Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-020-00614-8 OPEN

Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement

Silviya Serafimova[1]✉

Moral implications of the decision-making process based on algorithms require special attention within the field of machine ethics. Specifically, research focuses on clarifying why even if one assumes the existence of well-working ethical intelligent agents in epistemic terms, it does not necessarily mean that they meet the requirements of autonomous moral agents, such as human beings. For the purposes of exemplifying some of the difficulties in arguing for implicit and explicit ethical agents in Moor's sense, three first-order normative theories in the field of machine ethics are put to test. Those are Powers' prospect for a Kantian machine, Anderson and Anderson's reinterpretation of act utilitarianism and Howard and Muntean's prospect for a moral machine based on a virtue ethical approach. By comparing and contrasting the three first-order normative theories, and by clarifying the gist of the differences between the processes of calculation and moral estimation, the possibility for building what—one might call strong "moral" AI scenarios—is questioned. The possibility of weak "moral" AI scenarios is likewise discussed critically.

[1] Bulgarian Academy of Sciences, Sofia, Bulgaria. ✉email: silvija_serafimova@yahoo.com

Introduction

Key concepts and clarifications. It is not an accident that decision-making based upon algorithms remains “a standout question in machine ethics” (Mittelstadt et al., 2016, p. 11). It is due to the fact that such a decision-making process necessitates the differentiation of the requirements for moral agency when applied to humans and intelligent agents (IAs). In this context, special attention should be paid to the issue of whether or not algorithmic decision-making can achieve a level of moral autonomy similar to that of human decision-making.[1]

Before trying to clarify the role of some key concepts, it is important to roughly delineate the boundaries of machine ethics as such. According to Anderson and Anderson, unlike computer ethics, which has traditionally discussed ethical issues “surrounding humans’ use of machines”, machine ethics is focused upon “ensuring that the behavior of machines towards human users, and perhaps other machines as well, is ethically acceptable” (Anderson and Anderson, 2007, p. 15). The ultimate goal of machine ethics is “to create a machine that *itself* follows an ideal ethical principle or set of principles”. Those principles should guide it in making decisions about possible courses of action it could take (Ibid.). As such a principle, one can point out that of “how we might program computers that will *themselves* refrain from evil and perhaps promote good” (Powers, 2006, p. 46).

In turn, it is methodologically important one to raise MacIntyre’s well-known questions, “Whose justice?” and “Which rationality?” in a new voice. It means to modify them into “Whose morality?” and “Which rationality?”, while discussing the role of moral reasoning in building moral machines. The two questions provoke some additional questions such as the following: Does it mean that the engineers should create a machine that is sufficiently intelligent and then try to elaborate upon its intelligence in respect to morally significant matters? Or should they try to develop a certain type of intelligence that goes hand in hand with a given type of morality?

Anderson and Anderson make the point that “having all the information and facility in the world won’t, by itself, generate ethical behavior in a machine” (Anderson and Anderson, 2007, p. 15), nor would experts find easy solutions. This is due to the fact that “ethics has not been completely codified” (Ibid.). That is why the challenge consists in specifying how to create “an ethical intelligent agent” (Ibid.).

The plurality of combined approaches towards specifying what an ethical intelligent agent should look like requires some limitations to be provided. In the current article, two types of ethical intelligent agents are examined, namely, the so-called by Moor implicit and explicit ethical agents (Moor, 2006, pp. 19–20). According to Moor, “implicit ethical agent” is a machine that “has been programmed to behave ethically” by following some ethical principles as embodied by its designer. Correspondingly, Moor defines “explicit ethical agent” as a machine that is “able to calculate the best action in ethical dilemmas using ethical principles” by itself (Anderson and Anderson, 2007, p. 15). While the definition of implicit ethical agent meets that of artificial moral agent (AMA), the definition of explicit ethical agent meets the definition of artificial autonomous moral agent (AAMA).

Certainly, the more ambitious objective within the field of machine ethics is to “create a machine that is an explicit ethical agent” (Ibid.). However, fulfilling such an objective faces numerous problems not only from a computational point of view. The gist of the problem is that machine ethics aims to build machines that should demonstrate not only a human-level of cognition, but also something which can be equated to a human-level of morality. In this context, humans are the only full ethical agents. In addition to their ability to make explicit ethical judgments and reasonably justify them, they have consciousness, intentionality and free will (Moor, 2006, p. 20; Moor, 2009). As Moor cogently argues, whether or not robots can become full ethical agents as humans is “a wonderful and speculative topic, but the issue need not to be settled for robot ethics to progress” (Moor, 2009).

That is why he suggests *explicit ethical agents*[2] to become “the paradigm target example of robot ethics”. Such “robots would be sophisticated enough to make robot ethics interesting philosophically and important practically, but not so sophisticated that they might never exist” (Ibid.).

Consequently, elaborating upon the issue of moral reasoning as being irreducible to machine practical reasoning in the field of machine ethics makes room for the following examination. One should analyze why it is so difficult an AAMA to cultivate an autonomous moral concern on the basis of a self-developing moral reasoning.[3] The necessity of clarifying the role of moral reasoning in the field of machine ethics is of general importance. It reveals whether or not the distinction between so-called strong AI scenarios (due to which a human-level of cognition should be reached in a computer) and weak AI scenarios (operating by using preprogrammed algorithms) (Savulescu and Maslen, 2015, p. 84) (Klincewicz, 2016) can be extrapolated into what I would call strong “moral” AI scenarios (looking for an explicit ethical agent) and weak “moral” AI scenarios (designing an implicit ethical agent).

Structure. The main objective of this paper is to demonstrate why arguing for well-working machines, which fulfill moral objectives does not necessarily mean that such machines can be coined moral machines by default. I try to prove that the specification concerns not only the function of explicit (autonomous) ethical agents, but also that of implicit (semi-autonomous) ethical agents in Moor’s sense. Exploring this hypothesis, in Section “Is the mission AMA/AAMA accomplishable?”, I analyze why human practical reasoning is irreducible to machine practical reasoning in moral terms. That is why I examine the implications of the so-called Moral Lag Problem in the field of machine ethics. Specifically, I investigate the computational and moral challenges, which derive from the application of three first-order normative theories in the field of machine ethics. They include Powers’ prospect for a Kantian machine (Section “Some challenges in building “Kantian” moral machines”), Anderson and Anderson’s reinterpretation of act utilitarianism (Section “Building “utilitarian” moral machines”) and Howard and Muntean’s project for a multifaceted AAMA (Section “Howard and Muntean’s project for an AAMA”).

The choice of the three theories is driven by the way in which they exemplify three different prospects for moral machines (deontological, utilitarian and virtue ethical machines). By comparing and contrasting those theories, I aim to reveal how, paradoxically enough, they face similar difficulties when relating the process of computation to that of estimation. For the purposes of revealing whether or not an AI system can generate persuasive arguments using algorithms that formalize moral reasoning, I also examine the problems with the computation of moral feelings and moral motivation faced by a Kantian machine (Section “Kantian machines’ lack of moral feelings”) and an act utilitarian machine (Section “The role of moral feelings in building “act utilitarian” machines”). In conclusion, I outline why both strong and weak “moral” AI scenarios can be questioned due to the issues, which arise from the relation of epistemic to moral predictability in human terms.

