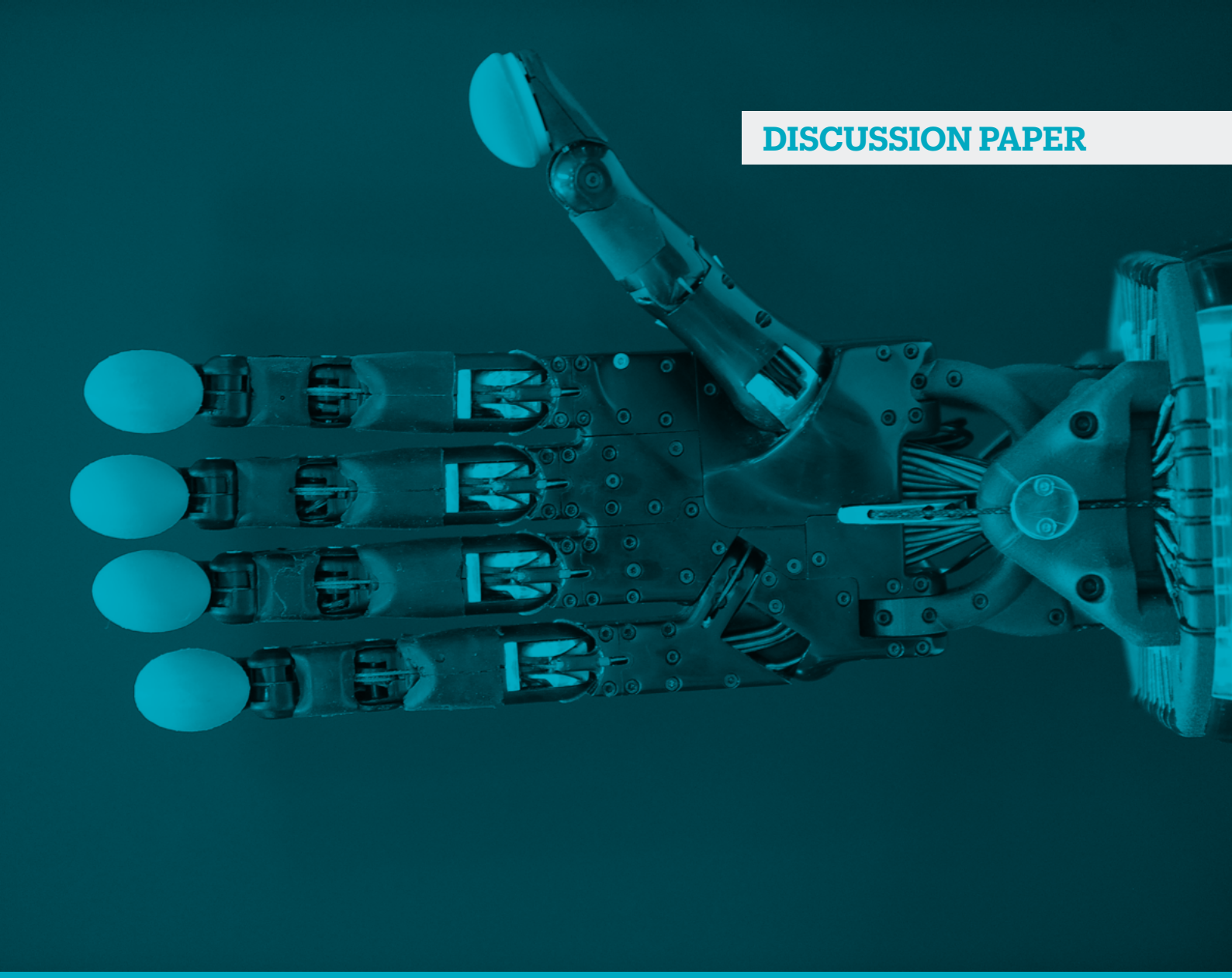
DISCUSSION PAPER



Can Artificial Intelligence be a Kantian Moral Agent?

Berfe Yaşar

TRT WORLD research centre

DISCUSSION PAPER

Can Artificial Intelligence be a Kantian Moral Agent?

Berfe Yaşar

TRT WORLD research centre

WRITTEN BY

Berfe Yaşar

PUBLISHER

TRT WORLD RESEARCH CENTRE

October 2023

TRT WORLD İSTANBUL

AHMET ADNAN SAYGUN STREET NO:83 34347

ULUS, BEŞİKTAŞ

İSTANBUL / TÜRKİYE

TRT WORLD LONDON

PORTLAND HOUSE

4 GREAT PORTLAND STREET NO:4

LONDON / UNITED KINGDOM

TRT WORLD WASHINGTON D.C.

