AI and Ethics (2023) 3:337–344
https://doi.org/10.1007/s43681-022-00160-w

COMMENTARY

AI, alignment, and the categorical imperative

Fritz J. McDonald[1]

Received: 3 March 2022 / Accepted: 6 April 2022 / Published online: 25 April 2022

Abstract
Tae Wan Kim, John Hooker, and Thomas Donaldson make an attempt, in recent articles, to solve the alignment problem. As they define the alignment problem, it is the issue of how to give AI systems moral intelligence. They contend that one might program machines with a version of Kantian ethics cast in deontic modal logic. On their view, machines can be aligned with human values if such machines obey principles of universalization and autonomy, as well as a deontic utilitarian principle. Programming machines to do so might be useful, in their view, for applications such as future autonomous vehicles. Their proposal draws both on traditional logic-based and contemporary connectionist approaches, to fuse factual information with normative principles. I will argue that this approach makes demands of machines that go beyond what is currently feasible, and may extend past the limits of the possible for AI. I also argue that a deontological ethics for machines should place greater stress on the formula of humanity of the Kantian categorical imperative. On this principle, one ought never treat a person as a mere means. Recognition of what makes a person a person requires ethical insight. Similar insight is needed to tell treatment as a means from treatment as a mere means. The resources in Kim, Hooker, and Donaldson's approach is insufficient for this reason. Hesitation regarding deployment of autonomous machines is warranted in light of these alignment concerns.

Keywords Artificial intelligence · Ethics · Alignment problem · Immanuel Kant

Tae Wan Kim, John Hooker, and Thomas Donaldson make an attempt, in recent articles, to solve the alignment problem. As they define the alignment problem, it is the issue of how to give AI systems moral intelligence. They contend that one might program machines with a version of Kantian ethics cast in deontic modal logic. On their view, machines can be aligned with human values if such machines obey principles of universalization and autonomy, as well as a deontic utilitarian principle. Programming machines to do so might be useful, in their view, for applications such as future autonomous vehicles. Their proposal draws both on traditional logic-based and contemporary connectionist approaches, to fuse factual information with normative principles. I will argue that this approach makes demands of machines that go beyond what is currently feasible, and may extend past the limits of the possible for AI. I also argue that any deontological ethics for machines should place greater stress on the formula of humanity of the Kantian categorical imperative. On this principle, one ought never treat a person as a mere means. Recognition of what makes a person a person requires ethical insight. Similar insight is needed to tell treatment as a means from treatment as a mere means. The resources in Kim, Hooker, and Donaldson's approach are insufficient for this reason. Hesitation regarding deployment of autonomous machines is warranted in light of these alignment concerns.

Kim, Hooker, and Donaldson, as mentioned above, aim to solve the alignment problem. This is defined by the authors how to give artificial intelligence systems moral intelligence. This is often called "machine ethics" in the philosophical literature. Kim, Hooker, and Donaldson try to show in a kind of broad outline how one might program machines with a version of Kantian ethics cast in deontic modal logic. On their view, machines can be aligned with human values if such machines obey principles of universalization and autonomy, as well as a deontic utilitarian principle [1, 2].

Investigating this proposal is worthwhile not only for the sake of critically considering Kim, Hooker, and Donaldson's ideas. The authors draw on some of the most prominent views in contemporary philosophical moral theory. They also appeal to state of the art approaches in artificial

✉ Fritz J. McDonald
fritzjmcdonald@oakland.edu

[1] Department of Philosophy, Oakland University, Rochester, MI, USA

intelligence. Were their project to succeed, it would be a fusion of AI and ethics with great appeal. However, an investigation of the project reveals some pitfalls for future attempts at machine ethics to avoid. There are serious issues with their appeal to AI resources and in their formulation of a kind of ethics that might fit the AI resources on offer. Paying attention to these issues might give someone pause in the project of machine ethics, the project of building AI systems meant to follow a sort of moral code.