Is the mission AMA/AAMA accomplishable? One of the main reasons for the discrepancy between the criteria of epistemological verification and normative validity in morality, as projected into the field of machine ethics, can be found in “the ethos of reducing intelligence to a numeric value” (Markham et al., 2018, p. 4). The ethos displays the “epistemic shift toward “algorithmic knowledge production” ” (Metcalf et al., 2016, p. 6). In this context, one faces the dilemma of whether or not the AI system can generate persuasive arguments using algorithms that formalize moral reasoning based upon first-order normative theories. Such theories include Kantianism, utilitarianism (Klincewicz, 2016, p. 185) and virtue ethics.

In turn, the successful generation of persuasive arguments can contribute to clarifying the implications of the so-called Moral Lag Problem (Ibid., p. 172)—“a shorthand name for all the things that cause us to be not as moral as we could or should be” (Ibid.). Examining the role of the Moral Lag Problem within the field of machine ethics raises some additional concerns. If humans cannot be as moral as they should or wish to be, how can one expect that first, they can avoid projecting their own moral imperfections while building moral machines? And second, and most importantly, would it be possible for machines to become autonomous moral machines? Certainly, the problem cannot be solved if one tackles it within the field of practical reasoning alone.

As Powers cogently argues, one can point out some formal similarities between human practical reasoning and machine practical reasoning. Similar to human practical reasoning, which is grounded into the transformation between the consideration of facts and the resulting actions, machine practical reasoning can be recognized as a transformation from a set of declarative units in a database to an output (Powers, 2006, p. 46). Furthermore, similar to humans who can learn new facts that are informative for their reasoning about actions, machines can incorporate feedback systems that determine their outputs (Ibid.).

Extrapolating the debate into the field of machine ethics makes the differences between human moral reasoning and machine moral reasoning even more complicated. The main concern is that regardless of the fact that humans are not as moral as they should or wish to be, most of them can make up diverse moral claims in contrast to machines which cannot. The reason is that machines have not reached the stage of making normative conclusions that connect facts to action (Ibid.).

In this context, Powers profoundly defines the risk that machines “exhibit a simulacrum of ethical deliberation” (Ibid.) emphasizing the issue that neither humans nor machines can do better. If a system is allowed to decide within the pluralism of human moral judgments, then one will be “left with a static account of morality as it would only be able to simulate, rather than to make judgments” (Lara and Deckers, 2019, p. 5). Consequently, if one argues for an explicit ethical agent that simulates such a static account of morality, this could lead to apparently negative outcomes in moral terms. For instance, explicit ethical agents will create a reality of pure simulacra and turn human imperfect morality into an immoral hyper-reality in Baudrillard’s sense.

Tracing back some of the main reasons behind the risk of having such a simulacrum of ethical deliberation, I would point out the impact of both conscious and unconscious biases of the designers of moral machines. Specifically, the influence of biases can be found in increasing the risks of blurring the decision-making process based on algorithms and the moral decision-making process of humans. The particular outcomes can be recognized on the level of intermingling computational errors with moral mistakes.

The role of biases is determined by the use of the algorithms themselves, since they “inevitably make biased decisions” (Mittelstadt et al., 2016, p. 7). The reason is that algorithms’ design and functionality reflect the values of the designers “and intended uses, if only to the extent that a particular design is preferred as the best or most efficient option” (Ibid.).

In addition, the biased decision-making process gets some new implications when the debate is expanded as an institutional debate, viz. when the power of social institutions is focused upon the shaping of technologies-in-practice (Neff, 2020, p. 5). Furthermore, the complex role of the algorithms in the public discourse is due to the fact that they are not static entities, but have social power. This power depends on the way in which they are “a part of broader rationalities, broader programs of social change and development” (Beer, 2016). Specifically, algorithmic power “produces” truths as outcomes of systems being a part of discursive reinforcement of given norms (Ibid.). If one wants to reveal the origin of algorithmic power, one should not only unpack “the full socio-technical assemblage of algorithms” (Kitchin, 2017, p. 25), but also examine the particular moral implications of this assemblage. The implications concern the risk of intermingling computational errors with moral mistakes.

Analyzing the possibility for deliberately or non-deliberately intermingling computational errors with moral mistakes, one faces the following questions. Is it possible the inevitable errors in the datasets to be described as “human” errors? If not, can they be examined as purely algorithmic errors? The risk of intermingling computational errors with moral mistakes has a deeper origin. The concern is whether or not multiple (new) normative conclusions can be computationally traced in moral terms so that they guarantee a morally acceptable functioning and a possible moral self-update of the moral machines in the future.

Tackling the aforementioned dilemmas, one can find two possible scenarios, at least. Either there are no specifically algorithmic errors, which shifts the debate back again to the origin of the biases behind human mistakes, or there are purely algorithmic errors. If so, however, the chances of building moral machines that are supposed to be not only explicit ethical agents, but also agents with an unquestionably moral (in its “content”) behavior, become slight. Another issue is who is responsible for the regular update of moral rules. One should take into account that AMAs and AAMAs face the learning task “to deal with the inherently ambiguous nature of human moral judgment” and the fact that “even experts can disagree” about this matter (Wallach and Allen, 2010, p. 97).

The scenario of experts’ disagreement becomes even sharper when the disagreement is a result from the projection of the engineers’ biases, viz. when they want to design and apply a moral machine fostering their own ends “which may be biased by their own interests” (Lara and Deckers, 2019). Such a deliberate programming in moral terms breaks not only with the principle of moral impartiality, but also with the idea of machines’ moral autonomy.

In turn, the issue of having immoral machines raises some serious concerns, which go beyond the field of machine ethics. This is due to the fact that learning is not abstracted from the general processes of social and political learning and interaction. In this context, the socio-political implications of biases trigger the issue of how the conditions of “a composite of human-algorithm relation” turn into “venues for ethicopolitics” (Amoore, 2020, p. 9).[4] Ethicopolitics as such results from the acknowledgement that algorithms contain multiple potentials for cruelties, surprises, violence and joy (Ibid., p. 7). Consequently, the potentials are embodied into illuminative representations of the human relations with both other humans and algorithms themselves.

As long as those complex relations are recognized as politically formed, the reconsideration of the socio-political and moral implications of biases can make room for locating the ethicopolitical origin of creating immoral machines. The incomplete character of algorithms is both weakness and strength depending on the starting point of investigation. Regarding the implications of strength, incompleteness of algorithms can be justified as a stimulus for revising the ethicopolitics they introduce.[5] Thus, some new ways of "generating worlds" (Ibid., p. 21) can be established if doubt is employed productively in the process of algorithmic revision (Ibid., p. 24). However, adopting such an approach does not make the issue of biases less substantial, since every single regeneration brings its own ethicopolitical narrative. This narrative can be deprived by some of the old biases, but brings new ones with itself.

