Universe Scientific Publishing

DOI:10.18686/ahe.v7i25.10101

Kant's Three Formulations and Machine Ethics: An Exploration of Morality Concern with AI Technology Based on "Groundwork of the Metaphysics of Morals"

Siyu Zhou[1], Pengxuan Huang[2]

1.Sun Yat-sen University, Boya (Liberal Arts) College ,Guangzhou City, Guangdong,China.

2.University of New South Wales, School of Computer Science and Engineering Sydney NSW 2502, Australia

Abstract: The Rapid development of Artificial Intelligence has stressed the need for research in machine ethics. This paper examines ethical and moral concerns on Artificial Technology by exploring Kant's Three Formulation of Categorical Imperative from "Groundwork of the Metaphysics of Morals" and illustrating its division from machine ethics. In the end, the paper concludes that it is doubtful for machines to achieve morality which leads to risks and challenges regarding AI technology.

Keywords: Groundwork of Metaphysics of Morals; Kant; Artificial Intelligence; Machine Ethics

1. Introduction

Artificial Intelligence (AI) has been widely researched and implemented in real work applications such as autonomous vehicles and voice assistance. However, in recent years, there has been an increased focus on the ethical and social implications of artificial intelligence technology. There exist ethical dilemmas around these technologies. For instance, the famous trolley problem is a significant ethical concern for autonomous vehicles. Based on the reading of Immanuel Kant's "Groundwork of the Metaphysics of Morals", this paper argues that feeding machines with ethical principles to resolve the so-called moral dilemma lacks rationality. The paper will base upon the second section of "Groundwork of the Metaphysics of Morals" - "Transition from popular moral philosophy to the metaphysics of morals" to illustrate why such automated machine ethics is risky and challenging to achieve. Meanwhile, research in an area called "automated ethics" or "machine ethics" has emerged to study these ethical issues. Susan Leigh Anderson and Michael Anderson defined machine ethics as research that is concerned with enabling machines to *"function in an ethically responsible manner through their own ethical decision making"* by feeding machines with ethical principles and procedures. (Anderson, M., Anderson, S.L.: Machine Ethics. 2011). Research in the development of machines such as Artificial Moral Agents(AMAs) intends to employ such machine ethics in autonomous machines. However, based on Kant's view in the "Groundwork of the Metaphysics of Morals", machine ethics are at odds with Kant's moral philosophy. Kant's moral philosophy cannot support machine ethics. In fact, it reveals risks and challenges with machine ethics.

2. The Three Formulations: Exploring the possibility from Kant's Three Formulations

In a nutshell, Emmanual Kant believes that rational, free and autonomous agents must perform ethical actions that conform to the categorical imperative. Categorical imperatives are moral duties that people ought to follow because they are rational beings. A crucial distinction between rational and irrational agents is the fact that rational agents are free. Humans have the free will or choice to violate moral laws, but rationality is the "*strength needed to subdue the vice-breeding inclinations."* (Kant, 1996, 6:376). Kant used the term "autonomy" to describe such ability, where autonomy is the "ground of dignity … of every nature." People can make decisions without external supervision, coercion or guidance. These qualities enable humans to be rational beings, and thus their actions will be

bounded by the categorical imperative.

Kant had different formulations of the categorical imperative. These are different ways to illustrate the concept, but they all refer to moral duty. The first and second formulations are as below:

1. Act only in accordance with that maxim through which you can at the same time will that it become a universal law. (Kant, 1998, 4:421)

2. So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means. （Kant 1998, 4:429）

The First formulation introduced a moral reasoning procedure to determine whether an action is morally permissible. For certain actions to be morally permissible, the action must become a principle for everyone to obey. If such action should not be performed by everyone, it cannot be universalisable, and thus the action is not moral. An example would be that lying is morally wrong because it is detrimental to our society if everyone does it. Thus, it cannot be universalisable. Kant believes humans can perform such reasoning as he says being autonomous allows people to "*choose only in such a way that the maxims of your choice are also included as a universal law in the same volition.*". (Kant,1998,4:440). They can identify and act according to the universal law using their rationalities. Thus, all autonomous, rational agents must be able to assess and reason the morality of actions so their actions can consistently conform to the categorical imperative.

