Take-home Essay

Li Hantao, G2101725H, MSAI, hli038@e.ntu.edu.sg

***Moral philosophy* is traditionally subdivided into *metaethics*, *normative theory* and *applied ethics*. What are the possible ethical standards that we find described in normative theory and how might they provide guidance on the AI research front?**

I. Moral Philosophy

Moral philosophy is the study of what morality is and what it requires of us. As *Socrates* said, it's about "*how we ought to live*"—and why. [1] *Ethical premises are typically treated in the manner of mathematical propositions: directives supposedly in dependent of human evolution, with a claim to ideal, eternal truth.* [2] Moral philosophy is limited by the gap between "is" and "ought" so fat; thus, the resulting life facts themselves cannot provide a moral blueprint for future action. For this reason, philosophers often believe that ethics has nothing to do with science to a great extent. Nevertheless, with technology development, especially AI technology, moral philosophy can provide knowledge or norms for AI researchers.

Moral philosophy is traditionally subdivided into metaethics, normative theory and applied ethics. In terms of philosophical definition or concept, the three subcategories are very different. **Metaethics** can study whether a moral principle should be adhered to and, more deeply, what morality means. *It is the attempt to understand the metaphysical, epistemological, semantic, and psychological, presuppositions and commitments of moral thought, talk, and practice.* [3] **Applied Ethics** is the study of ethics about a specific real-world behavior, which *refers to the practical application of moral considerations.* [4]

Ethicists have no standard view on what **Normative Theory** is and what it discusses. [5-6] Although there have been various normative theories for a long time in history, as a separate research field, it was formed after metaethics and applied ethics were independent of the moral philosophy. The definition put forward by *Lewis Vaughn* and *Theodore Schick* in *Doing Philosophy* is that normative ethics can tell us what moral principles should be used to make correct moral judgment and determination. [7] We can see that the definition here shows that normative ethics does not study what kind of person is moral, but only what kind of behavior principle is moral. We can explain the difference between the three classifications through the following football match example.

Let us assume that a football match is being held. The football rules are formulated by FIFA in advance; the referee on the court is responsible for determining whether the player breaks the rules; there are sports commentators off the court commenting on whether the rules and penalties are reasonable. In this scenario, FIFA is the normative theory, focusing on telling whether a principle is moral (whether a player's behavior is foul in rules). The referee is applied ethics, which focuses on whether a specific behavior is ethical in a specific scene (judging whether a specific action of a player is a foul at a particular circumstance). Football commentators are like metaethics, which can study whether a moral principle should be persevered (commentators can comment on whether a football rule is reasonable).

All in all, ***Normative Theory*** *is the part of philosophy devoted to elucidating and defending very general ethical principles, commonly contrasted with* ***Metaethics*** *(part of philosophy concerned with the ultimate status and grounding of ethics), and* ***Applied Ethics*** *(which is devoted to offering and defending solutions to practical moral problems).* [8]

II. Normative Theory

Contemporary, we generally believe that normative theories can be divided into three categories: **Consequentialism**, **Deontology** (obligation theory), and **Virtue Ethics**.

Consequentialism *is the view which claim that the moral value of any act consists in its tendency to produce things of intrinsic value.* [9] This theory holds that whether an action is moral depends on the consequences of it or the consequences of things (such as the motivation or the rule behind the action) related to it. **Deontology** *is the theory that the morality of an action should be based on whether that action itself is right or wrong under a series of rules, rather than based on the consequences of the action.* [10]

In the discussion on the difference between consequentialism and deontology, the subjective arguments of philosophers can be expounded in the following points:

1. Consequentialism looks only at the result, while deontology looks only at the motivation.
2. Consequentialism pays attention to the consequences of behavior, while deontology pays attention to the characteristics of behavior itself. [11]
3. The deontology holds that '*Right*' takes precedence over '*Good*', while the consequentialism is on the contrary.
4. Consequentialism demands too much on people who doing the behavior, but deontology does not.

Nevertheless, we have to make it clear that these arguments are widely controversial. For example, *Larmore Charles* believes that both Consequentialism and deontology take '*Right*' over '*Good*.' [12] We can expand the previous examples of football to deepen our understanding of the two theories further.

