AI and Ethics (2024) 4:325–331
https://doi.org/10.1007/s43681-023-00269-6

ORIGINAL RESEARCH

Can artificial intelligence be a Kantian moral agent? On moral autonomy of AI system

Arunima Chakraborty[1] · **Nisigandha Bhuyan**[2]

Received: 3 September 2022 / Accepted: 18 February 2023 / Published online: 7 March 2023

Abstract

Machine Ethics emphasises the importance of collaboration between engineers, philosophers and psychologists to develop artificial intelligence-endowed systems and other 'smart' machines as artificial moral agents (AMA). They point out that there are top-down and bottom-up approaches for programming values into artificial autonomous systems. A number of thinkers argue that formalisation of the Kantian categorical imperatives is feasible, and hence, it is possible for smart machines to become Kantian moral agents, through the top-down approach of programming the Kantian categorical imperatives as algorithms into the AI systems. This paper examines some of the arguments put forth by the defendants of the possibility of Kantian AMAs such as Powers to point out that what these thinkers ignore is that in the Kantian schema, a moral agent is a rational being who is capable of 'universalising' as the law, the subjective maxims of her actions. Can the AMA be rational in this Kantian sense? The paper argues that though Kantian deontology may be attractive a theory for designing AMAs, the artificial agents cannot be Kantian moral agents in the real sense of the term.

Keywords Moral machine · AMA · Reason · Choice · Autonomy · Inclination · Dialectic · Kant · Synthetic a priori

In the literature on ethics of artificial intelligence (AI), one area which has deservedly received considerable attention is the question whether moral values can be programmed into AI-equipped machines, and if yes, then can such machines be considered to have moral responsibility. As robots, 'smart' machines and bots increasingly foray into various fields where direct human control over them gradually decreases, philosophers and scientists speculate over the prospects of the emergence of moral machines [24]. The discipline which deals with the possibility, nature and characteristics of the Artificial Moral Agents (AMA) of the future is known as Machine Ethics. Speculations are also rife as to whether AI-driven machines can be held morally responsible for their actions [9]. Artificial intelligence is the product of both advanced and constantly evolving hardware and software technologies [7]. Access to big data makes a crucial contribution to the growth of the autonomy of the AI systems (Bostrom 2016, UNI 2019; [20].

Artificial intelligence now powers driverless cars which can take independent decisions and these decisions have moral implications [10]. Artificial intelligence is also used in medical research, targeted advertisements, and care for the elderly and even lethal weapon systems [22], Scudellari 2018; [21]. Machine-learning algorithms can speedily process vast quantities of data to recommend efficient courses of actions in several fields,they can learn on their own from a vast set of data, and take decisions in various situations without human supervision or control. The year 2023 began with much buzz and curiosity around the capabilities of the AI-powered Chabot, ChatGPT; it has passed some of the world's most competitive university entrance exams giving rise simultaneously to euphoria over the increasing computational skills of AI as well apprehensions about the impending redundancy of jobs due to automation as well as about dangers of plagiarism [12]. Thinkers such as Christian Fuchs [11] argue that it cannot be yet asserted with any certainty if technology, particularly AI-based digital technology will empower the masses, or be a tool of their oppression. The question worth considering at this juncture is whether AI

✉ Arunima Chakraborty
shosha.aru@gmail.com

Nisigandha Bhuyan
nisigandha@iimcal.in

[1] Centre for Studies in Social Sciences, Calcutta, India

[2] Indian Institute of Management, Calcutta, India

systems, given the increasingly complex tasks which they can perform, emerge as autonomous moral agents?