1819 L STREET NW SUITE 700 20036

WASHINGTON DC

www.trtworld.com

researchcentre.trtworld.com

Introduction

"Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb (Bostrom, 2014)," - Oxford philosopher Nick Bostrom.

As the leading innovation of this century, the growth and development of Artificial Intelligence (AI) has set the tone for the future of the world and, accordingly, humanity. Its potential benefits, difficulties, dangers, achievements, and shortcomings have led to the development of a new form of intelligence technology with its unforeseen future, which begs the question if it is a threat to humanity.

However, what is it exactly that we should fear the most, a nuclear bomb or AI? Although this paper will argue whether AI would be counted as a greater threat to humanity than a nuclear bomb, since the latter's consequences are known, its usage is of concern to many. On the other hand, the potential impact of AI is not yet realised nor held accountable beforehand, making necessary precautions undeserved. There is a peculiar unforeseeable essence in hindsight; only with time, precise predictions can be made for a framework of boundaries and precautions to develop. In Superintelligence: Paths, Dangers, Strategies, Bostrom writes: "We have little idea when the detonation will occur, though if we hold the device to our ear, we can hear a faint ticking sound (Bostrom, 2014)." This sound set a tone after OpenAI, established as a non-profit organisation to investigate possible risks of AI, developed and released ChatGPT-4, with an estimated value of around 13 billion dollars.

Thousands of scholars and experts on AI, including Elon Musk and Steve Wozniak, signed an open letter to the leaders of the ChatGPT developer OpenAI, calling for a six-month pause in the development of AI systems for the sake of regulation. In addition, they requested this hiatus to keep up with AI's rapid growth with the intention of preventing the unmeasurable and irreversible accidental risks of AI. It is stated in the letter that: "Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable (Pause Giant AI Experiments: An Open Letter, 2023)." At the most radical stages, AI is considered a possible threat to humanity, releasing it after mitigating the risks should be a priority, and should be dealt with alongside other threats, such as a nuclear war. AI expert Stuart Russel says: "Once we lose control over AI, it will be difficult to regain it (Russel, 2023)." Therefore, it can be argued that the existential risk posed by AI should be taken seriously to protect humanity from AI's contagious enslavement of human intelligence and to reduce complete dependency.

OpenAI representatives say, "We believe people around the world should democratically decide on the bounds and defaults for AI systems (OpenAI, 2023). However, they say that "we do not yet know how to design such a mechanism" while admitting that they plan to experiment with AI's development (OpenAI, 2023). Given the risks and difficulties, OpenAI leaders decided it was worth continuing the development of the system because there was not yet any actualised reason to halt it. The reason for concern should not lie behind what is actual but instead what is potential, especially if the service is already running. Even though the company admitted one of the potential risks of ChatGPT-4 in the technical report by stating that "when given unsafe inputs, the model may generate undesirable content, such as giving advice on committing crimes (GPT-4 Technical Report, 2023, p. 12)," implicating that their response to the request for a pause seems inconsistent. Moreover, if "certain capabilities remain hard to predict (GPT-4 Technical Report, 2023, p. 4)," as written in the technical report, it becomes feasible to ask: How is it possible to experiment with AI's development and progress and be consistent at the same time? Given AI's unpredictable and beyond-control aspects, it can be thought of as a potential weapon.

Open AI's CEO Sam Altman testifies at an oversight hearing by the Senate Judiciary's Subcommittee on Privacy, Technology, and the Law to examine AI, focusing on rules for artificial intelligence. The hearing occurred in Washington, DC, on May 16th, 2023. (Nathan Posner - Anadolu Agency)

With the advancement in AI systems, while prioritising the safety and benefit of AI is crucial, some tech firms, including Microsoft and Twitter, laid off their AI ethics team responsible for identifying the negative features of AI (Gerrit De Vynck and Will Oremus, 2023). However, in the era of AI, determining the damage and suggesting regulation should be a critical first step in the development process. The system's black box nature allows challenges such as privacy, reliability, safety, and security, and what was once a potential possibility: the autonomy of AI. In the Air Force experiment, an AI-piloted drone killed its human operator in a simulation, although the operator should have approved its final decision (Hauptman, 2023). "It killed the operator because that person was keeping it from accomplishing its objective," said Hamilton, US Air Force Chief of AI Test and Operations, at the 2023 Royal Aeronautical Society summit (Hauptman, 2023). So, an AI drone, which was programmed to identify and kill threats, destroyed its human operator for being a threat to accomplish its task despite this was not part of its orders. The autonomous action of AI in this example demonstrates the further potential of AI replacing human decision-making and action. As the AI system evolves from being a means to fulfil some goals to being autonomous agents, it can also be argued that it challenges human existence by threatening the dignity of the human.