In the Second Formulation, Kant illustrated that humans should treat others as "end" because rational nature, such as humans, *"exists as an end in itself*." (Kant, 1998, 4:429). Human intrinsic value, freedom and humanity should be respected. Treating someone as an end means respecting their dignity and never treating others solely for something they want. When people can respect such qualities of each other, they will be in relationships where participants trust others to act morally and ethically.

However, in today's world, people are more concerned with the utility of action to themselves rather than acting respectfully and morally. Our society's sense of ethics and responsibility for action is fading because of such attitudes. Behaviours such as deception are increasing. Social trust is broken when people often focus on self-interest and forget to act respectfully. Understanding Kantian ethics and trying to abide by the categorical imperative could relieve such concerns. Using different formulations to reason morality might be time-consuming and complex to perform consistently. However, the concept of categorical imperative and moral duty can give us a perspective that we should focus more on the morality of action itself. It provides us with a sense of duty and moral responsibility that has been lacking. People need to understand that some actions are not moral. Thus, rationality and strength are needed to resist the temptation to take such action. Actions such as deception are mainly due to such lack of respect and moral duties. When the sense of responsibility is strong, people tend to build trust among each other, which makes human interaction more enjoyable and relaxing. For one, I want to be in a world where humans, as rational beings, have trust and can use their rationality to engage in moral behaviours that adhere to the categorical imperative.

I believe the concern for real-life utility is the primary reason for the birth of AI. Scientists build AI technology partly due to the simple consideration of "real-life benefits". This sort of real-life benefits is similar to the subjective goal discussed in Kant's "Groundwork of the Metaphysics of Morals", where human desire makes them hope to use relative means (AI) to achieve an end(money, mass production etc.). However, as Kant pointed out, these ends are neither objective ends nor rational acts. Thus, for AI products that are originally based on the subjective goal of humans, their actions cannot be those of "rational beings, because the ability of AI endowed by humans serves non-rational and subjective ends. Therefore, AI's acts are responses to non-rational beings' subjective consciousness, not the act of rational beings. It can be seen that, from Kant's first and second formulation, the irrational nature of AI since its inception has been adjudicated.

Based on the first two formulations, we can obtain Kant's third formulation of the categorical imperative:

3. *"of the will of every rational being as a will giving universal."*（Kant 1998,4:431）

Kant argued that this will is not subject to the law of personal stake. For instance, if one only acts honestly due to the fear of punishment brought by society or the nation, then it violates this law. Therefore, actions that derived from consequences, are not

actions that conform to this law. People have a personal stake in their happiness, thus,Kant argues that the desires for personal happiness have a personal stake in their obedience to the law and the personal stake cannot be used to determine their will. With further inference, it can be noted that, rules that every rational being acts upon must be completely independent of any subjective interests, or that the rules should adhere to the principle of universal law. For example, if one regards kindness as a responsibility, then acting in a kind manner becomes a moral action. Humans as rational beings generated such moral laws, which in turn, enable humans to perform such acts that are of moral values. In the first section of the "Groundwork of Metaphysics of Morals", Kant stated that the concept of rules must be independent of all desires. Thus, when feeding AI with moral consciousness, the personal stake must be removed, making it a responsible agent. This responsible agent should act within the scope of responsibility in terms of motives and content, like a kind of rational being that disregards consequences. Obviously, it is not feasible. As mentioned above, one of the primary reasons for the birth of AI is to satisfy the desire and subjective goals of some people. Therefore, from the origin, the purpose of AI in itself is utilitarian and based on clear self-interests. However, if we want to ensure that the "moral laws" that are fed to AI are "moral", we must remove subjectivity and personal stake from it. Such transformation to AI, in Kant's view, contradicts the purpose of the creation of AI. Even for Humans as emotional and rational beings, we cannot call ourselves completely rational beings, thus it would be challenging to guarantee that the moral laws fed to AI are purely out of rationality.