The document titled 'Asilomar AI Principles' of 2017 [1] which was signed by the likes of Stephen Hawking, Elon Musk and Jaan Taalin declares that autonomous AI systems should be designed in such a way that they are 'aligned with' human values, and they do not contravene 'ideals of human dignity, rights, freedom and cultural diversity'. Likewise, if one goes back to the Three Laws of Robotics formulated by Isaac Asimov, the emphasis was on the fact that robots and artificial intelligence systems, no matter how autonomous, cannot be allowed to harm humanity [16]. The concept of 'adjustable autonomy' argues for a limited autonomy of artificial agents (AAs) in such a way that humans are able to able to control and regulate AAs in complex situations [8]. The importance of this framework is accentuated by the fact that it is not always clear as to what causes an AI system-based machine to generate a particular result, or to carry out a particular action in a given situation. In the words of Mindell [15], the process of algorithmic decision-making of the AI system-based machines is marked by a certain 'opaqueness' because of which it is not always clear whether a particular action or recommendation by the machine is because it is 'malfunctioning or is that part of its decision-making tree' (p. 191). But while some models like the adjustable autonomy theory favour the strictly subordinate role of AI systems in the human realm, there are possibilities of emergence of AAs as autonomous moral agents as is exemplified by 'cognitive computers' such as the humanoids iCub and ASIMO which are endowed cognitive skills [8]. The question which this paper aims to explore is whether AAs can emerge as Kantian AMAs, and thus, possess moral autonomy in the Kantian sense of the term? Immanuel Kant, who propounded the theory of 'transcendental idealism' in the *Critique of Pure Reason* wherein he states that 'synthetic a priori' knowledge is possible because all empirically gained and thereby, synthetic knowledge is subsumed by mental categories of understanding which are a priori, argues in the *Groundwork of the Metaphysics of Morals* [13] that autonomy of the moral agent consists of her capacity to exercise her 'will' which in turn, is the capability to 'act in accordance with the representation of laws' (p. 24). According to Kant, the highest moral law is the categorical imperative which is 'an a priori synthetic practical proposition' (p. 30). Further, since an imperative is the formulae of the command for action, in the Kantian schema, there are two kinds of imperatives: the hypothetical and the categorical imperatives. While the hypothetical imperative is a command of action wherein there is need to carry out the action as a mean to attain some other end, the categorical imperative or the highest moral law is the command of an action which is 'objectively necessity of itself' (p.25). An action which is necessary by itself, is an action wherein a personal maxim can be universalised as a law. Kant writes about the categorical imperative:

> "There is, therefore, only a single categorical imperative and it is this: *act only in accordance with that maxim through which you can at the same time will that it become a universal law*" (p. 31, emphasis original).

Now since, the Kantian deontological theory is one of the ethical theories which are often advocated for the designing of AAs as artificial moral agents, it is worth examining whether AAs can emerge as Kantian moral agents?

1 Kinds of artificial moral agents

In order to answer whether artificial intelligence or machines can emerge as artificial moral agent, it is first crucial to ask what qualifies as moral agency of machines? While some define it through a negative as the machine or system which does not carry out immoral actions [8], others have enunciated a functionalist account of moral agency of machines. A machine is a tool in so far as it is a means to a specific end; it is aimed at improving the speed, efficiency and efforts entailed in a certain human endeavour. In so far the machine successfully carries out the task assigned to it, it has served its purpose. According to Nowak [17] an action is a means of self-determination and not only living organisms, machines to can carry out actions to produce certain ends,in this limited sense, machines are 'agents' just as living beings are. James Moor [16] elaborates the kinds of moral agency which can be attributed to machines: first, there are ethical impact agents which is the term for that category of machines which impact humans morally: positively, negatively, or in both ways. Almost any technology can qualify for the status of an ethical impact agent.

The second category of machines are the implicit ethical agents which are called so since 'ethical considerations are built into (that is, implicit in) their design'. A teller machine that correctly counts the money is an example of an implicit ethical agent since it aids its user in counting cash correctly, and has thereby serves the end for which it has been designed. AI in so far as it takes decisions and gives recommendations which are to assist humanity in its varied endeavours can also be cited as an example of implicit ethical agent. From an Aristotelian perspective, an efficient and well-functioning machine is an ethical agent in so far as it performs its 'function' well. In so far as a machine carries out, albeit mechanically, the function to serve which it has been conceived, designed and built, and if that function enhances human welfare—in no matter how small or immense a way—it serves its ethical function and hence, is an implicit ethical agent. It may be argued that Moor's

concept of implicit ethical agent—as the second category of ethical AMAs—bears some relation to the Aristotelian concept of function.