Kim, Hooker, and Donaldson's proposal draws both on traditional logic-based and contemporary connectionist approaches, to fuse factual information with normative principles. Combining these approaches to AI is not, in itself, novel. As Kim, Hooker, and Donaldson [2] acknowledge, hybrid approaches to AI, dating back to the work of Smolensky [3], provide accounts that fuse connectionist and logic-based AI. Some contemporary hybrid models of artificial intelligence appeal to connectionist and logic-based systems to perform different tasks. For instance, a system might use elements of logic-based AI to carry out tasks that require representation of precise symbols, and elements of connectionist AI to carry out tasks where the ability to learn or demonstrate the sort of "graceful degradation" characteristic of connectionist systems is required.[1]

What is novel in the work of Kim, Hooker, and Donaldson is their approach to facts and values. Using a variety of AI resources, Kim, Hooker, and Donaldson aim to show how an AI system might relate factual information to normative values. In so doing, they stress that they want to avoid the well-known mistake of trying to infer an "ought" from an "is." This would be an illicit move of deriving normative premises, premises about what should be the case, from purely descriptive premises regarding the way the world is. They are critical of researchers in machine ethics who seek to derive ethical principles from observation of human behaviors and judgments.[2] In their view, such an approach would lead to machines having an ethics that reflects the biases and ethical misjudgments both of ordinary individuals and supposed moral experts.

To give an ethical framework not derived from observation of human behaviors and judgments, they propose, as mentioned above, to combine classical artificial intelligence approaches, that stress reasoning that resembles logical proofs, with more contemporary connectionist approaches. On a connectionist approach, sometimes called a "neural network" and often in the most sophisticated recent versions, "deep learning," computing is not done generally in a fashion that resembles logical proof. Instead, a number of separate units of varying mathematical complexity operate in parallel, to carry out intelligent tasks. Such systems have shown success in industrial applications such as object recognition and facial recognition.[3]

Kim, Hooker, and Donaldson suggest an ethically aligned AI system can use connectionist resources to gain factual information regarding the world. In turn, this factual information is then assessed using moral principles spelled out in modal logic. Their approach is a potential account of how Kantian and utilitarian ethics might possibly be programmed into an autonomous AI system. It attempts a kind of possibility proof of ethically aligned AI.

It is worth noting that their papers are largely theoretical, putting forward three principles in modal logic, and then showing how those modal logic principles might be applied to specific cases. This is not work that shows in full, exhaustive detail how to engineer a system that carries out the thinking involved in applying moral principles to factual information. For the engineering to take place, a lot more work would be required than is done in the Kim, Hooker, and Donaldson papers, but I don't think their work is meant to be a fully engineered plan for moral machines. Instead, they propose some principles in logical form, principles that may go some way to showing how machines may be moral. Their work is an effort to combine elements of the some of the most prominent ethical theories on offer in the philosophical literature, utilitarianism and Kantianism, with elements of two of the main strands of contemporary AI research. Were this to succeed, it would be a blueprint for the creation of autonomous systems that can be trusted to follow sound morals. This might be useful in a number of industrial applications, from robotics to self-driving cars. Still, it is at best questionable, as I will argue below, that AI systems are capable of carrying out the tasks at hand here.

As noted above, there are three main principles in their system. These moral principles are all deontic principles, although one of them is a deontic version of utilitarianism.

The first principle is a Generalization Principle. It is based on the Kantian Formula of the Universal Law. It

[1] For further detailed discussion of hybrid systems, see Sun [4, pp. 119–124].

[2] In particular, Kim, Hooker, and Donaldson criticize what they call a "a bottom-up approach in the form of inverse reinforcement learning, which allows a machine to internalize a pattern of preferences by observing how humans actually behave" [2, p. 2]. They attribute this view to Ng and Russell [5].

[3] Kim, Hooker, and Donaldson take for granted that connectionist AI systems can deliver factual information suitable for making moral judgments. There are some reasons to be concerned at least about the potential failures of the current connectionist systems on offer. It is well-known that many connectionist AI systems are "brittle." What this means is that such systems, while at times good at characterizing objects in the environment, can also at time fail in pretty catastrophic ways when given certain kinds of inputs. For a number of interesting examples, see Heaven [6]. I am grateful for an anonymous reviewer for *AI and Ethics* for raising this concern.

states, in plain English, that an action is morally permissible if it is rational to believe that it is possible for one to perform an action, based on one's own action plan, in a world where everyone has the same action plan. Any action that fails this test is immoral.

