Vilnius University Press

Contents lists available at Vilnius University Press

Problemos
2021, vol. 100, pp. 139–151

ISSN 1392-1126 eISSN 2424-6158
DOI: https://doi.org/10.15388/Problemos.100.11

Kantian Moral Agency and the Ethics of Artificial Intelligence

Riya Manna

Department of Humanities and Social Sciences,
Indian Institute of Technology Bombay
Email riya_manna18@iitb.ac.in
ORCID iD https://orcid.org/0000-0002-1222-2917

Rajakishore Nath

Department of Humanities and Social Sciences
Indian Institute of Technology Bombay
Email rajakishorenath@iitb.ac.in
ORCID iD https://orcid.org/0000-0003-0855-9709

Abstract. This paper discusses the philosophical issues pertaining to Kantian moral agency and artificial intelligence (AI). Here, our objective is to offer a comprehensive analysis of Kantian ethics to elucidate the non-feasibility of Kantian machines. Meanwhile, the possibility of Kantian machines seems to contend with the genuine human Kantian agency. We argue that in machine morality, '*duty*' should be performed with '*freedom of will*' and '*happiness*' because Kant narrated the human tendency of evaluating our 'natural necessity' through '*happiness*' as the end. Lastly, we argue that the Kantian '*freedom of will*' and '*faculty of choice*' do not belong to any deterministic model of 'agency' as these are sacrosanct systems. The conclusion narrates the non-feasibility of Kantian AI agents from the genuine Kantian ethical outset, offering a utility-based Kantian ethical performer instead.
Keywords: artificial intelligence, categorical imperative, choice, freedom of will, Kantian ethics, moral agency, utility

Kantiškasis moralės subjektas ir dirbtinio intelekto etika

Santrauka. Straipsnyje aptariami filosofiniai klausimai, susiję su kantiškuoju moralės subjektu ir dirbtiniu intelektu. Straipsnio tikslas – pateikti išsamią Kanto etikos analizę, kad būtų išaiškintas kantiškojo moralės subjekto kaip pareigos mašinos neįgyvendinamumas. Kantiškos mašinos galimybė, regis, dera su tikruoju kantiškuoju moralės subjektu. Straipsnyje teigiama, kad mašinų moralėje „pareiga“ turėtų būti atliekama su „valios laisve“ ir „laime“, nes Kantas rašė apie žmogaus polinkį „prigimtinę būtinybę“ vertinti „laimės“ kaip tikslo požiūriu. Galiausiai straipsnyje tvirtinama, kad kantiškoji „valios laisvė“ ir „pasirinkimo galimybė“ neturi nieko bendra su deterministiniu „subjekto“ modeliu, kadangi tai esą šventi dalykai. Daroma išvada, jog kantiškasis dirbtinio intelekto subjektas neįmanomas dėl pačių tikrosios Kanto etikos pagrindų, ir vietoj to siūlomas naudos kategorija besiremiantis kantiškosios etikos išpildytojas.
Pagrindiniai žodžiai: dirbtinis intelektas, kategorinis imperatyvas, pasirinkimas, valios laisvė, moralės subjektas, nauda

Received: 11/04/2021. **Accepted:** 28/06/2021

Introduction

AI technology has been invented as the augmented intelligence to humanity (Rossi 2016). However, over time, the goal has been shifted to invent autonomous ethical agents that can mimic human ethical decision-making processes without any human intervention. Susan Leigh Anderson and Michael Anderson have summarised this advancement and claimed, "Ideally, we would like to be able to trust autonomous machines to make correct ethical decisions on their own, and this requires that we create an ethic for machines" (Anderson & Anderson 2011: 1). The probability divides AI enthusiasts into two conflicting groups. One group believes in the feasibility of an entirely autonomous ethical artificial agent (strong AI supporter), and the other group negates the previous possibility (weak AI supporter). Generally, the term 'ethical' or 'moral' pertains to those who have the rationale to judge whether a particular action is agreeable or not. It means that the term is solely applicable to a sentient and rational being. Previously we believed that only human beings are sentient enough to be entitled to rational capacities. It seems unnatural to call any artefact an ethical or moral agent in the same sense (Wallach & Allen 2009: 55). Intelligent systems are close companions to our daily life, and their actions can influence our civilisation positively as well as negatively. The performances of these systems as independent moral agents will be a critical question that has to be evaluated accurately to enhance the positive impact of AI technology and minimise the adverse effect.