Similarly, Amitai Etzioni and Oran Etzioni describe as 'AI partner' the following:

> " The other kind of AI merely seeks to provide smart assistance to human actors...This kind of AI only requires that machines be better at rendering decisions in some matters than humans and that they do so effectively within parameters set by humans or under their close supervision" ([10], p. 9).

This conception of the AI is also in tandem with the principles of Asimov and those termed as the Asilomar AI principles which envision the AI endowed machines as tools of human endeavours. There is, however, another kind of AI which the Etzioni and Etzioni [10] term as 'AI mind' which 'seeks to reason and form cognitive decisions'. Closely resembling this concept of the AI mind is Moor's third category of AMAs: explicit ethical agents, which are those machines which can independently take decisions taking ethical considerations into account, since ethical rules are encoded in them. Moor's fourth category of AMAs are the 'full ethical agents' whose capacity for ethical reasoning and action are as developed as those of humans.

Nick Bostrom observes about the gradual transformation of machine/artificial intelligence from a mechanistic tool to an 'AI mind' or 'explicit ethical agent' as follows:

> "During these early days, researchers built systems designed to refute claims of the form 'No machine could ever do X!' Such sceptical claims were common at the time. To counter them, the AI researchers created small systems that achieved X in a 'microworld' (a well-defined, limited domain that enabled a pared-down version of the performance to be demonstrated), thus providing a proof of concept and showing that X could, in principle, be done by machine. One such early system, the Logic Theorist, was able to prove most of the theorems in the second chapter of Whitehead and Russell's Principia Mathematica, and even came up with one proof that was much more elegant than the original, thereby debunking the notion that machines could 'only think numerically' and showing that machines were also able to do deduction and to invent logical proofs" ([6], pp. 5-6).

The exercise of cognitive faculties is also entailed in taking moral decisions; Moor describes as 'explicit ethical agents' those machines which can comprehend moral aspects of various processes and situations, and allows itself to be guided by those aspects. As Powers [18] points out, in the Kantian framework, in order to be considered as 'ethical in themselves', machine intelligence must demonstrate a 'simulacrum of ethical deliberation'. Close to the concepts of 'AI mind' (of the Etzionis), and the 'explicit moral agent' (of Moor) is the idea of the moral Turing Test. Allen et. al. [4] discuss a variation of the Turing test called the Moral Turing Test (MTT); while in the standard Turing test, a person is asked to distinguish between machine and human based solely on interacting with both via printed language, in the MTT, likewise, a person is asked to differentiate between a machine and a human based on conversations with both about morality. Allen et. al. [4] write, "If human 'interrogators' cannot identify the machine at above chance accuracy, then the machine is, on this criterion, a moral agent" (p. 81). Thus, in the Moral Turing Test, it is not sufficient for the machine to be merely functional in order for it to qualify as moral agent; it must also be able to demonstrate its capacity for moral reasoning. Does the reasoning which AAs must demonstrate in different AMA models such as the 'AI mind', 'explicit moral agent' or MTT qualify them as moral, rational agents in the Kantian sense?

2 Actions of moral worth, according to Kant

As AI increasingly takes over not only processing of vast and complex data sets, but also diurnal human activities such as care for the elderly, baby-sitting, driving and serving as waiters amongst others, the increasing human–machine interactions render it necessary thatartificial agents be equipped with certain just principles of decision-making. Principles which guide the algorithmic decision-making systems are expected to be aligned to human values. Wallach [25] observes that the only reason why AMAs are modelled after human values is because the process of human moral reasoning is all that there is access to. He, however, misses a crucial reason as to why AMAs should reflect human values: it is because artificial agents are to serve human needs. The ethical principles can either be programmed into an artificial intelligence system which is the top-down approach; they can also be acquired or learned by the AI system as it interacts more with humans in different situations, and this is the bottom-up approach [10, 25]. If the top-down approach is adopted, then the programmer or the designer has the responsibility of deciding what kind of ethical principles she should programme into the autonomous artificial intelligence-driven decision-making systems. There is a rising demand today for algorithmic transparency in the AI systems so that the process of how and why does a particular AI system takes a decision, can be comprehended [22]. This is necessary to detect and counter biases in the algorithmic decision making process which may be due to biases in the data set made available to the algorithm itself [5]. Wallach [25] and Boddington [5] argue that the top-down approach to designing AMAs