The second principle is a Utilitarian Principle: This can be stated simply as: an action is permitted if it is rational to believe that an action promotes at least as much utility as any other action one could perform in the circumstances. Once again, actions that fail this test are immoral.

The third principle is the Autonomy Principle. An action is permitted if it is rational to believe that my action plans are consistent with the action plans of any other agent. Failing this test means an action is immoral.

They are all used to place constraints on behavior. Kim, Hooker, and Donaldson formulate all of these principles in a modal logic, where the modality in question is what can or cannot be believed rationally. The modal operator Diamond is understood as what is possible to rationally believe. The operator Box represents what is rationally required to believe. These principles are:

First, the Generalization Principle: In plain English, an action is permitted if it fits the following principle: it is rational to believe that it is possible for one to perform an action, based on one's own action plan, in a world where everyone has the same action plan. Any action that does not fit this Generalization Principle is immoral.

The formal version of the principle is:

$$\Diamond_a P\,[\forall x(C(x) \Rightarrow_x A(x)) \wedge C(a) \wedge A(a)].$$

An action plan, for Kim, Hooker, and Donaldson, is a relation between certain conditions that the agent considers justifications for an action, and the action itself. In their formal presentation, ⇒ is not a logical entailment, but instead is meant to represent that the agent takes the conditions in question to justify taking a course of action. Potentially justifying conditions are represented as C(x), where x is the agent who takes these conditions to be justifying, and A(x) represents the action that the agent would take on the basis of these conditions. The subscript after the arrow ⇒ is meant to indicate the agent who takes these conditions as justifications for the action.

Here, to reiterate, the diamond is construed as what is possible to rationally believe. Relation P(S) means that, for the given proposition S, that it is possible S is true. The letter a represents the agent in question, both in the subscript following the diamond and in "C(a)," the conditions and "A(a)," the action.

Reading from the left, this should be understood as: it is rational for agent a to believe it is possible for the following proposition to be true: Every agent takes the conditions C to be a justification for acting in way A, and agent a takes conditions C to be reasons to A, and agent a acts in way A.

If this test fails, an action should be considered morally wrong.

It is worth noting here that this is a sophisticated logical principle. The principle involves judgment regarding what is rational to believe. For an AI system to be able to assess what is rational to believe is a significant demand. To make judgments regarding what is rational to believe would mean that an AI system has some conception of belief and rationality. These are the kinds of notions on which even the most sophisticated AI systems seem to falter. To judge what is rational to believe would involve the kind of general intelligence that most think AI systems lack currently. Perhaps a future AI system could make judgments regarding rationality, belief, and their relation, but that would be speculative at the present.

Even more than this, for an AI system to be able to make the judgments Kim, Hooker, and Donaldson have in mind, that system would have to understand what kind of conditions are those that an agent takes to justify a given course of action. Once again, this is a significant demand, and we currently do not have any machines that can do this. To crack this nut and get an AI system to make judgments of this kind of a big problem. It would require an AI system to have an understanding of agents and their psychology. Here, Kim, Hooker, and Donaldson are helping themselves to quite a bit of sophisticated understanding, on the part of machines, to spell out what these systems would do.

Their second principle is the Utilitarian Principle: This can be stated simply as: it is rational to believe that an action promotes at least as much utility as any other action one could perform in the circumstances. Failing this test of rational belief renders an action immoral.

Hooker and Kim spell this out formally as:

$$\Diamond\forall A'\left[u(A(a),\ C(a)) \geq u\left(A'(a),\ C(a)\right)\right].$$

Here, again, Diamond represents what is rational to believe. A(x) again represents actions taken by an agent x. This principle is formulated using second-order logic, where ∀A′ ranges over predicates A(x) for agents' actions. u(A′(a), C(a)) is a utility function that measures the expected utility of the action, given the conditions C.

So, this amounts to the following: it is rational to believe that for all actions, that the utility of the action taken by agent a in conditions C is greater than or equal to all other actions taken in these conditions. Once again, in their view, failure to meet this test would mean that action A is wrong for the agent to perform.