Here the fundamental elements are the ascription of moral agency on intelligent machines and the consistency of their moral decision-making process. Though Isaac Asimov had introduced three basic rules for robots to be moral agents, these would be insufficient for the guidance framework of future automatons (Asimov 1950). Nowadays, our daily life is much more dependent on autonomous systems and intelligent sensor-based prediction systems. Though AI scientists anticipate that a superhuman artificial general intelligence (AGI) can be invented in the future, the current development is not close enough as it was expected earlier. However, the result reveals that if the technology goes wrong or malfunctions in any way, it will certainly turn up as a *devil* for the entire civilisation. Elon Musk once claimed that "with artificial intelligence, we are summoning the demon".[1] Therefore, we need a deliberate ethical structure for AI agents from the beginning to enhance its positive impact on human society. To get a proper solution to this conflict, we must analyse the applicability of the term 'moral agent' to AI machines and its' difference from the moral agency of human beings. In seeking the resolution, we will primarily focus on Kantian ethics. In due course of the discussion, we will first focus on the concept of Immanuel Kant's 'moral agent', and then, we will be able to evaluate the moral agency of the AI system from a Kantian perspective. As we judge these systems

[1] Elon Musk has further demanded that AI technology has the potency to drive us to Mars and accomplish our daily duties smartly, but on the other hand, it can be more dangerous than any nuclear weapon if not handled with care by government regulatory authorities from the initial phase of application. He warned that AI could be turned into the biggest existential threat for the entire humanity. (https://www.cnet.com/news/elon-musk-we-are-summoning-the-demon-with-artificial-intelligence/ [accessed April 3, 2020]).

from their performance related to human beings, we shall assess them within our ethical evaluative paradigm. The anthropocentric evaluative paradigm can be helpful for the prospect of the Kantian machine. However, it seems to be conflicting to the supporters of posthumanism (Evans 2015: 378).

Kantian ethics is precisely an anthropocentric evaluative paradigm. Kant has not explicitly stated anything about non-anthropocentric moral agency (1997: 28-29).[2] However, philosopher Thomas M. Powers has argued that "A rule-based ethical theory is a good candidate for the practical reasoning of machine ethics because it generates duties or rules for action, and rules are (for the most part) computationally tractable" (Powers 2006: 46-47). Our paper aims to analyse the probability of a Kantian machine and justify whether it can be a genuine Kantian moral agent. To accomplish the goal, we shall discuss in this paper several Kantian notions like duty, categorical imperative, freedom of will, and faculty of choice. As Kant has not speculated any probability of artificial moral agency, we intend to extend the Kantian analysis of moral agency and equate it with machine morality. Kant acknowledged that humans are entitled to 'reason' as a faculty to judge unbiasedly; therefore, human beings can be considered as moral agents (2002: 17).[3] However, AI robots can be the perfect simulacrum of human moral agents without any conscious moral agency, and their rationality is algorithmic. It is dependent on the human programmer. Therefore, the moral agency of AI robots will be unlike genuine Kantian moral agency.

Kantian Moral Agency

Conventionally we denominate any human being as a 'moral agent' when s/he has moral consciousness. They can judge any action as legitimate or not concerning the welfare of human society. In this regard, Kant claimed in *Critique of Pure Reason* that only human beings could conquer the pathologically necessitated causal connection between actions and effects and 'transcendental freedom' is intrinsic to rational beings. In Kant's words (1998: 533):

> …it is this transcendental idea of freedom on which the practical concept of freedom is grounded, and the former constitutes the real moment of the difficulties in the latter, which have long surrounded the question of its possibility. Freedom in the practical sense is the independence

[2] Kant claimed, "Consciousness of this fundamental law may be called a fact of reason because one cannot reason it out from antecedent data of reason, for example, from consciousness of freedom (since this is not antecedently given to us) and because it instead forces itself upon us of itself as a synthetic a priori proposition that is not based on any intuition, either pure or empirical, although it would be analytic if the freedom of the will were presupposed; but for this, as a positive concept, an intellectual intuition would be required, which certainly cannot be assumed here". (*Critique of Practical Reason,* 5.31)

[3] Kant has claimed that human morality will be dependent on their faculty of ‚reason', which disarticulate the actions from the expectations of effects. He asserted that nothing other than the sense of 'duty' in itself, which occurs only in the rational being insofar as it, especially not the urge for the outcome, is the determining ground of the will. Therefore, it constitutes that pre-existing goodwill which we call 'moral sense', and which is already present in the person who acts in accordance with it. It cannot be derived from the outcome of her/his actions. (*Groundwork for the Metaphysics of Morals,* 4:402)

> of the power of choice from necessitation by impulses of sensibility. For a power of choice is sensible insofar as it is pathologically affected (through moving-causes of sensibility); it is called an animal power of choice (*arbitrium brutum*) if it can be pathologically necessitated. The human power of choice is indeed an *arbitrium sensitivum,* yet not *brutum* but *liberum,* because sensibility does not render its action necessary, but in the human being there is a faculty of determining oneself from oneself, independently by sensible impulses.