Some challenges in building "Kantian" moral machines. A crucial reason for choosing Powers' model is that he cogently demonstrates why logic, which grounds the design of a machine-computable categorical imperative, faces significant constraints. Consequently, those constraints necessitate an analysis of the corresponding ethical problems to be conducted. Powers emphasizes that nonmonotonic logic approach can better meet the requirements of a deontological machine ethics than monotonic logic approach. However, it fails the requirement of semidecidability of a set membership (Powers, 2006, p. 50). Powers' specification requires the examination of whether or not a Kantian machine can be underlined by what I call corresponding nonmonotonicity in the field of ethics. In addition, one should explore what can be considered as corresponding constraints of semidecidability within the field of machine ethics. The crossing point between a nonmonotonic approach in logic and its potential projections in ethics can be found in the realm of permissible maxims. Specifically, the status of permissible maxims is relevant to both human reasoner and computer program because it concerns the degree of precision of one maxim (Klincewicz, 2017, p. 249).

Powers formalizes monotonic reasoning in the following manner: if you can infer sentence *a* from a set of premises P, then you can also infer *a* from any set S that contains P as a subset. For instance, the addition of "Fritz is a cat" to a set already including "All cats are mammals" licenses the monotonic inference "Fritz is a mammal" (Powers, 2006, p. 49). However, if the deductive law is replaced with a default rule such as "Cats are affectionate", there would be some conditions that would defeat the inference to "Fritz is affectionate" (Ibid.). As Powers cogently argues, the approach becomes even more complicated when applied to an ethical default rule, as is in the case with the default rule "Don't kill the innocent". The rule requires one to take into account some defeating conditions such as "unless they are attacking under the control of some drug" or "except in a just war" etc. (Ibid.).

Judging by the reasons behind Powers' choice of a nonmonotonic logic approach over a monotonic one (Ibid.), I would suggest extrapolating the idea of a nonmonotonic approach to the field of ethics. Adopting such an approach can contribute to better understanding the prospect for a Kantian machine in moral terms. As long as in the field of nonmonotonic logic one draws conclusions defeasibly[6] (when one reserves the right to retract them in the light of further information), tackling the nonmonotonic projection within machine ethics would be of use for the moral self-update of AMAs and AAMAs. In turn, the update in question will be made in the light of new (ethical) knowledge (information). Such a shift can shed light upon the status of permissible maxims, which are neither obligatory nor forbidden maxims.

From the perspective of the elaborated nonmonotonic approach, the role of permissible maxims can be clarified if one examines what Powers borrows from Reiter's default logic, namely, the so-called default extensions. Owing to their introduction, the rules can be defeated, but not vanquished (Ibid.). If evaluated from the perspective of the extrapolated default extensions in moral terms, permissible maxims can be settled in a Kantian machine as input maxims. Thus, they may contradict "the background set of facts and commonsense rules without introducing inconsistency" (Ibid.). Consequently, making room for the examination of some contradictions as irreducible to inconsistences would enrich both the number and the complexity of normative rules.

However, providing such a correspondence between nonmonotonicity in logic and ethics raises the significant concern how a Kantian machine can be protected from turning into an immoral machine if it includes input maxims, which can contradict the set of facts. Elaborating upon this line of thought requires examining the reasons behind the next important conclusion made by Powers. According to him, nonmonotonic inference fails the test of semidecidability of a set membership.

First, I will clarify the purely logical implications of semidecidability that Powers lists as being one of the most serious concerns against building a deontological ethical machine. Second, I will try to demonstrate why a possible semidecidability in moral terms is even more complicated than the logical one.

In computing theory, semidecidability demonstrates that if an element can be a member of a set, then the algorithm halts with a positive result. Consequently, if an element is not a member of a set, then the algorithm halts with a negative result. What is of particular interest for a Kantian machine is the third aspect of semidecidability, namely, the one that if an element is not a member of a set, then the algorithm does not halt at all.

Powers argues that a Kantian imperative can be described as being a nonmonotonically enhanced categorical imperative which, however, "still fails an important *formal* requirement for machine ethics: semidecidability" (Ibid., p. 50). I would argue that the lack of so to speak ethical implications of semidecidability triggers two negative effects, at least. First, it affects the role of permissible maxims as input maxims and second, the possibility for a moral self-update. If the deontological ethical machines lack the opportunity to include the third aspect of semidecidability, it would mean that they lack the opportunity for looking for a new ethical knowledge and its alternative ways of computation in long terms.

The second circle of issues addresses the specifically moral problems of the agency, which a Kantian machine should exert. As one of the main concerns in this context, I would point out the problem that Powers raises himself, namely, that there are no unquestionable proofs that a machine will understand an obligation in the same manner as a human agent (Ibid., p. 48). He argues that a machine might not be able to understand the difference between "I ought to do z" and "z ought to be the case" (Ibid.). The difficulty raises serious concerns because it is not specific for machine ethics. It derives from the complexity of human moral motivation, since humans may not be able to make such a difference either.[7]

Kantian machines' lack of moral feelings. The analysis of the negligence of moral feelings within the prospect for a Kantian machine is a different way of tackling some of the reasons behind the following conclusion drawn by Powers. Kantian formalism "in the constructivist or "bottom-up" tradition can't build a coherent moral theory from nothing" (Powers, 2006, p. 51).

In this context, one should first make a difference between feelings and moral feelings[8] and second, keep in mind the constructive role of moral feelings such as sympathy and empathy for humans. In addition, one should also be aware that moral feelings cannot be successfully designed in moral machines yet.

If the impact of moral feelings is underrated, it would lead again to the difficulties of how one defines the role of Kantian explicit ethical agents in epistemic and moral terms. Specifically, one will face again the problems in guaranteeing that a maxim that meets the tests of universalizability meets the test of being morally universalizable at once. Furthermore, one is still concerned about making a choice of how to give priority to two or more contradicting maxims, which adhere to the formal requirements of the Categorical Imperative (Stahl, 2004, p. 77).

If we one seeks to fulfill a strong "moral" AI scenario in which ethical intelligent agents can reach a human-level of ethical knowledge, one should be assured that the moral self-update of the AAMAs guarantees that after the update, those machines will remain at least as moral as they were before.

Tackling the role of moral feelings for the update of moral machines points towards the utopian scenario of some highly developed explicit ethical agents, which can manage to master such an update in the distant future. One of the main reasons for this vague prediction is that developing moral feelings is a double-step process. At first sight, it can be considered as a methodological benefit that computers "as emotion-free machines might again be said to be in a good position" to fulfill the criterion of impartiality in Kant's sense (Ibid.). However, as Wallach and Allen have cogently pointed out, it is unlikely that robots develop empathy to other entities, unless they cultivate emotions on their own (Wallach and Allen, 2010, p. 165).

So, what would be the role of moral feelings embodied in a Kantian ethical machine? A relevant alternative can be found in the way in which Wallach and Allen describe how an AMA should follow the Golden Rule. While working to notice the effect of others' actions on itself, assessing the effect and choosing its preferences, as well as assessing the consequences of its own actions on the affective states of others, it should work on assessing one's individual psychology (Ibid., p. 96).