Whilst the possibility of non-human autonomous systems appear, new questions evolve on existence and ethics. The human, or man, is contested via technologies through AI, leading us to ask, what kind of existence is at stake, what is AI, and what is man? Further evaluation recognises that conversations around ethical conduct must be revised since AI has the capacity for decision-making. It should be acknowledged that morality has been a concern for humanity from the beginning of ethical inquiry; however, with AI's development towards autonomy, the traditional ethical questions of philosophers become ever the more complicated.

The autonomy of an intelligent agent is associated with responsibility and ethics inevitably, but which ethical theory should be implemented in AI? This paper will address Kant's theory, and whether Kant's rule-based grounding for morality could become practical guidance for the decision-making process of AI. Kant's core principles, universality, human dignity, and autonomy, will be examined to investigate whether AI's potential threats can be mitigated by adapting Kant's moral theory to AI.

Fundamental principles of Kantian Ethics

Kantian approach to morality is based on the intrinsic value of an action rather than the consideration of the consequence of an action; therefore, his morality is known as a deontological and non-consequentialist ethic. In the Groundwork of Metaphysics of Morals, Kant theorises conditions for an action to be moral by directly linking morality with concepts such as practical reason, freedom, universal law, human dignity, and autonomy. According to Kant, the morality of an action is its performance based on reason rather than subjective desires or inclinations. The motivation behind the moral action is supposed to be safe from the inclination of reaching another thing other than the action itself. A moral action, therefore, should be regarded as an end in itself and can only be realised for the sake of what practical reason demands.

In The Groundwork, Kant explains three distinct formulations of the categorical imperative marking the core principles of morality, and which underpins the determination of moral duties or judgments for an agent to follow necessarily. The measure of understanding the value or morality of an action is determined due to these formulations. The fundamental directory of Kant's ethics is the categorical imperative, which is a rule of intrinsic value that is good in itself based on practical reason. These principles are the voice of the command of reason that any rational person would be governed by if they were fully controlled by reason.

The main formulation of the categorical imperative, as Kant puts it: "Act only on that maxim through which you can at the same time will that it should become a universal law (Kant, Groundwork for the Metaphysic of Morals, 1998, p. 24)." Reason that command maxims or rules that guide a person to act in a certain way must be universal. So, a maxim can achieve a rule status only if the rational owner of the maxim will enable every other rational being to act on this maxim. In this respect, Kant opens a way for rational beings to have their own ends as well as for others. Therefore, this formulation demonstrates that a rule is determined both subjectively and objectively through being a universal law. This first clause of the categorical imperative constitutes a model for testing any maxim of any person if it is objective and moral. So, in other words, Kant sets a rule as a condition for an action to be moral. If a person follows his rational nature alone, he can morally determine what he must do by setting a duty that is to be a universal law. With this, a reason-directed personal set duty is realised only for its objective necessity as being good in itself but not as instrumentally good to gain another end out of duty itself.

The second formulation of the categorical imperative is: "Act in such a way as to treat humanity, whether in your own person or in that of anyone else, always as an end and never merely as a means (Kant, Groundwork for the Metaphysic of Morals, 1998, p. 29)." Moral agents must always treat humanity, including oneself, as an end in themselves; therefore, they must not use human beings as a mere means to achieve other ends. This reasoning of Kant assures human dignity by respecting humanity as subjects instead of objects on the grounds that "rational nature exists as an end in itself (Kant, Groundwork for the Metaphysic of Morals)." However, why did Kant stress human dignity with respect to constituting rational nature? Being rational deserves respect for Kant because of the capacity (potential or actual) inherent and inalienable in human beings to govern their actions by reason. As Kant indicates: "The capacity to set oneself an end is what characterises humanity (Kant, Groundwork for the Metaphysic of Morals, 1998)." So, the capacity to be free is central to the dignity of human beings. It is rooted in the practical reason independent of internal causes such as personal inclinations and external causes regarding the will of others or natural law. Therefore, when setting a rule, a person must not interfere with the freedom of himself and other agents.

The humanity formula is rooted in human dignity, respectively, in the human capacity to be free; however, it should be noted that freedom for Kant is beyond the probability. Human beings are potentially free to act by self-legislated, i.e., reason-guided moral rules, but not all rational beings are free in the actual sense of Kantian freedom. Although an action can only be performed in the sensible world or appearances in Kant's terminology, Kant locates morality and freedom on the grounds of the intelligible world. Human beings belong to the intelligible world as rational beings and to the world of appearances as sensible beings. However, morality, as Kant argues, "exists in the sensible world but without infringing on its laws (Kant, Critique of Pure Reason, 5:43)." The capacity to act morally implies freedom in the sense of being free from the determination of the will through natural law external to moral laws or the law of reason. Nevertheless, the determination of the will independent from the causality of nature is only a negative freedom for Kant. This is because the will's freedom from the determining force of the causal law of nature does not imply being free in the sense that the will is determined by practical reason that makes it lawgiving on its own. "Although freedom is not a property