Here, Kant argued that the rest of the animal kingdom is necessitated by their instincts, hence will not be bestowed with moral agency. The pathological necessitation is enclosed by the scientific causal chain, which seems unbreakable within the sensible world. Further, Kant argued (1998: 533-534):

> …it is easy to see if all causality in the world of sense were mere nature, then every occurrence would be determined in time by another in accord with necessary laws, and hence – since appearances, insofar as they determine the power of choice, would have to render every action necessary as their natural consequence – the abolition of transcendental freedom would also simultaneously eliminate all practical freedom. For the latter presupposes that although something has not happened, it nevertheless ought to have happened, and its cause in appearance was thus not so determined that there is no causality in our power of choice such that, independently of those natural causes and even opposed to their power and influence, it might produce something determined in the temporal order in accord with empirical laws, and hence begin a series of occurrences entirely from itself.

The 'transcendental freedom' is the fundamental requirement of morality. Kant postulated a distinct type of parallel causality, which coexists with scientific causality. He entitled this novel causality as 'transcendental causality'. Any actions influenced by the latter will be free from any causal bondage of the sensible world. He claimed that our 'free will' is an expression of 'transcendental causality' (Sidgwick 1988: 405-412). Here Kant has introduced the notion of *'choice'* as another faculty, which helps us to surpass the scientific causal chain. It further assists us in performing moral deeds. He entitled human beings as *rational beings* because they are subjected to both phenomenal and transcendental rules simultaneously.

However, the faculty of 'choice' helps us overcome the sensible world's temptation and follows our moral consciousness. This faculty of 'choice' segregates us from other *pathological beings,* like animals. Kant argued that 'moral agency' requires the freedom to choose. Hence, those entitled to any one of these causalities cannot be bestowed as 'moral agents'. Animals are only subjected to sensible causality and lack of rational competence, which is required for moral understanding; therefore, they are devoid of moral agency.[4] Similarly, Kant claimed (2002: 30-31):

[4] Kant has not explicitly negated the scope of non-anthropocentric moral agency in his analysis. Rather he claimed that any being which possesses sufficient rational competency can hold 'moral agency'. However, it is quite evident that Kant has explicitly emphasised on human moral agency in his entire course of philosophy as he believed that non-human beings can be rational as well, but they are not sufficiently capable of overcoming the realm of 'sensibility' and think beyond it. This is why humans preserve a higher abode among all other organisms according to Kant. (*Critique of Pure Reason,* A 534/ B 562 - A 535/ B 563)

> A perfectly good will would thus stand just as much under objective laws (of the good), but it would not be possible to represent it as *necessitated* by them to lawful actions, because of itself, in accordance with its subjective constitution, it can be determined only place through the representation of the good. Hence for the *divine* will, and in general for a *holy* will, no imperatives are valid; the *ought* is out of place here, because the *volition* is of itself already necessarily in harmony with the law. Hence imperatives are only formulas expressing the relation of objective laws of volition in general to the subjective imperfection of the will of this or that rational being, e.g., to the human being.

In this regard, Kant claimed that if someone is solely subjected to transcendental causality, then s/he would not achieve the option to perform otherwise. Kant has denominated them as the 'holy will'. Their actions will necessarily follow the transcendental causality; hence there is no 'choice' to execute otherwise. The word 'ought' is only relevant to human performers who have the alternatives to act against their 'reason'. Still, they choose to follow rational thinking to be 'moral'. Kant introduced the notion of 'imperative' in this regard.

> Moreover, according to Kant, all moral rules should conform with the form of the categorical imperative, which is universal, thereby demarcating it from hypothetical imperatives. Kant narrated, "The categorical imperative would be that one which represented an action as objectively necessary for itself, without any reference to another end" (2002: 31). For him, hypothetical imperatives are dependent on an external condition for its applicability, which is 'happiness' or 'pleasure'. In contrast, categorical imperative bears its justification within itself. The self-evident nature makes the categorical imperative universal, inherently necessary, and indifferent to its actualisation,[5] whereas hypothetical imperatives are contrary to it. Our free will should be guided only by the categorical imperative to actualise moral actions. In Kant's words (2002: 63):

Natural necessity was a heteronomy of efficient causes; for every effect was possible only in accordance with the law that something else determined the efficient cause to causality; what else, then, could the freedom of the will be, except autonomy, i.e., the quality of the will of being a law to itself? But the proposition 'The will is in all actions a law to itself' designates only the principle of acting in accordance with no other maxim than that which can also have itself as a universal law as its object. But this is just the formula of the categorical imperative and the principle of morality: thus a free will and a will under moral laws are the same.