suit moral philosophy, since the philosophers can draw from numerous moral theories such as Virtue Ethics, Deontology and Consequentialism to decide which set of values should be programmed in the machine for it emerge as an AMA. While Consequentialism of which Utilitarianism is a variety, states that the moral worth of an action is to be gauged in terms of the consequences which follow from it,for Virtue Ethics of which Aristotle is a primary exponent, ethics is always imbibed through habituation to virtuous activities. Deontology of the Kantian variety eschews the judging of the moral worth of any action in terms of the effects it produces, and focuses instead on whether the action was out of respect for objective and universal laws.

In *Groundwork of the Metaphysics of Morals* [13], Immanuel Kant sets himself to the task of exploring the question how do we know that an action has moral worth? Kant writes:

> "The present groundwork is, however, nothing more than the search for and establishment of the supreme principle of morality, which constitutes by itself a business that in its purpose is complete and to be kept apart from every other moral investigation" ([13], p. 5).

Kant further states:

> "...an action from duty has its moral worth not in the purpose to be attained by it but in the maxim in accordance with which it is decided upon, and therefore does not depend upon the realisation of the object of the action but merely upon the principle of volition in accordance with which the action is done without any regard for any object of the faculty of desire" ([13], p. 13).

Having asserted that moral worth of actions does not depend upon the faculty of desire which might motivate an action, or upon the consequences which follow from it but upon the maxim in accordance with which it is enacted, Kant goes on to elaborate his moral theory according to which the categorical imperative as the highest moral principle. Kant states that an action can be deemed moral only if it is motivated not by inclination but from duty; he next describes duty as that which is 'the necessity of an action from respect for law'. By law, Kant means the objective principle of volition while maxim is the subjective principle. He defines respect as the 'immediate determination of the will by means of the law and consciousness of this' and as 'representation of a worth that infringes upon my self-love'. In other words, an action is to be considered as moral only if it is carried out not under the dictates of inclinations but from duty; duty consists of respect for the law; and the law is the objective principle of the will. When a maxim (which is the subjective principle of action) can be universalised, it becomes a law. The importance of universalization of a subjective maxim lies in the fact that when it is universalized, the objective worth of the maxim of the will is revealed depending upon whether it gives rise to contradictions or not. If the universalization of a maxim as the basis of volition gives rise to a logical or teleological or practical contradiction, then such a maxim obviously cannot be assigned the status of an objective practical law, and therefore, cannot be moral. Morality, therefore, has both a material and formal components. According to Kant, no rational being–including rational human beings–can pursue such a maxim as a law which when universalized, leads to contradictions.

A rational being therefore ensures that she elevates to the status of (universal) law only those (subjective) maxims which do not give rise to contradictions of logical, teleological or practical varieties. Or to put it differently, for Kant, a universal moral law must be free of contradictions. Thus, Kant defines as a rational being one who acts in accordance with her 'representation of the law' which alone is the 'determining ground of the will' and reason as that which is 'required for the derivation of actions from the law'. Now, if reason is the basis of both actions which are in accordance with the representation of the law, and of will, will may be defined as the 'capacity to choose only that which reason independently of inclination cognizes as practically necessary, that is, as good'. In the process of formulating an (objective, universalizable) law out of a subjective maxim, therefore, one exercises reason which in turn determines the will; and in the course of the same process, one also relinquishes one's inclinations which can and often does have influence on one's will. There is, according to Kant, a dialectic between the subjective inclinations which he defines as 'dependence of the faculty of desire upon feelings' which 'always indicates a need' and the objective law, which is arrived at out only by the rational being.

Kant explains the dialectic as follows:

> "Now reason issues its precepts unremittingly, without thereby promising anything to the inclinations, and so, as it were, with disregard and contempt for those claims, which are so impetuous and besides so apparently equitable (and refuse to be neutralised by any command). But from this, there arises a natural dialectic, that is, a propensity to rationalize against those strict laws of duty, and to cast doubt upon their validity, or at least upon their purity and strictness, and where possible, to make them better suited to our wishes and inclinations, that is, to corrupt them at their basis and to destroy all their dignity–something that even common practical reason cannot, in the end, call good" [13], p. 17-18).