Let us consider the formulation of a 'foul tackle' in football rule. We have two approaches to set the penalty criteria. On the one hand, we take the consequences caused by the tackle as the standard of punishment: it will be a foul when the tackle causes the interruption of attacking or others injured, which is the way of **consequentialism** method, focusing on the consequences of an action to determine whether it is moral (whether it is a foul). On the other hand, we take the motivation of the tackle as the standard of punishment: it will be a foul when the player intends to stop the attacking (pull the attacking player when counterattacking) or to hurt others (tackle behind or lift feet off the ground), which is the way of **deontology** method, focusing on the motivations and characteristics of the behavior itself to justify whether it is moral (whether it is a foul).

In addition, contrasted with the above two major approaches in normative theory, consequentialism and deontology, which make the goodness of outcomes of action and the concept of moral duty central, **Virtue Ethics** differs from both of them as it focuses on being over doing. *Virtue ethics is a class of normative ethical theories which treat the concept of moral virtue as central to ethics.* [13] Possessing these virtues is what makes one moral, and one's actions are a mere reflection of one's inner morality. [14]

With the progress of modern moral philosophy, more and more scholars believe that the previous classification method, shown above, is wrong, at least inaccurate. For example, *Kagan Shelly* proposed that the mainstream trisection has significant defects: '*This seems to me exactly incorrect.*' [15-16] He divided the normative theory into two levels, the **Foundational level** and the **Factor level**. Factor level is used to explain which behavior or factors determine the mora; the foundational level is used to defend or explain why these factors can justify that. In Kagan's view, consequentialism and deontology are typical theories of the foundational level.

III. GUIDANCE ON AI RESEARCH

The continuous progress and broad application of AI bring notable benefits. However, to make AI profitable to human society, we cannot ignore the ethical problems behind them. Building Ethics is an essential topic of AI governance. Researchers propose a taxonomy that divides the field into four areas: 1) exploring ethical dilemmas; 2) individual ethical decision frameworks; 3) collective ethical decision frameworks; 4) ethics in human-AI interactions. [17] When we establish an AI system, we are the formulator of rules and principles. Therefore, relevant ethical issues can be studied and explained under the framework of normative theory.

Corresponding to the trisection method introduced earlier, scholars have explained AI ethics in this way: [18]

1. ***Consequentialist ethics****: an agent is ethical if and only if it weighs the consequences of each choice and chooses the option which has the most moral outcomes. It is also known as utilitarian ethics as the resulting decisions often aim to produce the best aggregate consequences.*
2. ***Deontological ethics****: an agent is ethical if and only if it respects obligations, duties and rights related to given situations. Agents with deontological ethics act in accordance to established social norms.*
3. ***Virtue ethics****: an agent is ethical if and only if it acts and thinks according to some moral values. Agents with virtue ethics should exhibit an inner drive to be perceived favourably by others.*

Our goal is to design an ethical AI, that is, to embed the legal, moral, norms, and values of human society into the AI system. Thus, we have two ideas for establishing AI ethics:

On the one hand, we can use a large number of pre-set moral model data to establish the relationship between behavior and morality through deep learning to make AI learn human moral preferences and rules. For example, MIT researchers launched the *Moral Machine Experiment* to collect data on various moral dilemmas. [19] On the other hand, the corresponding reward function can be established according to the moral models mentioned above to make the goal of the agent consistent with the goal of humans through reinforcement learning. Duke University used game theory to model moral dilemmas and make AI's moral values consistent with humans. [20] MIT designed a computational model to describe moral dilemma as a utility function, introducing a Bayesian model representing social structure and norms. [21]

However, although we have sorted out a reasonable way to formulate ethical standards for AI, it is still challenging in the practical application scenario: the real world is not a BoxPushing grid-world game like we implemented in the AI6101 project.

Firstly, the real world has an infinite number of possible circumstances instead of a grid-world game. We cannot guarantee that the feature in DNN or the state in reinforcement learning can be extended to all situations that the AI system will encounter, especially those affecting the safety of human life and property, rather than simply classifying cats and dogs or playing chess.

Secondly, problems like the tram dilemma have not been solved even in human society, where utilitarianism and absolutism give different moral choices. Such problems will also be encountered in the automation scene, such as auto-drive, where we cannot warrant that AI can provide satisfactory answers.