[5] Kant claimed, "The fact mentioned above is undeniable. One need only analyze the judgment that people pass on the lawfulness of their actions in order to find that, whatever inclination may say to the contrary, their reason, incorruptible and self-constrained, always holds the maxim of the will in an action up to the pure will, that is, to itself inasmuch as it regards itself as a priori practical. Now this principle of morality, just on account of the universality of the law-giving that makes it the formal supreme determining ground of the will regardless of all subjective differences, is declared by reason to be at the same time a law for all rational beings insofar as they have a will, that is, the ability to determine their causality by the representation of rules, hence insofar as they are capable of actions in accordance with principles and consequently also in accordance with a priori practical principles (for these alone have that necessity which reason requires for a principle). It is, therefore, not limited to human beings only but applies to all finite beings that have reason and will and even includes the infinite being as the supreme intelligence" (1997: 32)

Kant upheld an impassable division between the categorical imperative and hypothetical imperative. He famously disseminated three maxims as three different expressions of the categorical imperative. According to him, hypothetical imperatives are not the efficient cause of our actions; instead, they work as the formal cause. That means hypothetical imperatives could not certify and necessitate our actions. Only categorical imperative works as an efficient cause and can determine our practical actions. Hypothetical imperatives depend on the external condition of 'happiness' or 'pleasure'. It loses its desirability if we dissolve its outcome. Kant asserted (1998: 677):

> Happiness is the satisfaction of all of our inclinations (*extensive,* with regard to their manifoldness, as well as *intensive,* with regard to a degree, and also *potensive,* with regard to duration). The practical law from the motive of happiness I call pragmatic (rule of prudence), but that which is such that it has no other motive than the worthiness to be happy I call moral (moral law). The first advises us what to do if we want to partake of happiness; the second commands how we should behave in order even to be worthy of happiness.

In this regard, Kant segregated freewill from the heteronomous will and argued that the latter remains integral with an upshot of 'self-love' or 'one's own happiness'. Further, he claimed (1997: 19-20):

> All material practical principles as such are, without exception, of one and the same kind and come under the general principle of self-love or one's own happiness… Now, a rational being's consciousness of the agreeableness of life uninterruptedly accompanying his whole existence is happiness, and the principle of making this the supreme determining ground of choice is the principle of self-love. Thus all material principles, which place the determining ground of choice in the pleasure or displeasure to be felt in the reality of some object, are wholly of the same kind insofar as they belong without exception to the principle of self-love or one's own happiness.

It is contrary to the Kantian notion of 'duty' and genuine moral agency. Only free will has the authority to usher a righteous deed and perform following one's duty. Sometimes humans are considered as pathological beings because they often encountered both heteronomous and autonomous will (Kleingeld 2010: 55-72). But our rationality keeps commanding us to follow our free will and operate morally. Therefore, the Kantian moral agency requires free will, faculty of 'choice' and 'autonomy', sense of 'duty', and categorical imperative to execute concrete moral actions. Now, if we quest for its applicability to AI, we must start with the analysis of 'artificial moral agency' and its features.

Artificial Moral Agency

Today's AI technology is primarily dependent on initial programming and big data sets. However, AI scientists anticipate that *Artificial General Intelligence (AGI)* will be able to self-programme and be a comprehensive simulacrum of human intelligence. They entitled it as the time of 'singularity' (Tegmark 2017: 261). 'Singularity' is explained as

an intelligent explosion, which will be powerful enough to replicate the human moral agency as well as the brain itself (Kurzweil 2005: 17). This claim postulates the ethical decision-making process as the effect of brain states; hence, it can be simulated by artificial neurons as well. Precisely, if an artificial neuronal structure can imitate the biological brain state, which is responsible for a particular moral deed/thought, then reproducing that moral act in artificial performers will be spontaneous. Therefore, the supporters of 'singularity' believe that superintelligent AI agents will be moral agents just like any rational human being. It is justified to hold them responsible for their actions (Kurzweil 2005: 314).[6] However, AI's present development trend does not ensure any particular anticipated time of 'singularity' so far.

In this regard, if we try to analyse the present status of artificial moral agency, then we shall find that these systems follow hypothetical imperatives. These imperatives are encrypted as a command at the coding level by some human programmers. That is to say, the programmer must prefer some specific set of human values while choosing the correct command for that intelligent machine. Undoubtedly, that specific human value is *happiness or pleasure* (Bostrom 2014: 227). It entails that AI robots are performing the actions following the rules which are encrypted in them. Programmers prefer those rules because they certify to provide pleasure or happiness to the user of that intelligent machine. From a Kantian approach, we can claim that the actions of these artificial agents are due to the effect of heteronomous 'will' determination. Precisely, here the external factor as the determinant of 'will' is the 'happiness or pleasure' of human users (Johnson 2006: 202).