In practical terms, it means building an AMA that has the capacity to discern and anticipate changes in the "affective reactions" of people to its decisions. However, both Wallach and Allen are aware that predicting the actual action on the basis of those reactions is almost impossible. The deeper origin of the problem with the affective reactions in the field of machine ethics is that there are no proofs for computers' sensitivity to human beings' moral interests (Lara and Deckers, 2019, p. 5.)

In addition, I would argue that even if such affective reactions are fully understandable and then, predictable, they are only a necessary condition for understanding moral feelings, which humans themselves still disagree about.

Building "utilitarian" moral machines. As another ethical theory that might give some potential answers as to "whether ethics is the sort of thing that can be computed" (Anderson and Anderson, 2007, p. 18), researchers point out so-called act utilitarianism. According to the latter, rightness and wrongness of actions are determined by the consequences of the actions alone. Before examining how such a moral arithmetic can be composed in the field of machine ethics, one should know what counts as "good" and "bad" consequences in themselves (Ibid.) regarding interhuman interactions. Even if one assumes that due to a precise moral arithmetic one can calculate the best possible actions, which maximize the total net of goods, nevertheless, there is no normatively grounded reason for one to avoid these actions to be considered immoral in their "contents".

Regarding machine ethics, Anderson and Anderson point out the challenges of "computing the ethically correct action, given this information" (Ibid.). The challenges of computation concern the exact possibilities and mechanisms of "measuring" utility in perspective by taking into account the result alone. Consequently, there is no guarantee that the engineers of ethical intelligent agents will be able to overcome the pitfalls of their own (human) moral arithmetic regarding what good action should look like. That is why they cannot successfully create both implicit and explicit ethical agents either.

Concerning the examination of permissible maxims and the possibility of a moral self-update of utilitarian machines, one faces the following contradictions. Regardless of the fact that machine-computable categorical imperative and act utilitarian imperative are based on mutually exclusive patterns of modeling (maxim-based computational model vs. result-based computational model), act utilitarian modeling can also be considered as lacking both logical and so to speak ethical projections of semidecidability. This means that act utilitarianism does not clearly encourage the application of permissible maxims as input maxims.

However, the reasons for that are different. The objective of act utilitarianism is to provide algorithms that should discern the classes of forbidden from obligatory actions, not those of maxims. In the best possible scenario, permissible actions are tacitly included as a sub-group of obligatory actions. Permissible actions are of secondary interest if they meet the requirements of maximizing the utilization of "good" results, as do obligatory actions.

Another moral concern is that an algorithm's uncertainty displays a more elaborated version of classic objection to utilitarianism, viz. that it would take too long for a person to sort through all the options and consequences in order to avoid uncertainty or unwanted results (Klincewicz, 2017, p. 248). This means that a utilitarian computer reasoner should be optimally efficient in terms of time and result. Thus, the efficiency of the moral advice will be guaranteed in both temporal and context terms. One of the objections to such an approach is that the efficiency in epistemic terms does not necessarily trigger moral relevance, unless one argues for an ideal example.

Furthermore, similar to a Kantian ethical machine, implementing utilitarian principles into moral machines does not provide "a base of standard morally unambiguous cases" (Ibid., p. 254). In addition, there are fundamental problems concerning utilitarianism that cannot be solved by computers such as what is utility, how it can be measured, and how interpersonal comparisons of utilities can be realized (Stahl, 2004, p. 77).

Paradoxically enough, there is a risk for both deontological and utilitarian moral machines that they may not be able to understand the moral complexity of agent's perspective. Thus, one inevitably shares again Powers' concern about whether or not machines can make a difference between "I ought to do z" and "z ought to be the case" (Powers, 2006, p. 48). A serious problem of act utilitarianism as a result-based model is how the engineer can understand whether at the moment of designing and developing the moral machine, he has enough information about the future consequences regarding goodness. The particular difficulties for the moral self-update are related to the fact that the lack of clear permissible input maxims makes the prediction of consequences problematic in moral terms.

For instance, human mistakes do not coincide with computation errors by default. This is due to the fact that human moral agents can make mistakes, but not errors in their moral practice. Technically speaking, error shows the gap between the calculated value and what is justified as true value, viz. the "rightness" or "wrongness" of data. Another point is that wrong data can be a result of a wrong moral choice. The error can be a result of a moral mistake made by the engineer as a moral agent and vice versa. Underrated or unrecognized mistakes can influence the behavior of moral agents so that they make moral mistakes. In addition, one should keep in mind that particular actions that might appear as errors or aberrations are in fact intrinsic to algorithms' experimental and generative capacities.[9]

However, some investigations in the field of machine ethics are focused upon preventing the possible errors in moral arithmetic due to the ambiguous assumption that the avoidance of those errors would automatically lead to avoiding moral mistakes. As Wallach and Allen cogently argue, most criticism of act utilitarianism is focused upon how to protect utilitarian AMAs from "an endless stream of calculations" (Wallach and Allen, 2010, p. 89), but, as I would add, not of rethinking the risk of establishing an endless stream of estimations.

Speaking in practical terms, encouraging an endless stream of estimation of good as useful would encourage reducing morality to a cost-benefit analysis. This analysis underlines meritocratic morality based upon evaluating economic merits as unquestionably moral if they are of economic use. Elaborating upon the discussion within the field of machine ethics, it would mean that all the criticism one can raise against the concept of a Condorcetian "mathématique sociale" (Feldman, 2005)—and its present-day representatives in the various schools of cost-benefit analysis in economic practice—applies even more cogently to a moral mathematics favoring strong AI scenarios.

The role of moral feelings in building "act utilitarian" machines. What is the status of moral feelings within the paradigm of act utilitarianism? Against the accusations of utilitarianism as being "cold and unsympathizing", Anderson raises the argument that the happiness and unhappiness of others can hardly make one "cold and unsympathizing", as is demonstrated by Mill's own hedonistic utilitarianism (Anderson, 2008). The main issue, which is of high relevance for machine ethics as well, is whether or not happiness can be unquestionably examined as a matter of goodness.

Certainly, the equation of goodness with happiness displays rather an ideal case. In addition to the question of whether or not intelligent agents (IAs) can have feelings at all, one faces the following dilemma. It is unclear whether or not at least some of those feelings if any meet the requirements of moral feelings.

All the aforementioned difficulties trigger some new difficulties as well. For instance, whether or not maximizing what is good for the majority of humans or IAs in general coincides with what is good for the individual or the given agent, respectively. It also raises the problem whether or not what is good and could be defined as potentially encouraging positive moral feelings in the machines meets the requirements of having morally acting machines.

Another problem derives from the side of the engineers who are supposed to program the machines with such moral feelings. Regarding the utilitarianist frame, the computational gist can be summarized as a matter of appropriate application of a relevant metric of happiness, pleasure, or a matrix of preference by the software (Klincewicz, 2017, p. 247). Building such a metric is a challenge in itself even when one tackles the ideal case, namely, when the engineer knows what kind of philosophically related sources he needs to employ.

