AI & SOCIETY
https://doi.org/10.1007/s00146-021-01229-6

OPEN FORUM

What overarching ethical principle should a superintelligent AI follow?

Atle Ottesen Søvik[1]

Received: 3 July 2020 / Accepted: 29 April 2021

Abstract
What is the best overarching ethical principle to give a possible future superintelligent machine, given that we do not know what the best ethics are today or in the future? Eliezer Yudkowsky has suggested that a superintelligent AI should have as its goal to carry out the coherent extrapolated volition of humanity (CEV), the most coherent way of combining human goals. The article discusses some problems with this proposal and some alternatives suggested by Nick Bostrom. A slightly different proposal is then suggested, which I argue solves the problems better than Yudkowsky's proposal.

Keywords Superintelligent AI · Overarching ethical principle · Coherent extrapolated volition · Yudkowsky · Bostrom

1 Introduction

In 2014, Nick Bostrom wrote the widely read book *Superintelligence*, which discusses what may happen if superintelligent machines ever appear (Bostrom 2016). By "superintelligence," Bostrom means "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" (Bostrom 2016, p. 26). In other words, the term is defined *quantitatively* in the sense of being much more of the same as human intelligence.[1] Bostrom is open to the possibility that such intelligence may become *qualitatively* different from human intelligence in the same way as human intelligence seems to be qualitatively different from that of monkeys (Bostrom 2016, p. 69), but this would be a kind of intelligence that the humans of today cannot understand or describe.

The term "superintelligence" does not imply the ability to solve all problems. Discussions today in the AI community about what superintelligence would imply resemble discussions that have gone on for a couple thousand years in philosophy of religion: what are the implications and limitations of the omniscience and omnipotence of God? Most philosophers of religion accept that there are many things an omniscient and omnipotent God could not know or do (Søvik 2011).[2]

Bostrom's main concerns with superintelligence are that it implies that a superintelligent agent would have vastly more power than any human being or group/organization/state and that there are many reasons to think that this power would be used to exterminate all life (Bostrom 2016 pp. 140–154). There are many immoral humans, groups, organizations or states, but they are not able to do the kind of damage that a superintelligent agent could do. Since we have no reason to think that we could control the actions of a superintelligent agent, our best alternative is to try to make it want to do benevolent actions from the beginning (Bostrom, 2016 pp. 157, 226). Since it is impossible to give a superintelligent agent a recipe for what to do in every possible situation,

[1] Human and artificial intelligence is here defined as complex problem-solving that would have been called intelligent if performed by humans (Tegmark 2017).

[2] God could not know the result of God's own free actions (given a libertarian theory of free will and a presentist theory of time); God could not know that there is not something that God does not know; God could not have indexical knowledge of other people (knowing what it is like for God to have the indexical experience "I am Donald Trump"); and a lot of other limitations that follow from having great knowledge and power (God could not remember a forgotten joke, learn a new language, learn to ride a bicycle, torture someone for fun)(Phillips 2004;Martin & Monnier 2003). The halting problem in computational theory resembles the problem of incompleteness that Gödel demonstrated in mathematics, which again resembles the general insight in philosophy that we can always expand a theoretical framework with infinitely many extra truths (Puntel 2008, p. 117). This is just yet another example of the many things that even a super-

✉ Atle Ottesen Søvik
Atle.O.Sovik@mf.no

[1] MF Norwegian School of Theology, Religion and Society, Oslo, Norway

Published online: 27 May 2021

Bostrom argues that the best we can do is to give it some overarching ethical principle to follow (Bostrom 2016, p. 226).

Given this, we get the following problem, which is also the topic for this article: *What is the best overarching ethical principle to give a possible future superintelligent machine when we do not know what the best ethics are?* As Bostrom points out, no ethical theory is supported by a majority of moral philosophers, which seems to imply that most moral philosophers have not yet found the right answer to this question (Bostrom 2016, p. 257).[3] Given that we are likely to be wrong about what are the best ethics, the question can then be formulated as follows: How can you build an AI which is more ethical than yourself? (Bostrom and Yudkowsky 2014, p. 332) Given that we do not know what the best ethics are for humans and machines today and in the future, and given that the ethics of today are not the best possible ethics, what should direct us in further developing (machine) ethics?