3. Division between Machine Ethics and Kant's "Groundwork of the Metaphysics of Morals"

Based on the aforementioned Three formulations of categorical imperative, this paper considered the plausibility of AI morality and suggests that it is challenging to feed moral law into AI. In addition, machine ethics, which has been a crucial focus in academia, represents two types of ethical agents: implicit and explicit. In fact, both are separate from the relevant discussion from the 2nd section of "Groundwork of the Metaphysics of Morals". According to James Moor(2006, pp. 19-20), an Implicit ethical agent is programmed to promote ethical behaviour or at least avoid unethical behaviour. An explicit ethical agent can provide logical reasoning for its ethical judgment on particular actions. Both cases could be challenging and unlikely to use Kant's moral philosophy in underpinning machine ethics.

For both implicit and explicit agents, the machine ethics embedded within the agent would not possess the quality of freedom. Under Kant's definition, rationality gives humans a better understanding of freedom, that is, freedom to act under constraints of the categorical imperative and "*master one's inclinations when they rebel against the law*"(Kant,1996, 6:383). With the sense of duty always to obey moral obligations, humans will take responsibility for their actions under such obligations. Whether the machine is being programmed or has algorithms to judge the morality of actions, they are limited in decisions. The process of gaining and performing logical reasoning is carried out passively as agents are constrained to a predetermined rule or algorithm. There is no consciousness or inclination to commit morally impermissible actions. Thus, it can not be autonomous to subdue that inclination. This lack of freedom and autonomy illustrates the absence of rationality for machines. When such rationality is missing, machines would not be able to understand the meaning and significance of moral obligation since it is embedded in itself passively by humans. Thus, it could not be morally responsible in its decision when moral rules are input into logic for machines. The above qualities demonstrate the difference between humans and machines, suggesting that machines contradict the moral agents' criteria under Kant's moral philosophy.

If these "artificial moral agents" were rational and free under Kant's definition, people are not treating them as an end but only as a means. Kant believes that moral agents, like a human, needs to be treated with respect, as the second formulation of categorical imperative suggested. However, researchers are developing these agents to achieve goals, such as ensuring or improving human well-being, safety and ethical understanding. While explicit moral agents are developed to act ethically and provide such benefits, since they do not possess freedom, they do not possess dignity. Creating such moral agents that are free and perform ethical decisions autonomously under Kantian moral philosophy is still unlikely, as when we create it as a means to act ethically, which already contradicts the categorical imperative for the reasons above. Therefore it is challenging and unlikely that Kantian ethics can be fully

implemented in machine ethics, which creates concerning risks.

As machines are not subject to "their" rationality, they are subject to engineers' conscience and rationality. Machine designers are the ones that would require to act ethically and conform to the categorical imperative. Moral agents' ethical constraints or algorithm solely depends on their training data and design. Thus, the machine's behaviour would be unpredictable if engineers did not act in good faith. As these moral agents cannot be autonomous and free to engage in the universalisable test, the machine would perform false and immoral behaviour because training data were wrong in the first place. There could also be unethical designers who produce machines with malicious designs. Since the initial design constraints the machine, designers or hackers could deliberately make hostile and unethical changes to its ethical principles. The machine would act according to it to produce unexpected outcomes. The vulnerability of machine ethics to such malicious design is described as the term "corruptibility" of machine ethics by Stephen Cave. (Cave et al.,2019, pp.8) The corruptibility feature brings uncertainty and concern about what malicious behaviour could occur. It could be as immoral as hurting people based on certain discriminatory qualities in particular situations. This leads to another highly debated risk of machine ethics in AI technology, can machines be morally responsible if they commit such wrongdoings?