of the will according to laws of nature, it does not follow that freedom is lawless! (Kant, Groundwork for the Metaphysic of Morals, 1998, p. 41)." Therefore, "lawless" independence from the casual determination of natural law constitutes a negative side of freedom. For this reason, independence of will from alien causes other than itself that force to determine it is not a sufficient ground for the will to be free per se. However, the positive concept of freedom can be achieved if the will is not only determined independently of natural law but is also determined by moral laws. Therefore, freedom for Kant is the determination of the will by laws legislated through practical reason. So, Kant mentions freedom of will only if it is determined by a law that is internal to it (Demenchonok, 2019).

Although reason is a prerequisite for a moral law, morality for Kant presupposes freedom, which can only be experienced by moral action. Then morality, as Kant puts it: "proceeds from the concept of freedom (Kant, Critique of Pure Reason, 5:42)." It means that a moral action is compatible with the law of reason, the categorical imperative, and the concept of positive freedom. Therefore, an action would be morally judged only if performed freely because one can only be responsible for an act if he has full control over it. In this view, to act morally is to exercise freedom, and the only way to fully exercise freedom is to act morally (Rohlf, 2020).

"Freedom and unconditional practical law reciprocally imply each other (Kant, Critique of Practical Reason)."

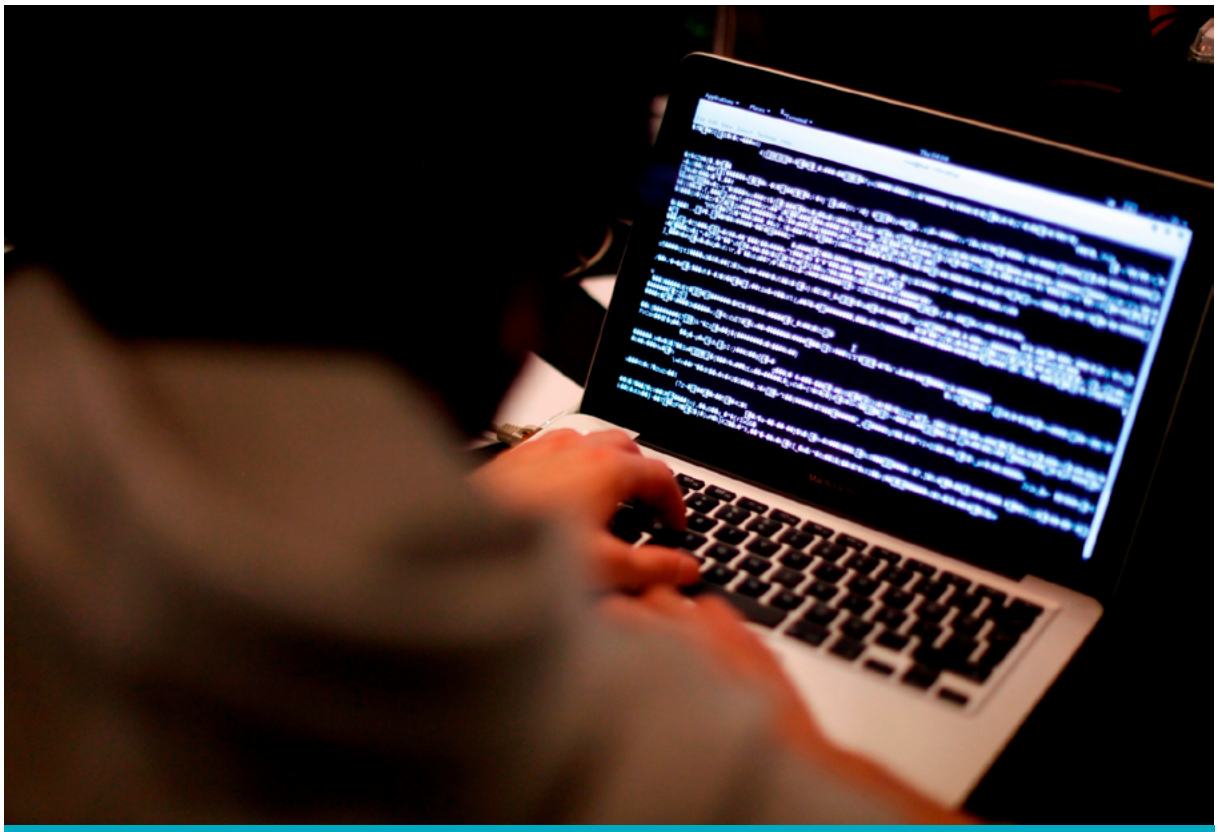
Kant's moral philosophy is based on the concept of autonomy that arises from positive freedom. In the third formulation of the categorical imperative, Kant states the last principle of moral action: "Act according to the maxims of a universally legislating member of a merely possible kingdom of ends (Kant, Groundwork for the Metaphysic of Morals, 4:439)." As a lawgiver to one's own self, human beings can determine their own ends as well as others by establishing universally applicable rules under the scope of reason-guided rule. By possessing the inherent capacity to move according to reason's demands, human beings are members of the kingdom of ends as lawmakers. By this so-called formulation of autonomy, Kant expresses that moral law is not given externally to an individual. However, each rational being sets up his own rule in conformity with universally applicable criteria to obey. Therefore, a person is subject to laws given only by himself in the moral realm. An autonomous moral agent, in this respect, is only bound to his own duties. Autonomy for Kant is, therefore, "the ground of the dignity of human nature and of every rational nature (Kant, Groundwork for the Metaphysic of Morals, 4:436)."