To help our thinking, Bostrom suggests a thought experiment. What if Archimedes in ancient Greece had been able to build an intelligent AI? Given that the ethics of ancient Greece were not perfect, what general advice would you give Archimedes as a strategy for developing ethics if you could not give any specific moral norms (Bostrom and Yudkowsky 2014, pp. 331–332)? Is there a principle which would have been good advice to give Archimedes a couple thousand years ago and which seems to be good advice for us and for intelligent agents in the future?

It may seem incoherent to ask how to build an AI which is more ethical than yourself since it seems to require that we already have the right moral theory at hand in order to be able to consider the quality of possible answers. On the other hand, it also seems like a meaningful project that we try to improve our ethics since philosophers have always tried to make better theories of ethics and truth.

The first thing to note here is that while humans at all times have disagreed on concrete ethics, there is much more agreement on the more abstract level. We can agree in large part that ethics is about making the world good for everyone, but disagree strongly on what such a world looks like concretely. More precisely, the question should thus be how to get closer to the abstract ideal of morality when we disagree so much on the concrete content.

But one can object to this as well, and say that both humans in general and moral philosophers in particular disagree completely on even the basic abstract principles. Ryan Muldoon and Gerald Gaus are examples of authors who have argued that since people disagree radically on both morality and how they understand the world, we should not have an ideal ethical standard but instead negotiate step by step an agreement that all involved will see as improvements (Gaus 2016; Muldoon 2016).

My response to this is that if we say that people really disagree about something – instead of just talking about completely different things – it presupposes some shared understanding of what they are talking about. If they really disagree about what is morally good or morally better, it presupposes at least some coarse-grained understanding of what the concept "morally good" means (one cannot think that it means "green").

Note how Muldoon and Gaus also must presuppose such a moral standard in order to think that it is a morally good suggestion that we should try to have negotiations that all involved see as improvements. We cannot think that Muldoon's and Gaus 'proposals are good unless we presuppose that it is good that everybody find their situation improved. They, like me, presuppose a coarse-grained abstract standard of what "morally good" means, even if they, like me, are uncertain about what concrete ethical rules are right or what the morally best world would look like concretely. Derek Parfit has argued that the main moral traditions are different roads to the same mountaintop (Parfit 2011), which indicates that most people actually have quite similar understandings of what morality implies.

I do not think that there is an objectively correct definition of moral terms, but instead I think that we need to agree on definitions we choose.[4] Of course, it is contested whether that is true, and if one agrees that it is true, it will be contested which definitions to choose. But everything is contested, so we need to start with some hypotheses and do our best to justify them as true and as solving the problems we want to solve. This is what I do in this article. If someone objects that I am suggesting an AI which does not have the correct understanding of "morally good", I will say that I am satisfied with the project of finding an AI that can actualize the best way to the best world. In this article, I argue that this

Footnote 2 (continued)

intelligent agent could not know – as far as we can tell. There is still a lively debate today among people defending positions like open theism, divine middle knowledge and divine foreknowledge on what God can and cannot know.

[3] To say that this implies that most moral philosophers are wrong presumes that there is a *best* ethical theory, which seems plausible. Maybe there is *not* a best theory of ethics, in which case most moral philosophers are wrong about that fact. In any case, most moral philosophers give a wrong description of how best to think of ethics.

[4] There is not room for this fundamental metaethical discussion in this article. For a recent defence of this position, see Dasgupta (forthcoming). The Meta-Ethics of Artificial Intelligence: Are We Beholden to Normative Joints? (draft of November 2020). Retrieved from http://shamik.net/papers/dasgupta%20the%20metaethics%20of%20AI.pdf.

approach solves more of the problems we want solved than do the alternatives I discuss.

The goal is thus to find a very general ethical principle which may work well even if many of the concrete ethical norms that are commonly defended today may be wrong. In the next section, I shall look mainly at a suggestion by Eliezer Yudkowsky, together with some alternatives discussed by Bostrom, to see what the weaknesses of these suggestions are.[5] In the third section I shall suggest an alternative similar to Yudkowsky, but avoiding some of Yudkowsky's problems. In the fourth and fifth sections, I answer two sets of objections to the new principle suggested.

Note that I am not discussing what the conditions are for making a machine a moral agent which could rightly be held responsible for its actions. I am only discussing which ethical principle a superintelligent machine should follow in order to actualize a morally good world, even if the machine itself does not have consciousness or feelings or would qualify as a morally responsible agent.