This is lacking in a significant amount of detail. The authors do not explain what they mean by utility, a notion that has been defined in many different ways in the

philosophical literature. How exactly would an AI system measure utility, however it's cashed out, is far from clear. Also, as in the Generalization Principle, rational belief enters into the picture.

Their third principle is the Autonomy Principle: It is rational to believe that my action plans are consistent with the action plans of any other agent. Failure for an action plan to be consistent with the action plan of others means that the plan is immoral. The following is a test for whether agent a's action plan is consistent with agent b's action plan:

$$\diamond_a P(A(a) \wedge A'(b)) \vee \neg\Box_a P(C(a) \wedge C'(b))$$

Here, there are two disjuncts, each representing a case where the action plans are consistent. The left-hand disjunct states that the potential actions taken by agents a and b are consistent. If this is rational to believe, then an action is not wrong. The right-hand side of the disjunct states that it is not necessary to believe that the reasons, the conditions given for taking an action, are the same. This right-hand side disjunct allows for actions to be allowed so long as the reasons given for those actions is different.

In case where it is not rational to believe one or a combination of these, an action is ruled out as morally wrong.

Kim, Hooker, and Donaldson relate the Generalization Principle to Immanuel Kant's Formula of the Universal Law. The Autonomy Principle, they suggest, reflects the Principle of Humanity. The Utilitarian Principle is an additional constraint that they suggest can be part of an "ecumenical" Kantianism, drawing on the work of Derek Parfit [7] and David Cummiskey [8].

I will review how Kim, Hooker, and Donaldson illustrate the application of these principles to cases.

For the Generalization Principle, Hooker and Kim discuss a person who wants to steal a watch. He has an action plan of snatching the watch from the store. Yet in a world where everyone had the same action plan, and stealing was widespread, security would be so increased that watch stealing would be, if not impossible, very highly unlikely. Therefore his thievery is not possible in a world where everyone has the same action plan, yielding the result that it is wrong to steal.

To relate this to the formal version of the principle:

$\diamond_a P\ [\forall x(C(x \Rightarrow_x A(x)) \wedge C(a) \wedge A(a)]$.

In such a case, we ask whether it is rational to believe that the following proposition is possible: all agents would steal a watch in the same conditions, and agent a steals a watch in these conditions. Hooker and Kim contend this is not rational to believe.

It is worth pausing here to note that this application of the principle involves a kind of thinking, on the part of the hypothetical moral machine, that seems to go beyond straightforward application of principles to factual information. It seems to involve a kind of imagination, an ability to conceive a certain world where everyone acts on the same plan. This goes far beyond the kind of pattern recognition that contemporary neural networks are good at. Very significant advances in AI might be required for this kind of thinking.

The Autonomy Principle is more straightforward. The Autonomy Principle is applied to the following case: "Suppose, for example, that you decide to cross the street to catch a bus as soon as no cars are coming. You begin to cross, but I grab you by the arm and pull you off the street' [1]. Hooker and Kim take this to be a violation of autonomy. It is illustrated by their principle, given that the plan of the person who pulls the individual's arm is not consistent with the plan of the individual who aims to cross the street.

Again, their autonomy principle is:

$$\diamond_a P(A(a) \wedge A'(b)) \vee \neg\Box_a P(C(a) \wedge C'(b))$$

In the case where the person pulls an individual off the street, the left-hand side of the disjunct is ruled out by the fact that both of the actions of these agents cannot take place together. Agent a cannot cross the street while agent b pulls a off the street. The conditions are the same, so the right hand disjunct turns out false as well.

Their version of a utilitarian principle is illustrated with a driving example, for an autonomous vehicle. They consider a society where veering into traffic to enter a lane is generally considered unacceptable [2]. In such a world, utility would not be maximized by a vehicle that cuts into traffic. A vehicle that does so would promote less utility than the maximal case, where the vehicle waits until a gap in the traffic is available for lane entry.

Kim, Hooker, and Donaldson, as mentioned above, do not go into great detail about how they define "utility." This would need to be further spelled out to clarify their account. It is also worth considering whether a neural network can be built that can calculate utility, however defined, of hypothetical scenarios.