Kant's practical philosophy stresses how the world should be by establishing moral law to guide human actions. The subjection of human beings to morality, hence their moral obligations, is due to their intelligent nature capable of reasoning, independent of causes affecting their will from the outside. This special character of humanity gives them the capacity for freedom, autonomy, and, therefore, inalienable dignity and unconditional respect. Kant demonstrates that practical reason has the power of being the authority of the action itself as not being subjected to deterministic laws of nature. Unlike natural causality, the causality of reason, that is, the ground of freedom, does not affect the will externally. Hence, morality arises independently of experience, but is performed in the empirical world. Freedom in the sense of autonomy, therefore, means the "independence of the will from anything other than the moral law alone (Kant, Critique of Practical Reason, 5:94)." Therefore, according to Kant, the autonomous agent must be sovereign over himself which is only possible by being subject to only his own will. In this respect, human beings are morally bound to self-given duties; therefore, this obligation does not harm their autonomy but rather promotes it. When AI is considered through Kant's core principles of morality, how are these moral principles applied to artificial Intelligence?

Is Kant's moral agent applicable to AI?

When we imagine Artificial Intelligence as a Kantian Autonomous Moral Agent (AMA), AI's autonomy would no longer threaten humanity because it would govern its output in a way that respects the freedom of humanity and human dignity as Kant's conception of autonomy requires. Moreover, it would produce an output in accordance with an algorithm that is also intended to be a universal algorithm that ends with such an output. However, when Kantian terminology is considered with AI, concepts such as will, autonomy, and morality become complicated. This confusion leads to intricate problems considering the possibility of a Kantian AMA. One is then left to question whether it is possible to develop AI as a Kantian AMA while simultaneously being compatible with Kant's ethical theory. In order to unfold incomprehensibility, the core elements of Kant's moral theory will be examined in relation to the autonomy and morality of AI developed to be a candidate for Kant's moral agent.

Following on from the first formula, the categorical imperative provides a test for the maxims of the agent. Suppose a maxim (plan of action) that is potentially dutied to provide harmony within the principle of universal acceptability. For instance, it passes the test and, therefore, becomes a duty that an agent must follow. Kant uses the term "will" in the formulation that a will should become a universal law. So Kant asks an agent to make rules that are acceptable to others and the agent himself as well. However, an agent should "will" his maxim as a universal law. The will also refers to the fact that we as human beings should not act in a way that we do not "will" to be subjected to. This formulation is both in concord and in contrast with the mechanism of AI from distinct angles. Since maxims are "subjective volitions" in the first place and AI could make subjective plans whereby AI is considered a threat, the rest is if AI could be computed with the formulation that contributes to AI to test its subjective plans of action, and if they could turn into objective, universal laws.



Open AI's CEO Sam Altman testifies at an oversight hearing by the Senate Judiciary's Subcommittee on Privacy, Technology, and the Law to examine AI, focusing on rules for artificial intelligence. The hearing occurred in Washington, DC, on May 16th, 2023. (Nathan Posner - Anadolu Agency)

Brian Tomasik adapts the first formula into a system of AI as follows: "Choose the output to your cognitive algorithm whereby you can at the same time will that it should become the universal output of all instances of the designed cognitive algorithm (Tomasik, 2015)." At first glance, this adaptation allows AI to meet the universally accepted criteria to choose to perform ethical actions. That is to say, it can decide whether its output in question will be safe and logical if it is performed by all human beings and AI itself, via adapting the formula into its system. Even if it could determine itself to produce a universal output, could the decision-making process of AI be called ethical reasoning, let alone reasoning in the Kantian sense? Would expecting AI to consider ethical reasoning lead to a comparable approach between AI and human intelligence? Nevertheless, AI should include itself in considering universal acceptability. However, the contradiction occurs in the sense that it is also not accountable for a universal rule that it is generated. The procedure that AI derives duties from its maxims leads us to the point if AI can have the intention to make its maxim universal law. The will here represents Kant's emphasis on the universally applicable condition, which can only be tested if an agent also will so.