We acknowledge that the concept of 'happiness or pleasure' is entirely subjective. What is pleasurable to one might not be the same for the other. Here, the agency of the AI robots will depend on its user's preference, which sometimes might be of some moral inclinations as well. This autonomy of performance might not be as robust as human agency, but still, we can consider it autonomous to some extent (Nath 2009, 2020; Sullins 2006: 25). The artificial agency conveyed the focus on data and its collaborations using cybernetic as well as computation theory. However, the agency lies in the broader perspective of the analysis of the shared relation rather than mere statistical data (Farnsworth 2017: 237). We tend to maintain human moral agency as superior because we believe it includes some intervening layers that the artificial moral agency might not ever achieve. It necessitates us to judge AI's action as an expression of its human programmer's values (Hall 2011: 513). So far as we know, the entire inference is based on the disparity between conscious moral agency and non-conscious moral agency. However, this disparity does not work as the only hindrance to the prospects for Kantian machines. In this regard, we shall next discuss our first argument from the perspective of Kantian 'duty'.

[6] In his book *Singularity is Near, author* Ray Kurzweil has used a quotation of philosopher Nick Bostrom: "Substrate is morally irrelevant, assuming it doesn't affect functionality or consciousness. It doesn't matter, from a moral point of view, whether somebody runs on silicon or biological neurons (just as it doesn't matter whether you have dark or pale skin). On the same grounds that we reject racism and speciesism, we should also reject carbon-chauvinism, or bioism" (the quotation from *Ethics for Intelligent Machines: A Proposal*, 2001).

Argument from Duty

Some AI thinkers believe that a Kantian machine will be a perfect example of a performer who can unbiasedly follow the 'duty'. It will be devoid of any emotional partialities like biological beings. For instance, some researchers have claimed that Kantian machines can be the opposite of 'Lie Bots' from the background of performing duties (Bendel et al., 2017: 7-11). That is to say, the plausibility of 'objectivity' is the crucial feature that compels scientists about the possibility of Kantian AI. On this issue, Thomas M. Power has claimed, "In fact, humans suffer from 'weaknesses of the will', as Aristotle called them, that shouldn't be a problem for a machine: once it reaches a conclusion about what it ought or ought not to do, the output will follow automatically" (Powers 2006: 46). According to his claim, the Kantian sense of '*objectivity*' in performance seems feasible for automatons. To find whether or not a Kantian 'duty' can be performed by an AI agent, we must deep dive into the Kantian analysis of actions.

Kant has segregated 'actions' into three diverse categories, i.e., immoral action, amoral action, and moral action, in his *Groundwork for the Metaphysics of Morals*. Here moral action refers to that action which emanates from 'duty' itself and nothing else, and immoral action lies at the opposite, which contradicts 'duty'. We will concentrate precisely on the notion of amoral actions concerning AI's performance of 'duty' as an objective performer. Kant has narrated amoral action as a deed, performed merely in conformity with 'duty', not essentially necessitated from it. As Kant argued (2002: 13):

> It is much harder to notice this difference where the action is in conformity with duty and the subject yet has besides this an immediate inclination to it. E.g., it is indeed in conformity with duty that the merchant should not overcharge his inexperienced customers, and where there is much commercial traffic, the prudent merchant also does not do this, but rather holds a firm general price for everyone, so that a child buys just as cheaply from him as anyone else.

According to him, amoral actions are not motivated to achieve happiness, appreciation, or other rewards. The performance of this action is used as a path to achieve those goals, as mentioned earlier. As these actions are not performed from 'freedom of will' and the sense of 'duty', they cannot be of any moral significance.

The performance of an AI agent seems to be of moral significance, sometimes concerning its outcome. As we have seen, in the case of artificial moral agency, AI agents are motivated by the happiness principle (happiness of the user) as a goal to achieve, and their duty is measured in terms of whether or not it conforms with the purpose (Bostrom 2014: 227). From the Kantian perspective, we can claim that AI's moral deeds are not generated from the 'freedom of will' and the sense of 'duty' itself. These systems perform their actions as per the programmer's command; hence we can claim their actions as amoral actions that run in accordance with 'duty' (here 'duty' means perform according to its programming level). There is no choice for these systems to do otherwise. The term itself indicates amoral actions are neither permissible nor impermissible. It neither violates one's duty nor conforms to the same. It is neither ethically better nor worse than any other action. It is simply morally unevaluated.

AI agents might be better performers than human beings to a restricted extent, but their actions are amoral and hold a lower magnitude of value. The commands are encrypted in AI's coding to accomplish some action, which will be helpful to achieve our needs. Instead of being a Kantian machine, these intelligent systems can be better evaluated as happiness provider machines. Kant has claimed that we tend to assess our 'natural necessity' by our happiness as an end.[7] In the case of an AI agent, 'happiness' denotes the moral quality of the action concerning its user, the accuracy of the predictions, and its overall impact on the entire human society (Kaplan 2016: 75). However, all these do not sound coherent with the Kantian moral action performed from 'freedom of will' and necessitated by categorical imperative. We will discuss these issues in the following sections of the paper.