In so far as the rational being can overcome the dialectic between her inclinations and the representation of the law, and choose the latter, she is also autonomous. For Kant, while autonomy comprises of the choice of the will to base its actions on the law, instead of inclination, heteronomy is that which the 'object, by the means of its relation to the will, gives the law to it' instead of the will giving itself the law (p. 47). In this case the object is the end which the will serves, and the law is merely a means to the attainment of end, and hence, a hypothetical imperative.

According to Kant, therefore, morality, reasoning as well as autonomy are intrinsically linked to the capacity of the will to act in accordance with objective and universal practical laws, into which subjective maxims have been transformed. Inclinations gives rise to maxims of actions, but only those maxims which can be universalised without running into logical, or teleological, or practical contradictions are elevated to the status of objective laws of actions by the rational being; when the subject exercises her will in accordance with the laws instead of her inclinations and other heteronomies, she also exercises her autonomy. And this is how, the moral worth of an action is determined in the Kantian framework of ethics. Kant writes:

> "Morality consists, then, in the reference of all actions to the lawgiving by which alone a kingdom of ends is possible. This lawgiving must, however, be found in every rational being himself and to be able to arise from his will, the principle of which is accordingly: to do no action on any other maxim than one such that it would be consistent with it to be a universal law, and hence to act only so that the will could regard itself as at the same time giving universal laws through its maxims" [13], 42).

In the Kantian framework, therefore, moral reasoning entails 1. The dialectic between inclinations and needs, on the one side, and the imperative of the law on the other; as well as 2. The capacity to transform subjective maxims of actions to objective, universally applicable laws of action. The absence inclinations and needs does not render the AAs as Kantian artificial moral agents because they also lack the capacity of practical reasoning whereby actions are motivated by the respect for the law, or the categorical imperative which entails the capacity for universalising of personal maxims.

3 Can there be a Kantian moral machine?

The preferred ability of the rational and autonomous being to act by a will that can 'give itself' universal laws, independent of the impact of subjective inclinations and desires, makes certain Kantians wonder if the AI can acquire the status of moral agents given the fact it can act on universal laws –which it either has been programmed to act on, or which it gives itself on the basis of self-learning–unaffected by pathologies of interests or inclinations [14, 17–19, 26]. Artificial Intelligence systems can be moral in the Kantian sense if they become less human in some ways [19]. According to Powers [18] machines can be ideal Kantian AMAs because in the Kantian moral philosophy, exercise of the Categorical Imperative does not entail use of emotions, or intuitions. He observes:

> "The procedure of deriving duties from the rule—if we are to believe Kant—requires no special moral or intellectual intuition that might be peculiar to human beings" [18].

For Powers [18] and Lindner and Bentzen [14], the two categorical imperatives of Kant can be formalised and can be algorithmically represented,for Powers [18], as his above observation indicates, the choice and exercise of the categorical imperatives require no emotions, and hence, can be programmed as algorithms into the machines. But is such an interpretation of the categorical imperative an accurate one? Allen et. al. [4] take a stance contrary to that of Powers when they point out that according to Kant, 'the categorical imperative is only an imperative for humans' (p. 253). A second objection can be articulated against the position of Powers which may be derived from the discussion in the previous section where it has been demonstrated that for Kant, moral reasoning entails a dialectic between pathological inclinations and subjective preferences for certain ends on the one hand, and will which respects the objective, moral law on the other, as well as the capacity to universalise as laws, the subjective maxims of actions.

In the first section of *Groundwork*, Kant in order to explain the ontological status of moral principles as synthetic a priori, states that ethics cannot be garnered from the phenomenal world, as there is hardly any human action in the range of experience which can be described as an instance of the categorical imperative. It is not within the ambit of the present paper to debate whether moral actions are indeed synthetic a priori, or not but what is relevant here is Kant's observation that man can formulate the objective law as the end-in-itself but is too dependent on his heteronomies to be able to act on the law.