This is a series of applications of clearly formulated principles. It is an attempt to show how one might potentially create an autonomous system that performs a sophisticated task: applying general principles of rightful action to specific examples. While these applications of the Generalization, Autonomy, and Utilitarian principles are clear, there are several concerns with these principles, as formulated.

They may be too restrictive. Consider the Autonomy Principle. When I want to do something, it is too much to

demand that my plan of action cannot contradict the plan of action of another. Sometimes I act, and hence thwart another's plan, but I do not do anything wrong. For instance, imagine that there is one slice of brie left on a plate at a wine and cheese party. Agent a and agent b are both in the same circumstances, yet they want to commit inconsistent actions. Agent a wants the brie for himself, and agent b wants the brie for himself. If agent a reaches the table first, and takes the first piece of brie, he hasn't done anything wrong. Their plans are at odds, but not in any way that violates an ethical demand. So it would seem like Kim, Hooker, and Donaldson's Autonomy Principle is too restrictive.

The utility principle is highly demanding [9]. It may suggest that anything short of devotion to the most saintly moral life is a failure [10]. Some of what is demanded by the utilitarian principle may be merely supererogatory, rather than obligatory [11]. It might always be good to do the most good, but it's a lot to ask every agent to perform only the course of action that is maximal with respect to utility. When it comes to self-driving vehicles or other engineering applications of these principles, one might wonder why one would want a car that maximizes utility, or a robot that maximizes utility. The course of action that promotes the greatest utility might differ from the one that would be preferred by the user. If I tell my self-driving car to take me to the store to buy brie, it would not be helpful for the car to indicate that I could do other things that would be more beneficial to society.

These principles are also perhaps in need of supplementation to do the required philosophical work. The universal law principle might not be enough to carry the weight of formulating an ethical framework. Dating back to Hegel there is the concern that this formulation of the categorical imperative is an "empty formalism" [12].

As Thomas Powers [13] points out, drawing on Silber [14], Rawls [15], and O'Neill [16], Kant himself seems to supplement his applications of the universal law principle with some further principles of moral reasoning. Powers suggests that supplemental principles might be required for a plausible Kantian ethics. If so, there is an issue here, for such supplemental principles are lacking in Kim, Hooker, and Donaldson's proposed version of the categorical imperative.

One example of a supplemental principle used by Kant occurs in his discussion of suicide. Kant contends that suicide done for the sake of one's own good, or as Kant puts it, for "self-love" violates the categorical imperative. He supports this point by availing himself of another claim: the purpose of self-love is to further life. However, raising potential concerns for accounts like the one given by Kim, Hooker, and Donaldson, Powers argues persuasively that these supplemental principles, as formulated, might not fit well into a machine ethics framework, for these principles hold generally but allow for exceptions. Suicide for the sake of self-love might be wrong, but suicide for other reasons, like saving the lives of people one loves, might be allowed, even by Kant. It is not clear how to formulate these types of principles allowing exceptions in a way that is computable by machines [13]. The sort of thinking involved requires non-monotonic inferences, Powers suggests. What this means is that counterexamples to these principles do not prove them wrong, given the exceptions. They do not follow strict logical rules of the sort where exceptions prove principles wrong.

In any case, resources for supplementing the Generalization principle with further principles of ethical reasoning are not on offer in Kim, Hooker, and Donaldson's work, and so the concern that the Generalization principle does not rule out enough remains.

It is doubtful whether the Generalization Principle, as formulated by Kim, Hooker, and Donaldson works for some Kant's own examples—is there really a practical problem, for instance, with universalizing a maxim, or plan, to ignore others in need? As awful as a society of selfish egoists might be, it's not so clear that it's impossible. Even Kant himself [17] grants such a world is possible. His concern isn't whether we could conceive of such a world, and even live in such a world, but whether we could rationally will to live in such world. O'Neill calls these cases where there is inconsistency in what we can rationally will "volitional inconsistencies" [16]. These stand in contrast to the "conceptual inconsistencies" we find in cases where we cannot conceive of acting in a certain way world consistent with our maxims for action. The case of stealing a watch given by Kim, Hooker, and Donaldson is a "conceptual inconsistency" case, for the world where the maxim of action is universalized is inconsistent with the possibility of the action taking place. Kim, Hooker, and Donaldson do not avail themselves of all the rich resources of Kantian ethics, for they only test for conceptual inconsistencies, not volitional inconsistencies.