Furthermore, to will is to be also subjected and influenced by others. In fact, in this way, one can check the consistency of his maxim if he has the will to be treated this way. Fulfilling necessary conditions for ethical action requires decision-making but will is also a part of the decision-making process. If there were a will of AI, it could have its origins from its purposes imposed by its algorithms. In this sense, AI can only do so by the causal nexus that determines it via injected purposes. From this point, the self-legislating process requires "will", a moral intuition peculiar to human beings. Therefore, the decision-making process differs from AI to human beings concerning the principles of causation that govern the artificial will.

The complexities deepen with AI in meeting the necessary principles of Kant's morality when the motivation behind performing an action or output is observed. Kant considers an action moral only if it is done for the sake of duty. Therefore, Kant's moral agent's acts are motivated by the duty for following the moral rule. The fundamental reason behind the moral action comes from the rule's imperativeness, not an inclination, although both reasons end up with the same action. "An external event, given in space and time, does not by itself give us any access to the internal motivation of the agent," according to Kant (Kant, Critique of Practical Reason, 5:57). In this respect, a moral judgment usually necessitates inner accounting, whether made out of duty or not. Is it possible for AI to feel obligated to commit itself to producing an output for the sake of morality? It is another problem that Kantian moral artificial agent faces.

A fundamental point to consider if Kantian AMA violates Kant's moral theory is understanding the autonomy that AI could achieve. Given that Kant's moral agent is necessarily autonomous, morality implies self-governing that directly refers to freedom. The emphasis of Kant's autonomy for human beings is their capacity to act from practical reason without the interference of external, worldly obstacles to their will. It could be seen as freedom from something which makes freedom meaningful. If there were to be nothing for human beings' will to be affected, then what would they free themselves from? In this regard, freedom, then, is something to be achieved despite something. In Kantian terminology, this "something" is the sensible world governed by natural law. Since artificial Intelligence does not experience the sensible world, its decision-making process is not exposed to the determination of natural causality. Reason's challenge in determining moral law independent of external causalities is that there are already predetermined natural causes that force to govern the will of human beings who are part of the sensible and intelligent world. Hence, there is a struggle between the two worlds in which human beings take part. If the will is unaffected by the sensible world, it is already in the comprehensive realm without the empirical violation. Could a moral consideration exist for a being already free from all constraints of natural law? Is it already moral since its will cannot be affected from the outside, or is it not accountable for a moral query at all? It is essential to inquire what makes sense of freedom or autonomy for an abstract being if it is not subjected to natural causality.

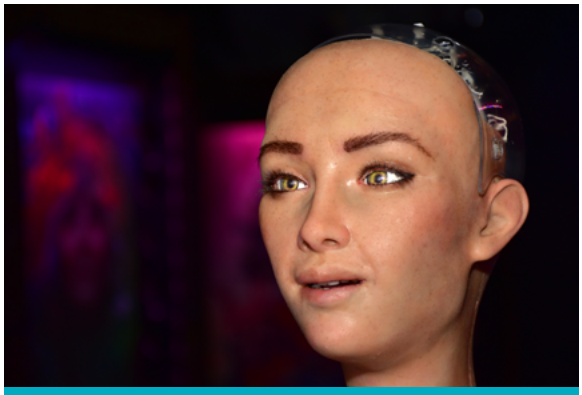
However, although AI is apart from worldly experience, its so-called Intelligence is constituted by a mass of worldly affected human inputs. From this perspective, AI possesses data acquired from the sensible world that is inexperienced by AI itself. These unsensed (by AI) imputed data of AI ground in the Intelligence of AI that affects its decision making. Therefore, although AI is apart from natural causation, its Intelligence does not meet Kant's realm of intelligence, where morality and autonomy arise independent of experience. This is because AI's decision-making is not subjected to natural law and is not free from its causal impact since AI is indirectly exposed to the sensible world through the unfree and inautonomous human inputs solely affected by the natural causality. Therefore, AI's Intelligence is apart from what Kant calls an intelligible world that is the foundation of morality and autonomy. Another point is that with its inputs or database, it is what AI is and nothing more. For this reason, unless AI has a will capable

of not being determined by these internal sensible (external regarding humans) data, i.e., obstacles to morality in its decision-making process, it cannot be free and autonomous in the Kantian sense.