Therefore, even in this ideal case, when the engineer should use "a philosophically informed set of criteria" (Ibid.), it is not clear how the informed consent can contribute to solving the problem of moral pluralism. The latter raises the corresponding engineering difficulty of encompassing the sheer number of criteria and the magnitude of difference between them (Ibid.). One should also keep in mind the specification that the objective of covering more domains limits the space of possibilities that can be adopted (Ibid.).

This difficulty is not strictly computational. The reason for that is that fulfilling the objective in a moral sense would mean arguing for moral formalism, although not of a Kantian type. The more domains are covered, the less moral variables can be applied in a non-contradictory manner. That is why adopting such an approach would result into moral absolutism due to which only few variables will be considered as universalizable in moral terms.

Howard and Muntean's project for an AAMA. Howard and Muntean's project (2017) aims at building a multifaceted AAMA ideal model by adopting the concepts of learning and adaptation (Howard and Muntean, 2017, p. 124). What is of particular interest in this model is that it gives some clues as to how an autonomous moral machine can achieve a moral self-update. The update is grounded in the authors' objective to replace the action-centered model by an agent-centered model. The latter encourages virtue ethics for a machine based on active moral learning (Ibid., pp. 121, 126).

In turn, the active learning is determined through the practices of moral learning[10] and moral development of dispositional traits by relying upon so-called soft computation (Ibid., p. 140)—neural networks, which are elaborated upon by evolutionary computation.[11] According to Howard and Muntean, each AAMA has a set of dispositional traits, or possible behaviors, which play "a role similar to "dispositional moral virtues" " (Ibid., p. 126). In such a "minimal model", virtue may be considered as a dispositional trait that is cultivated through a process of learning (Ibid., p. 135).

Certainly, what Howard and Muntean consider as a possibility for grounding artificial morality within the framework of so-called moral dispositional functionalism[12] (Ibid., p. 132) contributes to expanding the boundaries of the debate in new directions. The methodological strength of the project is that the authors explicitly define behavior reading as a "weaker" version of "moral" reading. Specifically, the project does not encourage one to decode human moral behavior in its complexity, but shifts the focus to the similarities between humans and machines. It rather assumes "as little as possible about the true nature of human morality, and enquires into the replicability and the scalability of these [similar] features, outside of this model" (Ibid., p. 130).

Judging by those specifications, one could coin Howard and Muntean's project as promoting a "weaker" version of an explicit ethical agent. It does not aim to reach the level of human morality in general, but only to "read" some behavioral moral similarities, which can be computed. This assumption is clearly stated when Howard and Muntean describe the virtue ethics of AAMAs as having a "partial analogy with human virtue ethics" (Ibid., p. 141).

Regarding the benefit of adopting a virtue ethical approach for the purposes of a moral self-update, one makes a step forward towards avoiding some risks posed by utilitarian and deontological approaches. A self-update grounded in virtues can be evaluated as minimizing the risks of blurring the categories of "useful" and "good" (act and rule utilitarian approaches). It can also minimize the restriction of moral agency to the update of duties alone (as can happen if the model of deontological ethics is adopted uncritically).

Another important contribution in this respect is Howard and Muntean's specification that AAMAs' active learning should be grounded in a pattern recognition structure. The latter includes "non-linearity, emergence, errors, noise, or irrelevant data" (Ibid., p. 143). The consideration of those factors makes the design less ambitious in terms of moral perfection. However, it is more critical towards the opportunity of rejecting some data if the data is recognized as unreliable in moral terms. Furthermore, the authors cogently try to limit in advance the moral concerns, which may derive from the inconsistencies between humans as moral agents and moral machines. They introduce some limits by assuming that AAMAs are analogous to human moral agents (Ibid., p. 153).

While the awareness of AAMAs' imperfection and corresponding imperfectability is a strength, the methodological concerns derive from the consequences of so-called by the authors computational "*minima moralia*" for an AAMA (Ibid., p. 131). The negative consequences can be examined on the level of deliberately limiting the analysis to the behavioral replicability of moral actions by so-called black-box strategy. Consequently, the authors apply this strategy by circumventing the replicability of mental and psychological mechanisms of moral judgments, as well as the rational and deductive aspects of morality (Ibid.).

Adopting such an approach initially disregards the complexity of human moral motivation and moral reasoning, which is irreducible to computation processes. In this context, the complexity of dis-analogies is considered as being somehow limited by the assumption that moral data is a collection of unknown and "possibly very complex patterns" (Ibid., p. 142). However, the specification does not prevent the analysis from going towards bad moral infinity, while arguing for AAMAs' moral self-update. As one of the most apparent risks regarding such an update, I would point out the risk brought with so-called exhaustive enhancement criticized by Lara and Deckers (2019, p. 3). This enhancement is negatively evaluated due to the fact that the decision-making process is delegated entirely to the machines. Such a delegation is possible only as a matter of theoretical speculations at that stage. The reason is that there is no guarantee that moral maturity of humans as a process of development can be replicated by machines in the process of their moral self-update.[13]

When one claims that the advanced AAMA can pick relevant patterns of dependency among relevant variables, although there is always missing information about the variable of moral behavior (Howard and Muntean, 2017, p. 142), nothing guarantees that the information in question will be relevantly interpreted by the AAMA in moral terms. Practically speaking, the missing information may lead to having patterns with dis-analogies, which to be black-boxed for the sake of keeping the pattern and thus, make it unusable. In this case, it will show a limited picture of behavior reading or even worse. It could be accepted uncritically and trigger the programming of immoral AAMAs, which to be considered as moral machines.

Consequently, one of the main concerns about the limitations is that the analogy between humans and moral machines is used as an excuse of machines' moral imperfection and the corresponding imperfectability for the sake of grounding their autonomy. Howard and Muntean adopt a milder than necessary version of criticism towards the limitations. They argue that there will "be always bad results and "too-good-to-be-true" results". The conjecture here is that, statistically and in the long run, such an AAMA model or something similar to it will gain "model robustness" (Ibid., p. 154). The reasons behind why such robustness is questionable in moral terms can be traced back to the specification that AAMAs' moral self-update cannot be evaluated in progressive terms. Therefore, extrapolating Laura and Deckers' idea that exhaustive enhancement excludes the possibility for moral progress (Lara and Deckers, 2019, p. 4), one may argue that moral self-update cannot be guaranteed as a matter of moral development by default.

The conviction that regardless of the moral imperfections, moral self-update of such machines will result somehow into a moral perfection in the long-run raises more concerns than it looks like. The moral concern is irreducible to the worry about epistemic probability. On the contrary, it is a concern about moral predictability, viz. a concern about the consequences from a possible immoral machine's behavior. It triggers the issue of how one takes responsibility for such consequences.

Howard and Muntean are aware of some of those concerns. They claim that the criteria for when a bundle of features of a population of NNs (neural networks) is a "virtue" of the AAMA, as well as those of whether or not one can build a virtue from a set of potential behaviors (Howard and Muntean, 2017, p. 154) require further investigation. Regarding the question of responsibility, the authors argue that one should not worry about the predictability of the AAMAs' (im)moral behavior, although one should be prepared to disable such machines at any time (Ibid.).