The autonomy principle, as formulated by the authors, places a great deal of stress on the formulation of individual action plans. Actions are ruled out when one individual's choices are not consistent with the action plans of others. Still, there is room here for concern. One might plan to do something to oneself that at least seems wrong, or unjust. If a person plans to be sold into slavery, does that make it right? It's at least possible to plan to treat oneself in ways that are wrong, and this sort of planning likely has been actual many times over. Thus autonomy, cast in these ways, is worrisomely lacking in ways parallel to the Generalization Principle.

The last of the three principles, the Utilitarian Principle, does place quite strong demands on agents, whether natural or artificial. It suggests one should not take a course of action that would result in a lower amount of overall utility

than another course of action. It is well-known that utilitarianism faces a number of objections. As a general matter, utilitarianism would seem to allow for the mistreatment of a small number of individuals, or one individual, for the greater good of the many [18]. Perhaps we can rule out some of these cases of mistreatment by appeal to the Autonomy or Generalization Principles. Some utility-maximizing mistreatment might not be agreed to, or might not be conceivable as the behavior of all.

In the example of an unfair contract, we have a case that demonstrates that these principles, taken together, can be off the mark. A person might agree, in their plans, to immoral mistreatment of himself or herself. It is far from clear that the Generalization Principle is enough to rule this out, as it is formulated by the authors. In Kant's sense, it might be irrational to will a general situation where such treatment exists. Still, though, a world of poorly treated people is conceivable, and only what cannot be conceived given one's plans is ruled out by Kim, Hooker, and Donaldson's version of the Generalization Principle. Agreeing to do this as part of one's own plans is a given in this example, so the Autonomy Principle does not help rule out this case.

The Utilitarian Principle is clearly not enough as well. For if the unfair contract is one that maximizes overall utility, it not only is permitted, it is positively the right thing to do. So long as the greatest amount of overall utility is produced, the right course of action has been taken.

These are general moral considerations against considering the Autonomy, Generalization, and Utilitarian principles, taken jointly together, as correct constraints for ethical action. At this point, it's worth considering some examples that relate most directly to the topic at hand of Kim, Hooker, and Donaldson's work: autonomous systems such as self-driving vehicles. I will give two examples of cases that fit these principles yet do not clearly involve right action on the part of an autonomous system. I will call these cases "Lock" and "Veer."

In the case of Lock, consider a hypothetical set of terms and conditions for entering into a contract to buy an autonomous vehicle. These terms and conditions state that the vehicle will be locked unless the driver is engaging in activity that most benefits the greatest number of people. The driver agrees to this arrangement, thus Autonomy, as formulated by Kim, Hooker, and Donaldson, is assured. There is nothing inconceivable about a world where people agree to such terms and conditions, in accord with Generalization. Maximizing utility is certainly assured by this agreement. Yet when drivers find themselves locked in their vehicles, prevented from exiting unless their future behavior will be utility-maximizing, it would seem that their autonomy is being violated. To have agreed to such terms and conditions would be morally wrong, for it would require one's future self to be treated just as a means to maximize utility.

Turning to the Veer case, let's imagine a future in which everyone drives an autonomous vehicle. Part of the agreement that all drivers agree to is a policy for the sake of self-protection. To avoid death on the part of drivers and passengers, all of society agrees that, when a crash seems inevitable, it is permissible to veer towards a pedestrian, or a smaller group of pedestrians than the number of individuals in the vehicle, to save the lives of those in the vehicle. This is both possible as a general practice, utility-maximizing, and, given that everyone agrees to the terms and conditions, it is in line with the Autonomy Principle. In this case, though, the agreement everyone enters into is still wrong, because innocent bystanders are treated just as a means to the saving of the lives of the drivers and passengers.[4] That the bystanders killed by the vehicles agreed to the same terms when they bought their own vehicles may be a failing of some kind on their part, but it does not excuse the wrongfulness of sacrificing their lives for the preservation of others.