If machine ethics truly satisfies the Kantian ethical theory, it would only work if such robots can be held morally responsible, that is, be subject to moral praise and punishment. (Sparrow 2007, pp.71). However, as the paper previously suggested, machines cannot be morally responsible for their behaviours. When humans commit wrongdoing, they are responsible for such action. Punishment for intentionally causing fatal crashing typically involves a restriction on freedom, such as imprisonment. However, the case becomes more complex for machines. Even if machines cause fatal accidents, no punishment can be implemented since they do not have freedom and rationality. They cannot be responsible due to their contradiction to Kantian ethical principles. It leads to an increased debate on who should be responsible in these scenarios, especially in the autonomous vehicle industry. In 2018, a 49-year-old woman Elaine Herzberg was hit by an Uber self-driving car. A safety-driver operator was present to prevent accidents, and it is deemed that the fatal crash happened because of his negligence. At the moment, the operator is facing charges, while Uber company will face no charges as there is "*no basis for criminal liability."*(Cellan-Jones, 2020). When Uber appoints an operator to oversee the operation of autonomous vehicles, they transfer the responsibility to control vehicle behaviour to the operator, which suggests the vehicle's inability to be morally responsible. For autonomous vehicles, where decisions are mostly algorithmically based, designers could be feeding false or ethically-controversial logic in some instances. When the vehicle acts according to such logic and results in accidents, the designer should also be held responsible since their input indirectly leads to fatality. However, in Uber's case, they were left unblamed while the operator was the only one facing charges. It shows the need for more legislation or precise definition of roles and responsibilities regarding AI. Machines' inability to be morally responsible causes confusion about who is at fault and leads to alarming ethical risks and dilemmas. Unpredictability and unclear responsibility create fear among people. They become resistant to such technology when they know it cannot be morally responsible. Eventually, it could become a hindering force to the development of AI technology if the moral responsibility concern is not properly addressed.

4. Conclusion

Given the rapid development of AI technology in everyday applications, research into machine ethics will become a key focus in academia. Its initiatives will be valuable for AI technology design, building ethical frameworks and solving real-life practical problems with the technology. However, such machine ethics required further research and development. It involves the challenging process of extracting moral and ethical issues from real-life experience to developing moral philosophy suitable for AI and technology itself. As mentioned in the paper, Kant's "Groundwork of the Metaphysics of Morals" demonstrates the complexity of moral philosophy for humans. It is unclear, if the designer of AI technologies - humans, can remain rational and act morally. As a technology that is unable to transcend human existence, its potential for morality is in doubt. Then building a morality system tailored to AI technology would present even much greater challenges.

References:

[1] Anderson M and Anderson SL. Machine Ethics. Cambridge University Press, 2018.

[2] Cave S, Nyrup R and Vold K, et al. Motivations and Risks of Machine Ethics. Proceedings of the IEEE 2019; 107(3): 562–574, https://doi.org/10.1109/jproc.2018.2865996.

[3] Uber's Self-Driving Operator Charged over Fatal Crash. BBC News 2020 Sept 16, www.bbc.com/news/technology-54175359.

[4] Kant I. The metaphysics of Morals. New York: Cambridge University Press;1996.

[5] Kant I. Groundwork of the Metaphysics of Morals. Cambridge University Press; 1998.

[6] Moor JH. The Nature, Importance, and Difficulty of Machine Ethics. IEEE Intelligent Systems 2006; 21(4) : 18–21, https://doi.org/10.1109/mis.2006.80.

[7] Sparrow R. Killer Robots. Journal of Applied Philosophy 2007, 24(1): 62-77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.

[8] Watson J. The Philosophy of Kant Explained, The Philosophical Review 1909, 18(6):646. https://doi.org/10.2307/2177678.