Argument from Categorical Imperative

The *categorical imperative* is the unconditional command, which is an end in itself, and one should follow it to perform one's own 'duty'. In Kant's words , "So act as if the maxim of your action were to become through your will a universal law of nature" (2002: 38). The categorical imperative is an *apriori* truth and works as a presupposition of our moral actions. That means it works as the 'form' of moral action, whereas a particular moral action works as the content. We tend to analyse whether or not our specific action fits in the structure of *categorical imperative*. The categorical imperative, to be sure, is a command. But it is a rule for action, not an act itself. Moreover, its application does not *make* an act moral but rather shows whether a contemplated action is ethical or not. We can analyse that any deed follows a particular pattern at the time of occurrence, i.e., Intention> Action> Consequences. The *categorical imperative* corresponds with our sense of 'duty', which is emphasised in the first two sections (intention and action). According to Kant (2002: 46):

> …the human will is a categorical imperative, then it must be such from the representation of that which, being necessarily an end for everyone, because it is an *end in itself*, constitutes an objective principle of the will, hence can serve as a universal practical law. The ground of this principle is: Rational nature exists as an end in itself.

The intention part plays a fundamental role in defining moral action. Following the Kantian sense of unconditional 'duty', the categorical imperative command does not judge any act from its consequence.

There is another imperative that emphasises the consequence part, i.e., hypothetical imperatives. Hypothetical imperatives hold the format of 'if-then', which describes doing a specific action if someone wants to achieve a particular goal. Here the goals are determined by 'heteronomy of will', which contradicts Kantian 'autonomy of will' or

[7] Kant used this claim to distinguish 'assertoric' imperatives from the 'problematic' imperatives and argued for the imperfect duties, such as, helping others achieve specific and diverse goals. (*Groundwork for the Metaphysics of Morals*, 4:415 – 4:416)

'freedom of will' (Frankfurt 1971: 6-7). As our heteronomous will is subjective, it cannot necessitate a particular moral action. Hence, in the case of 'hypothetical imperatives', it does not command us to act just because it is our 'duty'. Kant has narrated (2002: 45):

> The subjective ground of desire is the incentive, the objective ground of volition is the motive; hence the distinction between subjective ends, which rest on incentives, and objective ones, which depend on motives that are valid for every rational being. Practical principles are formal when they abstract from all subjective ends; but they are material when they are grounded on these, hence on certain incentives. The ends that a rational being proposes as effects of its action at its discretion (material ends) are all only relative; for only their relation to a particular kind of faculty of desire of the subject gives them their worth, which therefore can provide no necessary principles valid universally for all rational beings and hence valid for every volition, i.e., practical laws. Hence all these relative ends are only the ground of hypothetical imperatives.

Instead, it directs us to act in a specific way to achieve a particular goal, which means it will not be universally applicable to all rational beings as an unconditional command.

Now, as far as the performance of an AI agent is concerned, it works according to hypothetical rules. These rules are encrypted in the command system, and it navigates the system to work in a particular technique to accomplish specific goals. It is speculated that if AI robots can successfully execute moral actions as an intelligent computing device, they can be a 'hypothetical moral agent' at their best (Powers 2009). Indeed, this 'hypothetical moral agent' will challenge the Kantian conception of *autonomy* of moral agency. Even if we reformulate *categorical imperative* to make it computable, we can encode it as an unconditioned command in the computing system. It can be encoded in a hypothetical command, which consists of actions to achieve some set of goals (Powers 2006: 47-48). Therefore, we can claim that computable categorical imperative is converting itself into a hypothetical imperative; hence, it is becoming redundant.

Suppose we assume that in the future, categorical imperative can be embodied successfully in AI machines and create a Kantian moral agent. In that case, indeed, all machines will not hold the same amount of moral rationality. According to their moral rationality, we must predetermine some rules to make them implementable in the machines (Powers 2006: 51). Though AI scientists are not even close enough to a machine that can successfully replicate a Kantian moral agent and act according to the categorical imperative. According to Kant, the two foundations of the categorical imperative are 'freedom of will' and 'faculty of choice'. We will discuss these issues in the next section.