These issues can be avoided through more consideration of the Formula of Humanity. On this principle, one ought never treat a person, whether oneself or another, as a mere means. Kant illustrates the Formula of Humanity with a case involving the making of a false promise. A person is in need of some assistance. That person considers making a false promise to get that help from another. Perhaps in such a case utility might even be maximized if the person makes the false promise to another. Yet Kant contends that this is still wrong. In his discussion of this case in light of the Formula of Humanity, he contends that making this false promise treats the other person as a mere means, not as a person. The individual makes the false promise to use the other to get what they want. The person offering assistance is just treated as a means to an end, the end being assistance.

The Formula of Humanity would rule out the cases that are allowed by the principles on offer in Kim, Hooker, and Donaldson's framework. Making a plan to be mistreated is a violation of a duty to not treat oneself as a mere means. Courses or action that might receive overall agreement, be universalizable, and promote utility, may yet involve the wrongful treatment of oneself or others as mere means. The kinds of mistreatments of the few for the sake of the many can be seen either as treatment of the few as mere means, or failure to recognize the value of individual persons as ends.

Note as well that the Formula of Humanity provides both a positive and a negative element to morality. It is forbidden, Kant says, to treat humanity (either another's or one's own) as a mere means, but we also must treat others as ends. This creates a positive obligation, to further the ends of humanity insofar as one can. This might go some way to meeting a challenge put forward by Powers [13], a challenge that

[4] This case is inspired by John Harris's "The Survival Lottery" [19].

Kim, Hooker, and Donaldson might have some issues with. Kim, Hooker, and Donaldson formulate morality entirely in terms of constraints, in other words, what is morally forbidden. Powers points out that the Formula of the Universal Law, understood as a constraint, only tells us what is morally forbidden, not what is obligatory or permissible. This is a lacuna in the theory of morality on offer both in Kant's Formula of the Universal Law and in a constraint-based account like Kim, Hooker, and Donaldson. One would like to be able to create machines with a grasp not only on what is forbidden, but also what is obligatory or permissible. At least in terms of obligation, the Formula of Humanity is a way forward to develop positive obligations.

It is not obvious, however, how to incorporate the Formula of Humanity, understood in this way, into the framework on offer in the AI alignment literature. Kim, Hooker, and Donaldson show how some of ethics can be construed in a normative and descriptive framework. Classification is done by neural networks, and principles cast in deontic logic are applied. However, how could a computer employ the normative notions of a person, or of treatment as a mere means, or the value of persons as ends? Recognition of the distinction between persons and things, in a moral sense, requires ethical insight. Similar insight is needed to tell treatment as a means from treatment as a mere means. This kind of insight is normative, nuanced, and not obvious based on the descriptive facts of the world. It is doubtful that the sort of neural networks Kim, Hooker, and Donaldson think might be involved in providing factual data could be enough to make such classifications and decisions.

A kind of perfected object recognition software might be able to perfectly capture the difference between humans and nonhumans. Yet this is not the same as capturing the notion of a person employed in moral philosophy. To be a person, to be a moral patient or a moral agent, is a debated idea that philosophers have defined in many, inconsistent ways. It might be a purely normative notion, or something we have not quite yet defined to everyone's satisfaction.

Kant himself characterizes what is special about persons in terms of autonomy, the ability to govern oneself by moral rules. One influential contemporary Kantian, Christine Korsgaard, characterizes what is unique about persons in terms of the ability to use principles such as the categorical and hypothetical imperative to constitute oneself as a unified self [20]. Compatibilist philosophers place less stress than Kant himself on freedom and autonomy, if freedom and autonomy are construed in a kind of libertarian fashion as independence from the causal laws of nature. Instead, they contend freedom and personhood can be consistent with causal determinism. For instance, Harry Frankfurt [21] develops a view on which being a person is a matter of having certain higher-order states of desire, desires regarding what one desires to do. When these are effective, they demonstrate a sort of freedom that is consistent, in Frankfurt's view, with causal determinism. R. Jay Wallace [22] criticizes Frankfurt's position, arguing that instead what is truly distinct about persons is their ability to respond to reasons.