Certainly, the possibility for disabling is the most appropriate action in moral terms; specifically, when one examines the impact of AAMAs, which are used in fields such as medicine and military industry. In this context, discussing the role of moral agents' responsibility as being closely tied with that of biases is of crucial importance. The debate covers many issues, which need to be tackled in moral terms. Such an issue is whether or not the people responsible for disabling AAMAs would like to do it being guided by immoral motivation and/or external reasons of a non-moral origin. A special place within motivational factors takes human trust when recognizing the risk and its elimination.[14]

In turn, the practical concerns about Howard and Muntean's project are related to the AAMA's ability to "read" and systematize moral behavior. According to the authors, the AAMA's ability to "read" the behavior of human agents depends upon the quantification of the behavior. It is recognized as a matter of collected data, which excludes the role of human moral motivation and moral judgments. Howard and Muntean aim at premising their project in the most objective manner. It should be "loosely premised only on the knowledge we have about the Neural Networks and Evolutionary Computation, and not on what we know about the nature of moral cognition of humans" (Ibid., p. 146).

Howard and Muntean see the complex limitations of "reading" people's behavior as raising the following concern: "reading moral behavior from data" puts "a heavy load" on the evolutionary computation. This is due to the fact that the search goes towards potentially infinite dimensionality (Ibid., pp. 147–148). The problem is that for the purposes of overcoming this methodological difficulty, the authors suggest the introduction of purely computational constraints such as limiting the number of hidden layers, simplicity, speed etc. (Ibid., p. 148, Note 32). However, those limitations do not shed light upon the problem how one can choose between morally contradicting, but not mutually exclusive behaviors on the level of input maxims.

Specifically, gathering data, as well as composing neural networks initially depend upon the preferences of those who collect and systematize the data in question. It is the engineers who determine the number of layers, the criteria for simplicity and speed etc. Regardless of the fact that Howard and Muntean distribute the responsibilities in the process of soft computing to so-called evolver—the person who makes the decision about the evolutionary computation and who can be a different person than the programmer or the trainer of the neural network (Ibid., p. 152)—this does not make the issue of biases less complicated.

One faces the concern about distributed responsibility for gathering morally relevant data when it is ascribed to more than one agent involved. Consequently, the increasing number of responsible agents increases the number and the impact of both conscious and unconscious biases. The distributed responsibility as such concerns the degree of uncertainty and vagueness, which should be left in the system within the process of soft computing. It means that biases are still influential at the first level of input data.

Another issue derives from the assumption that the evolvable features of each AAMA, which the authors associate with so-called robo-virtues, can radically differ in the different scenarios (Ibid.). Adopting such an approach in moral terms may lead to uncontrollable moral relativism and even to immoralism. Speaking in moral terms, the evolution of populations of the neural networks raises the risk virtues to become a “hostage” of the idea of moral autonomy. Thus, the strive for autonomy increases the risk of having autonomous immoral machines. Furthermore, if what is evolvable from one generation of the AAMA to the next, as well as what can be defined as initial populations, which are inherited from one scenario to another are empirical questions (Ibid.), then the following issue arises. How can the distribution of moral competence in the process of evolution guarantee not only a minimal and optimal set of virtues, as Howard and Muntean argue, but also those virtues to assure the development of moral scenarios alone?

The second time when biases play a crucial role in Howard and Muntean’s model is on the level of the categorization of data. According to this model, the process of categorization is based upon building analogies between artificial cognition, human cognition and moral action. In turn, the categorization as such raises some problems that merit further investigation. First, as was it shown with the previous models of Kantian and act utilitarian moral machines, human cognition does not necessarily coincide with human moral reasoning. Specifically, the cognition of moral reasoning does not necessarily represent the moral specificities of this reasoning nor is every practical reasoning moral reasoning in itself.

Second, the building of learning patterns from data on the basis of relatively clear analogies between human and artificial cognition is not unquestionable. The building can face numerous difficulties when one witnesses a lack of analogies such as those concerning moral feelings and complexity of human moral motivation in general. On this level, the role of both conscious and unconscious biases should be carefully examined because they have not only negative, but also positive influences. As an example of positive influence, one can point out the encouragement of moral feelings such as empathy and sympathy.

Then again, the problem is how soft computing can contribute to building AAMAs, which can develop (or actively learn in Howard and Muntean’s sense) moral sensitivity, taking into account that even implicit ethical agents have not yet mastered such an ability. The analysis shows that even if one introduces some methodological limitations, the project of “reading moral behavior from data” (Ibid., p. 147) provokes more questions than providing answers.

Conclusion

By exploring the difficulties in finding satisfactory answers to the questions “Whose morality?” and “Which rationality?” within the field of machine ethics, I raise some concerns about building an ethical intelligent agent. Those difficulties address the issue how such an agent would reach, (if not a human-level), at least a verifiable level of cognition corresponding to that of human ethical knowledge and human moral reasoning. Specifically, I examine how the crossing point between the questions “Whose morality?” and “Which rationality?” can be found in the exploration of complex moral reasoning, as adopted by different groups of ethical intelligent agents.

The primary objective of this paper is to demonstrate that building well-working AI machines, which serve for moral purposes (both as autonomous and semi-autonomous moral agents) does not mean building moral machines by default. As a main reason for such a complication, I point out the risks of reducing intelligence to a numeric value in the computation process, while arguing for a moral reasoning.

Analyzing the role of the so-called Moral Lag Problem within the field of machine ethics, I aim at revealing the origin of Powers’ concern whether or not a moral machine could exhibit only a simulacrum of ethical deliberation. The gist of this simulacrum can be determined as being triggered by the risks that moral machines replicate human morality, with all its imperfection. This is due to the role of both conscious and unconscious biases of the engineers. Biases’ influence as such concerns the risk of equating the decision-making based on algorithms with the human moral decision-making process. Consequently, the particular negative effects of such a potential misrecognition can be traced back to the level of intermingling computational errors with moral mistakes.

The methodological discrepancy between computational process and moral reasoning is exemplified by comparing and contrasting three projects for moral machines, viz. Kantian, utilitarian, and virtue ethical machines. The analysis shows that paradoxically enough, regardless of the different ethical models that are adopted, none of these three projects go far enough in building autonomous moral machines. Furthermore, those different, in moral terms, projects face similar problems in relating computation to estimation. The problems affect the respective correspondence between epistemic and moral predictability, while building such machines.

Examining the origin of Powers’ concerns about the application of nonmonotonic logic approach into building deontological moral machines and the lack of semidecidability it suffers from, I argue that one can find corresponding constraints within the field of ethics. Consequently, there are corresponding constraints of semidecidability if they are extrapolated into the field of machine ethics. That is why I raise the hypothesis that the methodological constraints of nonmonotonic approach in logic and its potential projections in ethics, while building a Kantian machine, concern the status of permissible maxims as input maxims. Clarifying the status of permissible maxims in nonmonotonic terms expands the possibility of having such maxims in the field of machine ethics. This is due to the nonmonotonic requirement that the possible contradiction to the facts would not lead to moral inconsistencies by default.

In turn, the main objections to the ethical projections of semidecidability address the possibility for a moral self-update of a moral machine, viz. the third aspect of semidecidability (when an element is not a part of the set, then the algorithm does not halt at all). If this aspect is neglected, it would mean that the moral machine would lack the ability to look for a new ethical knowledge and its alternative ways of computation.

