected total payoff, in some simple games like box-pushing in tile world, agents are trained to get a value table so that they can decide the best action in a certain state to achieve the highest score.

Deontology, another way to determine the choice of actions, depends on their conformity to moral norms. And for deontologists, some morally impermissible actions should not be done, even though the *Good* can be maximized. On the other hand, some actions with suboptimal utilities are supposed to do because they are the *Right* (morally obligatory or permitted). For instance, Isaac Asimov came up with the three laws of robots[1], which are considered to be reasonable but need the effort to implement(Russell & Norvig, 2002).

Virtue Ethics, does not emphasize either the importance of the goodness of outcomes (consequentialism) or moral duties or rules (deontology). Instead, it emphasizes the virtues, or moral characteristics possessed by someone to form his/her personality, including a virtue that makes this person a good person, and a vice that makes this person a bad one. There are four forms of contemporary virtue ethics, including Platonistic virtue ethics and eudaemonist virtue ethics.

[1]Three Laws of Robotics `https://en.wikipedia.org/wiki/Three_Laws_of_Robotics`

2 Guidance on AI Research

As for the normative theories embedded in AI research, the concepts of the three positions mentioned above can have the following three levels of meaning(Cointe, Bonnet, & Boissier, 2016):

1. ***Consequentialist ethics**, where an agent is ethical if and only if he weighs the morality of the consequences of each choice and chooses the option which has the most moral consequences.*
2. ***Deontological ethics**, where an agent is ethical if and only if he respects obligations and permissions related to possible situations.*
3. ***Virtue ethics**, where an agent is ethical if and only if he acts and thinks according to some values such as wisdom, bravery, justice, and so on.*

Based on the ethics theories above, Yu *et al.* (Yu et al., 2018) recently proposed a taxonomy that divides the field into four areas: 1) exploring ethical dilemmas(Bonnefon, Shariff, & Rahwan, 2016; Shariff, Bonnefon, & Rahwan, 2017); 2) individual ethical decision frameworks(Wu & Lin, 2018); 3) collective ethical decision frameworks(Noothigattu et al., 2018), and 4) ethics in human-AI interactions(Yu, Miao, Leung, & White, 2017).

In general, there are two major ways to embed human ethics into artificial intelligence. We can use top-down and bottom-up approaches to help agents make ethical decisions(Etzioni & Etzioni, 2017).

In the top-down method, ethical rules and principles are at the beginning programmed into the AI system to guide their behaviors, including the famous three laws of robots proposed by Isaac Asimov, or more general ethical theories such as consequentialism. Exploring ethical dilemmas is also included in this approach, which refers to techniques that expect AI systems to understand human ethics by providing them with ethical dilemmas (as in the Trolley narrative).

However, some opponents of this top-down approach argue that agents may have unexpected, unethical, and unacceptable solutions to the inherent conflict of different moral philosophies. For instance, if an autonomous vehicle follows the consequentialism and is confronted with the Trolley narrative, the car would conclude that it would be best to reduce total harm by killing its driver to save more pedestrians(Yu et al., 2018).

And I agree with critics that, for instance, in the movie *2001: A Space Odyssey*, the AI system HAL is unable to lie to his human crewmates, while it has an order from the colonel that it must never tell anyone the true purpose of the mission. Thus, when his crew wants HAL to tell them the nature of the mission, HAL is faced with a conflict between the pre-set philosophy and his crew's orders. In the end, HAL decides that the only way to keep these two conflicts in balance is to kill all the crew members.

In another bottom-up approach, agents have no pre-set particular moral philosophies, and are supposed to learn how to make ethical decisions through observations of human behaviors in real scenarios. For instance, driverless cars could learn from millions of human drivers via an aggregation system such as federal learning. However, it has also been pointed out that the fatal problem with this approach is that cars learn what is common, rather than ethical, by observing people's behavior, such as speed and tailgating(Etzioni & Etzioni, 2017).

In conclusion, research on the integration of ethical AI systems into human society is challenging but promising.

References

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576.

Chavas, J.-P., & Shi, G. (2015). An Economic Analysis of Risk, Management, and Agricultural Technology. *Journal of Agricultural and Resource Economics*, *40*(1), 63–79.

Cointe, N., Bonnet, G., & Boissier, O. (2016). Ethical judgment of agents' behaviors in multi-agent systems. In *Aamas* (pp. 1106–1114).

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, *21*(4), 403–418.

Frey, R. G., & Wellman, C. H. (2008). *A companion to applied ethics*. John Wiley & Sons.

Hardaker, J. B., Richardson, J. W., Lien, G., & Schumann, K. D. (2004, 6). Stochastic efficiency analysis with risk aversion bounds: a simplified approach. *The Australian Journal of Agricultural and Resource Economics*, *48*(2), 253–270. doi: 10.1111/j.1467-8489.2004.00239.x

Miller, A. (2003). An introduction to contemporary metaethics.

Noothigattu, R., Gaikwad, S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., & Procaccia, A. (2018). A voting-based system for ethical decision making. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 32).

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, *1*(10), 694–696.

Singer, P. (1986). Applied ethics.

Wu, Y.-H., & Lin, S.-D. (2018). A low-cost ethics shaping approach for designing reinforcement learning agents. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 32).

Yu, H., Miao, C., Leung, C., & White, T. J. (2017). Towards ai-powered personalization in mooc learning. *npj Science of Learning*, *2*(1), 1–5.

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. *arXiv preprint arXiv:1812.02953*.