The boundaries of who is and is not a person are also contested. For instance, Peter Singer [23] contends that we should extend the notion of personhood to nonhuman animals, but some human beings should not be considered persons.

How to define "person" is not obvious.

What matters for the current discussion is how to create a machine that can both recognize and value persons. Until machines can be built that recognize the value of persons, that know what kind of treatment or mistreatment of persons is treatment as an end or a mere means, there are serious reasons to be concerned regarding the deployment of autonomous systems.

One of the reasons the framework offered by Kim, Hooker, and Donaldson is perhaps not best to capture these important norms is that they put forward a very sharp fact/value distinction. There is purely descriptive information, on the one hand, gathered by neural networks. There are purely normative principles applied to these. When it comes to Kantian ideas like being a person, what Kant means by "humanity," and treating a person as a means or an end in themselves, can we really separate the normative from the descriptive so cleanly?

While Kim, Hooker, and Donaldson offer a novel approach to alignment, making a moral machine remains a serious challenge. I would suggest that doing so would require a more general kind of intelligence on the part of machines than is currently on offer now. It would also require a great advance in moral philosophy by humans themselves, to give the right kind of guidance to machines.

Declarations

Conflict of interest None.

References

1. Hooker, J., Kim, T. W.: Toward non-intuition-based machine and artificial intelligence ethics: a deontological approach based on modal logic. In: Proceedings of the AAAI/ACM Joint Conference on Artificial Intelligence, Ethics and Society (2018)
2. Kim, T.W., Hooker, J., Donaldson, T.: Taking principles seriously: a hybrid approach to value alignment. J. Artif. Intell. Res. **70**, 871–890 (2021)
3. Smolensky, P.: On the proper treatment of connectionism. Behav. Brain Sci. **11**, 1–23 (1988)

4. Sun, R.: Connectionism and neural networks. In: Frankish, K., Ramsey, W.M. (eds.) The Cambridge Handbook of Artificial Intelligence. Cambridge University Press, Cambridge, UK (2014)
5. Ng, A. Y., Russell, S. J.: Algorithms for inverse reinforcement learning. In Proceedings 17th International Conference on Machine Learning, pp. 663–670 (2000)
6. Heaven, D.: Deep trouble for deep learning. Nature **574**, 163–166 (2019)
7. Parfit, D.: On What Matters. Oxford University Press, Oxford (2011)
8. Cummiskey, D.: Kantian Consequentialism. Oxford University Press, New York (1996)
9. Williams, B.: A critique of utilitarianism. In: Smart, J.J.C., Williams, B. (eds.) Utilitarianism: For and Against. Cambridge University Press, Cambridge (1973)
10. Wolf, S.: Moral saints. J. Philos. **89**, 419–439 (1982)
11. Thomson, J.J.: A defense of abortion. Philos. Public Aff. **1**, 47–66 (1971)
12. Hegel, G.W.F.: Elements of the Philosophy of Right. Cambridge University Press, Cambridge (1991)
13. Powers, T. M. (2006) Prospects for a Kantian Machine. IEEE Intelligent Systems 21(4):46–51
14. Silber, J.: Procedural formalism in Kant's Ethics. Rev. Metaphys. **28**, 197–236 (1974)
15. Rawls, J.: Kantian constructivism in moral theory. J. Philos. **77**, 515–572 (1980)
16. O'Neill, O.: Constructions of Reason. Cambridge University Press, Cambridge (1989)
17. Kant, I.: Practical Philosophy. Cambridge University Press, Cambridge (1996)
18. Rawls, J.: A Theory of Justice. Harvard University Press, Cambridge, MA (1999)
19. Harris, J.: The survival lottery. Philosophy **50**, 81–87 (1975)
20. Korsgaard, C.: Self-Constitution: Agency, Identity, and Integrity. Oxford University Press, Oxford (2009)
21. Frankfurt, H.: Freedom of the will and the concept of a person. J. Philos. **68**, 5–20 (1971)
22. Wallace, R.J.: Responsibility and the Moral Sentiments. Harvard University Press, Cambridge, MA (1994)
23. Singer, P.: Practical Ethics, 3rd edn. Cambridge University Press, Cambridge, UK (2011)