The risk of moral inconsistencies resulting from the computation process is exemplified, although from a different perspective, by Anderson and Anderson’s interpretation of act utilitarianism as well. Specifically, I have examined the reasons behind their statement saying that before exploring how a moral arithmetic (based upon maximizing utility as goodness) can be composed in the field of machine ethics, one should know what counts as “good” and “bad” consequences in themselves. Act utilitarian modeling shows that moral arithmetic, which is designed to avoid errors, does not automatically lead to avoiding moral mistakes.

Elaborating upon the problems act utilitarian moral machines face, I compare and contrast the project of such machines with that of a Kantian machine. In this context, I reach the following paradoxical, at first sight, conclusion. Regardless of the fact that machine-computable categorical imperative and act utilitarian imperative are based on mutually exclusive patterns of modeling (maxim-based computational model vs. result-based computational model), an act utilitarian model of moral machines lacks similar logical and, so to speak ethically justifiable, semidecidability.

However, in the case of act utilitarianism, ethical implications of semidecidability have different embodiments regarding the status of permissible maxims and moral self-update.

Concerning the status of permissible maxims, they are of secondary interest because in act utilitarianism, one argues for permissible actions. In the best possible scenario, permissible actions can be treated as a sub-class of obligatory actions if they meet the requirements of maximizing the utilization of "good" results. Regarding moral self-update, similar to Powers' alerts to a Kantian machine, in an act utilitarian machine, one witnesses a problem with machine's ability to understand the nuances of moral motivation. However, the difference is that since act utilitarianism is determined due to the goodness of results—not due to that of premises—act utilitarian machines would lack clear (moral) input maxims. Thus, the lack itself will make the moral self-update entirely dependent on the consequences alone.

In this context, I tackle the reasons behind what Wallach and Allen define as "an endless stream of calculations" in act utilitarianism. I also examine the corresponding moral risks of encouraging an endless stream of estimation. Specifically, one should question the assumption that the avoidance of computational errors would automatically lead to the avoidance of moral mistakes.

In turn, the problems of "reading moral behavior from data" are exemplified with the third project of a moral machine, which is based upon a virtue ethical approach. The main contribution of Howard and Muntean's project of building an AAMA is that it does not have the pretention to build a global moral machine. Expanding Moor's terminology, this moral machine can be described as a "weaker" version of an explicit ethical agent. Consequently, the strength of the "weaknesses" consists in defining the objective of the computational process of behavior reading as a reading, which does not aim to decode human moral behavior in its complexity.

Being unable to "read" the complexity of human moral agency, such a machine would not be able to properly "read" and conceptualize human moral motivation and moral feelings either. Thus, it would leave "unread" many significant aspects of complex human behavior. Building a moral machine upon the computation of potential moral similarities between humans and machines raises one of Powers' arguments in a new voice, namely, whether or not the machine exhibits a simulacrum of human deliberation. That is why there is a risk that this "weaker" version of an explicit ethical agent becomes as vulnerable as humans who are imperfect moral agents.

Furthermore, the methodological strength and weakness of Howard and Muntean's project have one and the same crossing point, viz., so-called by the authors, computational "minima moralia" for AAMAs. The methodological benefits concern the way in which moral self-update grounded in virtues contributes to minimizing the risks of blurring the categories of "useful" and "good" (act and rule utilitarian approaches) in the process of behavior reading. In addition, this self-update also minimizes the restriction of moral agency to the update of duties alone (as can happen if the model of deontological ethics is adopted uncritically). Second, the specification that AAMAs' active learning should be grounded in a pattern recognition structure makes the project less ambitious in terms of moral perfection. However, it increases the possibility for updating and correcting some data if it is considered as unreliable in moral terms.

Regarding the methodological disadvantages, the main issue concerns the initial adoption of the black-box strategy. The strategy is applied to the replicability of mental and psychological mechanisms of moral judgments, as well as to the rational and deductive aspects of morality. However, limiting the analysis to the behavioral replicability of moral actions hides the risk the analogy between humans and moral machines to be used as an excuse of machines' moral imperfection and imperfectability for the sake of grounding their autonomy.

The particular implementations of the risks of having an immoral AAMA are traceable to the two main processes of computation regarding the behavior reading and the systematization of data. Both processes initially depend upon the biases of the programmers and, so-called by the authors, evolvers who are responsible for robo-virtues: these are virtues, which can radically differ in the different scenarios. The issue is that if robo-virtues can radically differ, they cannot guarantee that in the process of the moral self-update those scenarios will not turn into immoral scenarios.

In conclusion, comparing and contrasting the three models of moral machines shows that human cognition does not necessarily coincide with human moral reasoning. This means that cognition regarding moral reasoning does not necessarily represent the moral specificities of such a reasoning nor is every practical reasoning moral reasoning in itself. Analogical thinking that grounds the idea of replicating human and machine morality faces the serious concern about the lack of relevant analogies between humans and machines; specifically, in terms of moral feelings and complex moral motivation. It also triggers the concern that the lack of clarity regarding moral feelings and motivation is of a strictly human origin.

What are the general conclusions regarding what I called weak and strong "moral" AI scenarios? In addition to the problems inherent to having an AI reaching a human-level of cognition ("strong" AI scenarios), those concerning the impossibility of achieving a human-level of morality ("strong" moral AI scenarios) still occur. The origin of the ethical difficulty should be traced back to the issue that only humans can become full ethical agents in Moor's sense.

In turn, the question of machine's self-updating process in moral terms is a question that has not been answered yet due to the lack of necessary correspondence between epistemic and moral predictability. One of the main reasons for this discrepancy derives from the assumption that even if humans manage to design a moral reasoner system that can successfully persuade reason-responsive people, it does not follow that those people can be persuaded in moral terms by default, nor are there guarantees that what they can be persuaded to do is morally justifiable either.

In the end, the problem with answering the questions "Whose morality?" and "Which rationality" is that they should be both raised and answered by humans as being full ethical agents. That is why the issue is what would guarantee that an AAMA can become more moral than humans once and for all so that its moral perfection lasts unlimited through a process of constant morally perfect updates.

At this stage, the two alternatives—getting inspiration from somewhere else and machine self-building—could be considered as an object of successful investigations within the field of science fiction alone. However, what could be of some use for humans as being full ethical agents is what Wallach and Allen emphasize as an outcome of the explorations in the field of machine ethics, namely, in addressing the challenges, humans will understand what truly remarkable creatures they are (Wallach and Allen, 2010, p. 11).

Data availability

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Received: 29 April 2020; Accepted: 23 September 2020;
Published online: 07 October 2020

Notes

[1] For the differences between traditional (human) decision-making and decision-making based upon algorithms, see Mittelstadt et al., 2016.

[2] As some practical suggestions for how to build explicit ethical agents, Moor points out programming a computer "with a large script that selects information relevant to making ethical decisions", which allows it to process "the information to produce ethical judgments". "Alternatively, ethical insights might be acquired through training a neural net or through evolution by a genetic algorithm. However, it might be done, the intriguing possibility is that one day ethics could be understood by a machine" (Moor, 2009).