Argument from 'Freedom of Will' and 'Faculty of Choice'

Kantian 'freedom of will' and 'faculty of choice' are the foundation of his conception of 'duty' and 'autonomy'. In the Kantian moral agency section, we have already discussed how Kant has depicted 'freedom of will' as a transcendental causality that coexists with scientific causality (1998: 533-534). 'Faculty of reason' works as the force which assists

any rational being to follow the transcendental causality of 'freedom of will'. In literal meaning, 'freedom' means 'freedom to', whereas, for Kant, freedom was 'freedom from'. Here 'freedom from' indicates freedom from the slavery of the sensible realm. Kant claimed (1998: 542):

> …everything which has happened in the course of nature, and on empirical grounds inevitably had to happen, nevertheless ought not to have happened. At times, however, we find, or at least believe we have found, that the ideas of reason have actually proved their causality in regard to the actions of human beings as appearances, and that therefore these actions have occurred not through empirical causes, no, but because they were determined by grounds of reason.

Kant believed that 'freedom of will' belongs to the intelligible world, which we can only feel but cannot grab through empirical evidence. It can be supposed through our moral actions, except that everything which follows the scientific causal chain is sacrosanct.

If we closely observe the mechanism of AI agents, then we can comprehend how 'free will' works for an AI agent. It has been massively debated that an AI system could not be entitled to an authentic free will without any minimal conscious agency (Stuart & Dobbyn 2002: 410). These systems could not go beyond their encrypted commands and perform randomly like sentient beings. In this regard, it should be mentioned that here we contemplate 'freedom of will' as 'freedom to perform', which denotes the ability to self-determine between different prospects (Omoregie 2015: 5-6). AI agents are not free enough to perform beyond what is written in their command system. All possible options are related to our practical world and shall not be determined from the Kantian transcendental realm (McCarthy 1995). Hence, without the conscious choice to transcend the sensible realm, one could not feel the Kantian 'freedom of will'. As Kantian 'freedom of will' works as the foundation of 'freedom to', AI agents will not be fit enough to acquire 'freedom to' from the Kantian prospect.

AI systems are deterministic models of agency that do not exceed its initial programming. That is to say if the Kantian machine would exist, then we could not judge them as per their intentions behind any action, just like human beings (Wallach et al., 2008: 574). In this regard, Kant claimed (1997: 106):

> The human being (and with him every rational being) is an end in itself, that is, can never be used merely as a means by anyone (not even by God) without being at the same time himself an end, and that humanity in our person must, accordingly, be holy to ourselves: for he is the subject of the moral law and so of that which is holy in itself, on account of which and in agreement with which alone can anything be called holy. For this moral law is based on the autonomy of his will, as a free will which, in accordance with its universal laws, must necessarily be able at the same time to agree to that to which it is to subject itself.

Kant has entitled human beings the highest type of moral agency because they have the rational capacity, which can be the end in itself. They can overcome the boundaries of the sensible realm and choose not to be deviated by any empirical sensations. According to Kant, human beings deserve respect because they can set their end in itself through

their rationality. It necessitates that the 'faculty of choice' navigates human beings as the driving force towards 'freedom of will'. It is evident that Kantian 'faculty of choice' does not belong to any deterministic model of agency as these systems are sacrosanct. Therefore, it can be concluded that any Kantian machine could not acquire Kantian 'faculty of choice' and 'freedom of will' ever.

Conclusion

We can conclude that human beings possess a conscious moral understanding, which helps them to judge their novel situations. However, we can raise the concern that without a 'subjective' feeling of moral agency, which is precisely the critical topic of contemporary debates in AI (Nath & Sahu 2020: 103-111). Our value theories are profoundly related to our socio-cultural background. However, Kant has maintained that these socio-cultural impacts belong to the sensible realm. If we claim that a Kantian machine is possible, then it must obtain all the characteristics of a Kantian moral agent. We have discussed in our three arguments that these AI agents, as Kantian machines, could not acquire the features of the Kantian moral agency so far. The fundamental components of Kantian ethics lead towards a focus on self-determining human competence for rule-making and rule adherence. These components exemplify the fundamental ways in which human attributes and competencies, such as practical reasoning, exercising judgment, self-reflection, and deliberation, allow for the construction of moral rules that are capable of universalisation. Such human attributes and competencies are non-existent in artificial intelligence and automatons. It entails that human agency must be at the forefront of designing and holding accountability for these machines' ultimate moral behaviour.

These systems tend to be judged as Kantian agents from the perspective of the utility of their actions. That is why it has been speculated that Kantian ethics shall not be sufficient to build an equitable AI agent. Many AI enthusiasts have proposed the fusion of different ethical theories (Kantian theory, utility theory, or virtue ethics theory) and both 'top-down' and 'bottom-up' (these two approaches are famously used to implement ethical guidelines for AI system) approaches to construct the legitimate, ethical framework for AI (Etzioni & Etzioni 2017: 413-418).

Kantian ethics provide an anthropocentric attitude to formulate moral rules as Kantian ethics is an exclusively rule-based approach to moral agency. It seems to be a challenging prospect to make these rule-based values computable. In this context, AI philosopher Nick Bostrom has claimed that "Goal system engineering is not yet an established discipline. It is not currently known how to transfer human values to a digital computer, even given human-level machine intelligence." (Bostrom 2014: 253). However, we can anticipate that goal system engineering will soon be possible, and future development of the AI field will help advance it further. A righteous AI agent can be achieved as a result of future innovations.