Thirdly, it is challenging for us to quantify AI rules accurately when using the criteria of normative theory; that is, to quantify Bentham's '*Pain*' and '*Pleasure*' or Ross's '*Right*' and '*Good*' with formulas. [22-23] Therefore, we cannot assert that our construction of the AI moral system is suitable. For instance, we cannot prevent the cheating reward mechanism in reinforcement learning (using the vulnerabilities in the reward functions to obtain more rewards than expected).

Last but not least, we cannot guarantee that the performance of ethical AI, established by guidelines like normative theory, is what we demand. Avoiding AI from becoming the Midas touch, in Greek mythology, or the robot repeating the circle, in Roundabout in Asimov's novel, is a task worth pondering. [24]

In addition, we should also pay attention to the necessary supervision: legal remedies need to be provided for the destruction caused by algorithmic decision-making and discrimination, which are moral issues hidden behind AI but need our attention urgently as well. The AI community is more involved by engineers, lacking the participation of philosophy, ethics, and other social disciplines. In the future, such interdisciplinary about AI ethics needs to be strengthened.

REFERENCES

[1] Rachels, J., & Rachels, S. (1986). The elements of moral philosophy (p. 9). Philadelphia: Temple University Press.

[2] Ruse, M., & Wilson, E. O. (2021). Moral philosophy as applied science (pp. 365-379). Princeton University Press.

[3] Sayre-McCord, Geoff, "Metaethics", The Stanford Encyclopedia of Philosophy (Summer 2014 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2014/entries/metaethics/>.

[4] Petersen, T. S., & Ryberg, J. (2019). Applied ethics. In Oxford Bibliographies. Oxford University Press.

[5] Rini, R. A. (2015). Psychology and the aims of normative ethics.

[6] Wikipedia contributors. (2021, August 14). Normative ethics. In Wikipedia, The Free Encyclopedia. Retrieved 09:19, November 6, 2021, from https://en.wikipedia.org/w/index.php?title=Normative_ethics&oldid=1038781260

[7] Schick, T. (2009). Doing philosophy: An introduction through thought experiments.

[8] Star, D. (2015). Knowing better: virtue, deliberation, and normative ethics. Oxford Philosophical Monograph.

[9] "Teleological Ethics." Encyclopedia of Philosophy. Retrieved October 25, 2021 from Encyclopedia.com: https://www.encyclopedia.com/humanities/encyclopedias-almanacs-transcripts-and-maps/teleological-ethics

[10] Wikipedia contributors. (2021, September 28). Deontology. In Wikipedia, The Free Encyclopedia. Retrieved 10:08, November 6, 2021, from https://en.wikipedia.org/w/index.php?title=Deontology&oldid=1046941839

[11] Sinnott-Armstrong, Walter. "Consequentialism." The Stanford Encyclopedia of Philosophy, 2015. http://plato. stanford.edu/archives/win2015/entries/consequentialism/

[12] Larmore, C. (1996). The morals of modernity. Cambridge University Press.

[13] Carr, D., & Steutel, J. (2005). Virtue ethics and moral education. Routledge.

[14] Wikipedia contributors. (2021, September 20). Virtue ethics. In Wikipedia, The Free Encyclopedia. Retrieved 12:44, November 7, 2021, from https://en.wikipedia.org/w/index.php?title=Virtue_ethics&oldid=1045494526

[15] Kagan, S. (1992). The structure of normative ethics. Philosophical perspectives, 6, 223-242.

[16] Kagan, S. (2018). Normative ethics. Routledge.

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

[18] Cointe, N., Bonnet, G., & Boissier, O. (2016, May). Ethical judgment of agents' behaviors in multi-agent systems. In Proceedings of the 2016 international conference on autonomous agents & multiagent systems (pp. 1106-1114).

[19] Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The moral machine experiment. Nature, 563(7729), 59-64.

[20] Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017, February). Moral decision making frameworks for artificial intelligence. In Thirty-first aaai conference on artificial intelligence.

[21] Kim, R., Kleiman-Weiner, M., Abeliuk, A., Awad, E., Dsouza, S., Tenenbaum, J. B., & Rahwan, I. (2018, December). A computational model of commonsense moral decision making. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 197-203).

[22] Bentham, J. (1996). The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation. Clarendon Press.

[23] Ross, D., & Ross, W. D. (2002). The right and the good. Oxford University Press.

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