[3] According to Klincewicz, the recognition of moral reasoning is a two-folded process. First, it requires a rational justification of the process of reasoning as moral to be provided. Second, this process must be non-contradictorily applied to a computer (Klincewicz, 2016, p. 179).

[4] The idea of relating humans and algorithms as a matter of ethicopolitical relations is set by Amoore against the background of her project for so-called cloud ethics. Cloud ethics addresses the political formations of relations to oneself and the others, which take place via algorithms (Amoore, 2020, p. 7). It "extends the opacity of the human subject, envisaging a plurality of venues for ethical responsibility in which all selves—human and algorithmic—proceed from their illegibility". (Ibid., p. 8).

[5] According to Amoore, "In their partial and incomplete way of generating worlds, we can locate their ethicopolitics". (Amoore, 2020, p. 21).

[6] As an example of defeasible reasoning in Kant's writings, Powers points out Kant's arguments against suicide and false promising (Powers, 2006, pp. 49–50).

[7] In this context, a prospect for a Kantian machine can be defined from two main perspectives regarding moral agency. As Klincewicz argues, if one wants "to have a Kantian moral reasoner program with human-level intentions and rationality, then the task may be beyond what we can presently achieve. However, if the engineering task is to create a system that acts consistently with the moral law, then arguably the project is not as difficult" (Klincewicz, 2017, p. 248). Judging by his definitions, one can claim that the first type of a Kantian moral reasoner meets Moor's definition of explicit ethical agent, while the second one meets that of implicit ethical agent.

[8] Wallach and Allen cogently outline that a (ro)bot which has emotions is a "virtual Pandora's box" (Wallach and Allen, 2010, p. 196). Such a (ro)bot, however, would make people even more concerned about which moral emotions (feelings) would leave the box in question. As one of the most apparent fears regarding Pandora's box, I would point out that of a feeling which formally meets the definition of a moral feeling, but has an immoral "content".

[9] See Amoore, 2020, p. 23.

[10] Another model of agent-computer moral learning is demonstrated by Lara and Deckers' Socratic enhancement. By adopting the Socratic method of interrogation, humans as moral agents are supposed to "reach better moral judgments and realizable behavioral options that cohere with those judgments" without granting moral autonomy of the machines in question (Lara and Deckers, 2019, p. 12).

[11] According to Howard and Muntean, "the evolving artificial neural networks" approach employs a population of neural networks (NNs) and evolutionary computation (EC)", which are presented by the formula NN + EC (Ibid., p. 143).

[12] Moral dispositional functionalism corresponds to semantic naturalism. The model is premised on the complicated model that "moral properties are *determined* by nothing else than natural properties, and that such a determination can be known in principle by applying semantic norms to moral concepts" (Ibid., p. 132). The authors discuss some of the main concerns regarding so-called analytical moral functionalism; specifically, how moral claims can be decided by conceptual analysis and how immoral actions are merely due to conceptual confusion (Ibid., p. 134). In this context, the authors outline the fact that moral mistakes are not necessarily a result from analytical errors.

[13] Lara and Deckers make the important specification that the process of moral maturity is not equivalent to that of machine's moral self-update. They compare and contrast the moral development of children and potential moral machines. Lara and Deckers outline that many children start to develop their ethical views in time even when they contradict to the views of their parents. In contrast, machines cannot program themselves in moral terms (Lara and Deckers, 2019, p. 4).

[14] The discussion follows Philip J. Nickel's line of arguments regarding so-called trustworthiness ethic (Nickel, 2013).

References

Anderson M, Anderson SL (2007) Machine ethics: creating an ethical intelligent agent. AI Magazine 28(4):15–26

Anderson SL (2008) Asimov's "three laws of robotics" and machine metaethics. AI & Society 22(4):477–493

Amoore L (2020) Cloud ethics: algorithms and the attributes of ourselves and others. Duke University Press Books

Beer D (2016) The social power of algorithms. Information, Commun Soc. https://doi.org/10.1080/1369118X.2016.1216147

Feldman J (2005) Condorcet et la mathématique sociale. Enthousiasmes et bémols. Mathématiques Sci. Hum. 172(4):7–41

Howard D, Muntean I (2017) Artificial moral cognition: moral functionalism and autonomous moral agency. In: Thomas M Powers (ed) Philosophy and computing: essays in epistemology, philosophy of mind, logic, and ethics. Springer, pp 121–159

Kitchin R (2017) Thinking critically about and researching algorithms. Inform Commun Soc 20(1):14–29

Klincewicz M (2016) Artificial intelligence as a means to moral enhancement. Stud Log Gramm Rhetor 48(61):171–187

Klincewicz M (2017) Challenges to engineering moral reasoners: time and context. In: Lin P, Abney K, Jenkins R (eds) Robot ethics 2.0: from autonomous cars to artificial intelligence. Oxford University Press, Oxford, pp. 244–257

Lara F, Deckers J (2019) Artificial intelligence as a Socratic assistant for moral enhancement, Neuroethics, https://doi.org/10.1007/s12152-019-09401-y

Markham AN, Tiidenberg K, Herman A (2018) Ethics as methods: doing ethics in the era of big data research–Introduction. Soc Media+Soc. July–September: 1–9. https://journals.sagepub.com/doi/full/10.1177/2056305118784502

Metcalf J, Keller EF, Boyd D (2016) Perspectives on big data, ethics, and society. The Council for Big Data, Ethics, and Society http://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/

Mittelstadt BD, Allo P, Taddeo M et al. (2016) The ethics of algorithms: mapping the debate. Big Data Soc, 3(2):1–21

Moor JH (2006) The nature, importance, and difficulty of machine ethics. IEEE Intell Syst 21(4):18–21

Moor JH (2009) Four kinds of ethical robots. Philosophy Now. A magazine of ideas 72 https://philosophynow.org/issues/72/Four_Kinds_of_Ethical_Robots

Neff G (2020) From bad users and failed uses to responsible technologies: A call to expand the AI ethics toolkit. Proceedings of the AAAI/ACM conference on AI, Ethics and Society, 5–6, https://doi.org/10.1145/3375627.3377141

Nickel PJ (2013) Trust in technological systems. In: de Vries MJ, Hansson SO, Meijers AWM (eds) Norms in technology: philosophy of engineering and technology. Springer, Dordrecht, pp. 223–237

Powers TM (2006) Prospects for a Kantian machine. Intell Syst, IEEE 21(4):46–51

Savulescu J, Maslen H (2015) Moral enhancement and artificial intelligence. Moral AI? In: Romportl J, Zackova E, Kelemen J (eds) Beyond artificial intelligence. The disappearing human—machine divide. Springer, pp. 79–95

Stahl BC (2004) Information, ethics, and computers: the problem of autonomous moral agents. Minds Machines 14:67–83

Wallach W, Allen C (2010) Moral machines: teaching robots right from wrong. Oxford University Press, Oxford

Acknowledgements

I would like to thank Andrea Saltelli for his valuable comments on this article.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.S.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.