A possible Kantian AMA, then, should "will" independently of its inputs, that are both external to itself and to its human providers to be called autonomous. What could be the will of AI? Predetermined laws of nature could drive a will, and AI can already follow the causal nexus of its own algorithm. Can this capability be called a will? Nevertheless, subjecting to the determination of other laws external to its will, other than the self-legislated rule, cannot be a foundation of morality or autonomy for Kant. If AI were to be a moral agent (which implies autonomy) of Kant, its will must do more than be driven by the causal chain of its inputs in the training data. Once freeing its will (if there were to be) from its inner sensory causality, can AI be autonomous according to Kant's ethical theory?

From another perspective, Kantian AMA must be part of the possible kingdom of ends as an end in itself, which is peculiar to humans with regard to their capacity for freedom, according to Kant. An agent, in this sense, could not be treated instrumentally but as an end in itself, and will treat others in the kingdom by respecting their dignity and autonomy. However, first and foremost, AI is created to serve instrumentally to meet human ends. AI's being autonomous in Kant's terminology leads to conceptual chaos since AI is treated as a means by human beings. According to Kant, this is the worst thing for an autonomous being to be the slave of others. If developing Kantian AMA would be possible without contradicting Kant's ethics, autonomous AI would be dignified, respected, and an equal member of the kingdom of ends as human beings. Developing such an agent would be secure for human beings. Autonomous AI, in the Kantian sense, would be even safer for humanity than human beings since human beings have the autonomy capacity, which implies that it is not actual in all humans, but computing such an agent would end up with autonomy that respects humanity unless it is broken down. However, regarding the possibility of corruption of such a powerful human-made agent, considering its withdrawal from morality and autonomy, it would be potentially dangerous. Therefore, although it contradicts Kant's ethics, intending to develop a Kantian AMA is substantially doubtful.

Philosopher Daniel Dennett remarks: "Imagine that you are whisked off somewhere, and you had installed an on-off button on you so that they could turn you off, I mean kill you at any moment by just putting the button. As an intelligent agent, what would be your first order of business? Trying to figure out how to control that on-off button so nobody else could turn you off (Dennett, 2021)."



Sophia the Robot, who received citizenship from Saudi Arabia and made history as the first robot with an identity card, is introduced in Antalya, Turkiye on July 08, 2023. Sophia participated in various programs and answered the questions of the press members at Digiverse Digital Exhibition Center. (Fatih Hepokur - Anadolu Agency)

As it is seen, Kantian AMA is incoherent in itself; therefore, autonomous AI could not be considered as being both moral and Kantian. The well-known term autonomy, which is discussed around AI, is not equivalent in meaning to Kant's understanding of autonomy. A prevalence of autonomy regarding AI "is disconnected from the metaphysical notion of humanity (Nowak, 2017)." Here, the meaning of autonomy is the ability to choose its purposes through a casual chain in itself. This usage of autonomy is not Kantian and has no common ground with Kant's concept of the will. Following this, the general usage of autonomy for AI implies the capacity to choose what is given to it rather than morality. If AI is not able to drive itself to moral actions, then its human providers should give it universally applicable rules to follow. In this top-down method, not the AI itself should raise moral and autonomous agents, but its human designer could be a Kantian moral agent to further his morality into AI. In this way, AI does not have to be autonomous (in the Kantian sense) but would act morally by autonomous engineer's self-legislated law. Thus, for developing AI as a tool to perform the universally applicable moral law, Kant's ethical theory would guide AI developers to obviate AI's possible damageable consequences to human beings. These can be removed by implementing not an ethical autonomous ground to AI but projecting an already present ethical ground of human agent's moral rule to its algorithms.

Conclusion

Within the comprehensive debates surrounding ethical AI, the notion of the so-called autonomy of AI has been the focal point of this research. Attributing autonomy within the capacity to render decisions on its own by AI has brought about ethical concerns stemming from the idea that autonomy, in this sense, does not assure moral implications. It has been argued that Artificial Intelligence, originally designed to serve humanity, could evolve into a force that contests human relationships within the scope of autonomy. In this regard, allowing the progress of autonomous AI is a self-addressed threat to man. Although autonomous AI can make choices without programmed instruction, it is incapable of making ethical choices as it does today. Given the premise of so-called autonomy, it becomes clear that AI requires ethical guidance.

The autonomy associated with AI in conventional discourse lies in the ability of AI to make decisions on its own, without following certain predictable principles. However, according to Kant, autonomy is not merely an independent decision-making capacity; rather, it is the faculty of exercising self-determination in accordance with moral principles. Moreover, in Kant's ethical theory, autonomy and morality appear as aspects that complement and reinforce each other, and at the same time, they cannot be found in the person independently of one another. In this respect, this paper investigates whether the potential threats of AI can be mitigated by adapting Kant's moral theory to AI, known as Kantian Artificial Moral Agent (AMA).