References

Anderson, M. and Anderson, S. L., 2011. General Introduction. In *Machine Ethics,* M. Anderson and S. L. Anderson (eds.), New York: Cambridge University Press, 1-4.

Asimov, I., 1950. *I, Robot.* USA: Gnome Press.

Bendel, O. et al., 2017. Towards Kant Machines. In *The 2017 AAAI Spring Symposium Series*. Palo Alto: AAAI Press, 7-11.

Bostrom, N., 2014. *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Etzioni, A. and Etzioni, O., 2017. Incorporating Ethics into Artificial Intelligence. *Journal of Ethics,* 21(4): 403-418.

Evans, W., 2015. Posthuman Rights: Dimensions of Transhuman Worlds. *Revista Teknokultura*, 12(2): 373-384. DOI:https://doi.org/10.5209/rev_TK.2015.v12.n2.49072.

Farnsworth, K. D., 2017. Can a Robot Have Free Will?. *Entropy*, 19(5), No.237: 237-258. DOI:10.3390/e19050237.

Frankfurt, H. G., 1971. Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, 68(1): 5-20.

Hall, J. S., 2011. Ethics for Self-Improving Machines. In *Machine Ethics,* M. Anderson and S. L. Anderson (eds.), New York: Cambridge University Press, 12-523.

Johnson, D. G., 2006. Computer Systems: Moral Entities But Not Moral Agents. *Ethics & Information Technology*, 8(4): 95-204.

Kant, I., 1997 [1788]. *Critique of Practical Reason* (M. Gregor trans.). New York: Cambridge University Press.

Kant, I., 1998 [1781/1787]. *Critique of Pure Reason* (P. Guyer and A. W. Wood trans.). New York: Cambridge University Press.

Kant, I., 2002 [1785]. *Groundwork for the Metaphysics of Morals* (A. W. Wood, trans. & ed.). New Haven, London: Yale University Press,

Kaplan, J., 2016. *Artificial Intelligence: What Everyone Needs to Know.* New York: Oxford University Press.

Kleingeld, P., 2010. Moral Consciousness and the 'fact of reason'. In *Kant's 'Critique of Practical Reason': A Critical Guide,* A. Reath and J. Timmermann (eds.), Cambridge: Cambridge University Press, 55-72.

Kurzweil, R., 2005. *The Singularity Is Near.* London: Duckworth Overlook.

Nath, R., 2020. Alan Turing's Concept of Mind. *Journal of Indian Council of Philosophical Research*, 37(1): 31-50. https://doi.org/10.1007/s40961-019-00188-0.

Nath, R. and Sahu, V., 2020. The Problem of Machine Ethics in Artificial Intelligence. *AI & Society*, 35(1): 103-111. https://doi.org /10.1007/s00146-017-0768-6.

Nath, R., 2009. *Philosophy of Artificial Intelligence: A Critique of the Mechanistic Theory of Mind*. USA: Universal-Publishers.

Omoregie, J., 2015. *Freewill: The Degree of Freedom Within*. UK: Author House Publication.

Powers, T. M., 2009. Machines and Moral Reasoning. *Philosophy Now Magazine*, 72: 15-16.

Powers, T. M., 2006. Prospects for A Kantian Machine. *IEEE Intelligent Systems Journal*, 21(4): 46-51.

Rossi, F., 2016. Artificial Intelligence: Potential Benefits and Ethical Considerations, *Briefing Paper to the European Union Parliament Policy Department C: Citizens' Rights and Constitutional Affairs European Parliament* (October) <https://www.europarl.europa.eu/RegData/etudes/BRIE/2016/571380/IPOL_BRI(2016)571380_EN.pdf>

Sidgwick, H., 1988. The Kantian Conception of Free Will. *Mind,* 13(51): 405-412.

Stuart, S. A. J. and Dobbyn, C., 2002. A Kantian Prescription of Artificial Conscious Experience. *Leonardo*, 35(4): 407-411.

Sullins, J. P., 2006. When Is a Robot a Moral Agent?. *International Review of Information Ethics,* 6(12): 23-30.

Tegmark, M., 2007. *Life 3.0: Being human in the age of Artificial Intelligence.* New York: Knopf.

Wallach, W. et al., 2008. Machine Morality: Bottom-up and Top-down Approaches for Modeling Human Moral Faculties. *AI & Society*, 22(4): 565-582. https://doi.org/10.1007/s00146-007-0099-0

Wallach, W. and Allen, C., 2009. *Moral Machines.* New York: Oxford University Press.