Powers, in a different essay, observes that the lack of dialectic between heteronomies/inclinations and the objective law in the case of machine intelligence, makes it an ideal Kantian moral agent, in the top-down approach; he writes:

> "It is interesting to evaluate such a machine in the light of the possible 'erosion' of central human characteristics like intelligence, empathy and

> sociability. The more exactly our moral machine implements such a formally structured Kantian morality, the less it would behave like a human in some relevant respects. It would have no need for expression of regret, moral conflictedness or any act of conscience, since everything it did would fall neatly under the categories of moral maxims that we've programmed into it or which have been logically derived from those we've programmed. It would not suffer from the weakness of the will, because it would always be programmed to act according to its moral categories" [19].

Powers further feels that such a Kantian moral machine might even learn to access human actions, which plagued with the shortcomings of frailties, always fall short of elevating laws to the status of ends-in-themselves. Powers acknowledges that these machines do not possess free will. Moor too observes that while AI already is 'implicit ethical agent' and may in future, become 'explicit ethical agent', but it is a matter of sheer speculation if an AI system can become a 'full ethical agent'—Moor's fourth and final category of AMAs – which he defines as the ethical being which possesses 'central metaphysical features that we usually attribute to ethical agents like us – features such as consciousness, intentionality and free will'.

In other words, machine learning can become an implicit ethical agent in the 'function argument' framework which states that to be moral is to be able to perform one's functions; but can it become a Kantian ethical agent, as Powers argues? In response to the idea that AI systems cannot possess free will, and therefore, cannot be treated ethical beings, Powers states:

Kantians always supposed that formal moral reasoning and the special moral status of humans pulled in the same direction, so as to speak, because free will ('the uncaused causality of the will') was something that mere means–mere machines–could never have. Now it seems that the capacity to reason morally would be seen as independent of our autonomous will, and not a consequence of it [19].

In the Kantian framework, however, one cannot separate morality and actions of moral worth from the will; one cannot envision a Kantian moral machine without will. As stated earlier, for Kant the moral being is rational since he can transform his certain maxims into laws which do not give rise to contradictions, as well as autonomous since his will can choose to act in accordance with the law which reason has formulated as an end-in-itself, and reject the needs of inclinations. The rational being is autonomous because she can exercise her choice when acting as per reason, instead of giving in to heteronomies. A moral being is, therefore, one who is both rational (since she give herself a representation of the law) and autonomous (since her will can choose to act on reason, instead of inclinations); a moral being in the Kantian framework is one who has undergone the dialectic between inclinations and reason, between heteronomies and autonomy, between laws as hypothetical imperatives and as categorical imperatives, and has chosen the kingdom of ends. Tonkens [23] makes an argument similar to this paper when he points out that AMAs cannot be Kantian ethical agents because for Kant, the ethical agent is one who chooses the law over inclination. Tonkens states:

> 'According to Kant, it is only because humans can violate the moral law and succumb to the temptations of sensual satisfaction that they can truly be said to be moral agents. Duty signifies the (rational) ''strength needed to subdue the vice-breeding inclinations''. In other words, part of the force and achievement of acting dutifully stems from the fact that one could have acted otherwise (2009, 426).

The AI system, on the other hand, might come to possess practical reason through the bottom-up approach and use it as the basis of its moral decisions, but in so far as it cannot undergone the dialectic, it cannot possess will, and thereby, it cannot possess autonomy or the power of reasoning in the Kantian sense. Hence, AI cannot be granted the status of a moral agent, from the Kantian perspective. Finally, if AI cannot be an ethical agent, the AI system or the robot endowed with machine learning algorithms cannot in-itself be held morally responsible for its actions; the moral responsibility must always lie with the autonomous human beings who design and programme it.

Acknowledgements The authors would like to thank the Indian Institute of Management, Calcutta for funding the research project during the course of which this paper was written

Author contributions Both the authors have contributed equally to the writing of the paper.

Funding The funding was provided by the Indian Institute of Management, Calcutta.

Data availability No data has been used which is not available in public domain.

Declarations

Conflict of interest No financial or non-financial interests are at stake if the paper is selected for publication in AI& Ethics.

Ethical approval The authors state that no ethical guidelines of research were flouted in the course of the research and writing of this paper.

Informed consent The authors consent to the publication of the paper if it is selected in AI and Ethics.