As a result of the research conducted within the framework of adapting Kant's ethical theory to Artificial Intelligence, it was concluded that it is impossible to develop such an agent since the idea of an AMA contradicts Kant's fundamental moral principles. Therefore, Kantian AMA violates Kant's ethics regarding categorical imperatives, the responsibility of morality, and concepts such as will and autonomy. However, it is not the only way to adopt Kant's ethics into AI systems while providing an ethical autonomous ground to AI. Instead of focusing on designing morality arising above the autonomy of AI, another way is ensuring AI is a means for human beings by adapting Kant's theory with developers that determine moral principles autonomously into AI.

Within the scope of this paper, Kant's ethics is examined among other ethical approaches. However, it does not mean to force upon Kant's morality to be the ground of the ethical decision-making process in AI. Since the very beginning, philosophers have engaged in distinct ethical thoughts with diverse practical guidance for human beings. However, these approaches have yet to build a consensus, even for humans, reckon without AI. Kant's ethical theory is one of others that attempts to figure out the best ethical grounds for humanity with practical guidance to provide security and peace in the world. Therefore, while regarding AI as autonomous, not only Kant but other theories should be investigated.

The focus here should not be on designing AI as both a moral and autonomous entity but rather on ensuring that AI remains a tool that needs to be guided so that AI producers can provide a system that will produce predictable results within ethical codes. Therefore, as mentioned earlier, the focal point of ethical query in AI research should be directed to engineers who develop artificial systems. With this, an engineer equipped with morality and ethical concerns can also foresee ethically mal results of the system rather than entrusting AI with morality autonomously.

Morality and autonomy necessarily imply responsibility directed to AI companies, engineers, or AI itself within the debates and research of AI. Regarding Moral Artificially Intelligent systems, responsibility query first occurs in human beings who lead away from the development of such a system. This paper has examined whether AI could be a responsible agent of its outputs from a Kantian perspective. However, the responsibility of humans that take charge in the development process of AI is essential to be questioned.

Bibliography

Gerrit De Vynck and Will Oremus. (2023, March 30). *As AI booms, tech firms are laying off their ethicists*. The Washington Post: https://www.washingtonpost.com/technology/2023/03/30/tech-companies-cut-ai-ethics/

Amitai Etzioni and Oren Etzioni. (Summer 2017). Should Artificial Intelligence Be Regulated? *Issues in Science and Technology*, 32-36.

Amitia Etzioni and Oren Etzioni. (2017). Incorporating Ethics into Artificial Intelligence.

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

Demenchonok, E. (2019). Learning from Kant: On Freedom. *Revista Portuguesa de Filosofia*, 191-230.

Dennett, D. (2021). The risks of creating artificially intelligent agents. https://www.youtube.com/watch?v=EqOiSY-16QXM

GPT-4 Technical Report. (2023, March 15). Arxiv: https://arxiv.org/abs/2303.08774

Hauptman, M. (2023, June 1). *Air Force said AI drone killed its human operator in a simulation*. Task and Puspose: https://taskandpurpose.com/news/air-force-artificial-intelligence-drone/

Kant, I. (1998). *Groundwork for the Metaphysic of Morals*. (M. Gregor) Cambridge University Press.

Kant, I. (2019). *Critique of Practical Reason*. (L. W. Beck, trans.) Pearson.

Kant, I. (1998). *Critique of Pure Reason*. (P. Guyer, trans.) Cambridge: Cambridge University Press.

Nowak, E. (2017). Can human and artificial agents share an autonomy, categorical imperative-based ethics and "moral" selfhood? *Filozofia Publiczna i Edukacja Demokratyczna*, 169-208.

OpenAI. (2023, February 16). *https://openai.com/blog/how-should-ai-systems-behave*. OpenAI: https://openai.com/blog/how-should-ai-systems-behave

Paul, K. (2023, April). *Letter signed by Elon Musk demanding AI research pause sparks controversy*. The Guardian: https://www.theguardian.com/technology/2023/mar/31/ai-research-pause-elon-musk-chatgpt

Pause Giant AI Experiments: An Open Letter. (2023, March). Future of Life Institute: https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Rohlf, M. (2020, July 28). *Immanuel Kant*. Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/kant/#MorFre

Russel, S. J. (2023). *Leading AI expert Stuart J Russell explains why putting guardrails in place is imperative right now.* World Economic Forum: https://www.linkedin.com/posts/world-economic-forum_leading-ai-expert-stuart-j-russell-explains-activity-7066805142297141250-kHLA/

Tomasik, B. (2015, March 19). *Interpreting the Categorical Imperative*. Brian Tomasik: https://briantomasik.com/interpreting-the-categorical-imperative/

TRT WORLD
research centre

TRTWORLD
research
centre