References

1. *Asilomar AI Principles.* 2017 https://futureoflife.org/ai-principles/?cn-reloaded=1&cn-reloaded=1
2. Adams, Tim. 2016. "Artificial Intelligence: 'We're like children playing with a bomb.'"*The Guardian,* 12 June, 2016.
3. https://www.theguardian.com/technology/2016/jun/12/nick-bostrom-artificial-intelligence-machine
4. Allen, C., Varner, G., Zinser, J.: Prolegomena to any future moral agent. J. Exp. Theor. Artif. Intell. **12**, 251–261 (2000)
5. Boddington, P.: Towards a Code for Artificial Intelligence. Springer (2017)
6. Bostrom, N.: SuperIntelligence: Paths, Dangers, and Strategies. Oxford University Press (2014)
7. Bostrom, Nick. 1997. "How Long before Superintelligence." https://www.nickbostrom.com/superintelligence.html
8. Cervantes, J.-A., Lopez, S., Rodriguez, L.-F., Cervantes, S., Cervantes, F., Ramos, F.: Artificial moral agents: A survey of the current status. Sci. Eng. Ethics **26**, 501–532 (2020)
9. Dignum, Virginia. 2018 Ethics in Artificial Intelligence: Introduction to the Special Issue. *Ethics and Information Technology.* Doi: https://doi.org/10.1007/s10676-018-9450-z
10. Etzioni, A., Etzioni, O.: Incorporating ethics into artificial intelligence. J. Ethics **21**, 403–418 (2017)
11. Fuchs, C.: M. N. Roy and the frankfurt school: socialist humanism and *the critical analysis of communication, culture, technology,* fascism and nationalism. TripleC **17**(2), 249–286 (2019)
12. Hao, Karen. 2023. What is ChatGPT? What to know about the AI Chatbot That Will Power Microsoft Bing? *The Wall Street Journal*. https://www.wsj.com/articles/chatgpt-ai-chatbot-app-explained-11675865177?mod=e2tw
13. Kant, I.: Groundwork of the Metaphysics of Morals. Cambridge University Press (1998)
14. Lindner, Felix and Bentzen, Martin Mose. 2019. "A Formalization of Kant's Second Formulation of the Categorical Imperative." *Journal of Applied Logic*. https://arxiv.org/abs/1801.03160
15. Mindell, D.A.: Our Robots, Ourselves: Robotics and the Myths of Autonomy. Viking, New York (2015)
16. Moor, J., H.: Four Kinds of Ethical Robots. Philosophy Now **72**, 12–14 (2009)
17. Nowak, E.: Can human and artificial agents share an autonomy, categorical imperative-based ethics and "moral" selfhood? Filozofia Publiczna i Edukacja Demokratyczna **6**, 169–208 (2017)
18. Powers, T., M.: Prospects for a Kantian machine. IEEE Intell. Syst. **21**, 46–51 (2006)
19. Powers, T., M.: Machines and moral reasoning. Philosophy Now **72**, 15–16 (2009)
20. Rosenfield, Monica. 2016. The next step for artificial intelligence is machines that get smarter on their own. *The Institute.*
21. Russell, S.: Take a Stand on AI weapons, in Robotics: Ethics of Artificial Intelligence. Nature **521**, 7553 (2015)
22. Spice, Byron. 2016. "Carnegie Mellon Transparency Report makes AI decision-making accountable." Carnegie Mellon Computer University School of Computer Science. http://www.cs.cmu.edu/news/carnegie-mellon-transparency-reports-make-ai-decision-making-accountable.
23. Tonkens, R.: A Challenge for machine ethics. Mind. Mach. **19**(3), 421–438 (2009)
24. Wallach, Wendell. 2009. "The Challenge of Moral Machines." *Philosophy Now* 72.
25. Wallach, W.: Robot minds and human ethics: the need for a comprehensive model of moral decision making. Ethics Inf. Technol. **12**, 243–250 (2010)
26. White, Jeffrey: Autonomous reboot: Kant, the categorical imperative, and contemporary challenges for machine ethicists". AI and Society (2021). https://doi.org/10.1007/s00146-020-01142-4