2 Some suggested ethical principles for a superintelligent AI to follow

Eliezer Yudkowsky has suggested that a superintelligent AI should have as its goal to carry out the coherent extrapolated volition of humanity (CEV), or the most coherent way of combining human goals. Coherent extrapolated volition is "our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted." (Yudkowsky 2004, p. 6).

There are some advantages of this suggestion: For one, it avoids that a superintelligent machine could hijack humanity. If you give a concrete rule for the machine to follow, it can easily backfire through some unforeseen problem. For example, a superintelligent machine set to make everybody happy could manipulate all brains into a permanent state of feeling happiness. Since this presumably is not what all people want the most, such an outcome is avoided with Yudkowsky's proposal. A safety turn-off switch is thus built into the system, so if a superintelligent machine starts planning something humanity does not want, it aborts the mission (Yudkowsky 2004, pp. 15–18).

Further, it encapsulates moral growth in the future. The current moral views prevailing in the world are presumably not optimal, and new contexts may need better solutions than we have thought of so far. By extrapolating the most coherent volitions as these change over time, the superintelligent machine will presumably make more optimal choices than if merely based on concrete decisions of today (Yudkowsky 2004, p. 14).

Bostrom discusses this proposal by Yudkowsky. One possible objection is that even a superintelligent machine may not know what we desire or would desire, but Bostrom argues that at least it could make very well informed estimates by having massive information about humans and their choices (Bostrom 2016, p. 261).

However, Bostrom argues that some questions remain unanswered. Whose volitions should be included in the extrapolated volition? Should those of embryos, fetuses, severely brain-damaged persons, people with severe dementia, people in persistent vegetative states, people in the past, people in the future, animals (which?), digital minds, or extraterrestrials? (Bostrom 2016, p. 265).

Yudkowsky has a short discussion of this question, arguing that the starting point must be humans existing today, and their coherent volition must decide which individuals to include in the coherent volition (Yudkowsky 2004, pp. 23–25). This does not answer exactly who should be counted as included in the starting group, although it says that it is humans (as opposed to animals) and presumably those humans have a volition at all (which severely brain-damaged, etc., do not have). The proposal does not justify why this group should determine future ethics instead of future ethics being determined also by what is good for those who are then left out.

Another objection Bostrom raises is that there are many evil preferences among humans living today, which could imply that the most coherent extrapolation of their volitions would nevertheless not lead to the best ethical decision (Bostrom 2016, p. 263). Yudkowsky discusses this objection himself. Since many people have destructive desires, why assume that a coherent extrapolation of this will be better? (Yudkowsky 2004, p. 25) Yudkowsky admits that it is a real possibility that a coherent extrapolation of the collective will of humanity might not be good.[6] He also thinks that the problem is difficult to solve. If one starts adding extra conditions, the guiding rules become more complicated, which again makes it more probable that some error or unwanted result occurs (Yudkowsky 2004, p. 26).

In addition to the objections raised by Bostrom, I would like to point out some other problems. These problems have

[5] The main article by Yudkowsky is not peer-reviewed, but the Yudkowsky-Bostrom debate is chosen since it is the most detailed debate on the topic that this author is aware of.

[6] Yudkowsky's proposal resembles Kant's Kingdom of Ends formula but is nevertheless clearly different. Kant presupposes rational agents while Yudkowsky bases his proposal on actual volitions, making him more of a preference utilitarian than a Kantian.

to do with how to determine how to extrapolate volitions. Coherent extrapolated volitions are what we would want if we knew more, were more the people we wished we were, etc. (Yudkowsky 2004, pp. 4–9). But our hypothetical wishes are very dependent on context, so what decides that? You may answer that coherent volitions should determine context as well, but it seems that humanity could change in many different ways and coherently wish many different things. People's desires (including who they wish to be) can change in different ways in different contexts; people could be happy overall with the different alternatives (having different desires in different contexts), so it would be preferable to have a clearer criterion for measuring in what direction volitions should be extrapolated. This point is thus related to the objection above that it seems there could be a coherent evil extrapolation of wishes.

A possible solution to the problem is to point out that extrapolated volition is the volition we would have if we knew more, and presumably greater knowledge would make us prefer contexts with less evil and more good. A problem with this proposal is that what many people would prefer if they had very much knowledge is probably quite different from what they would prefer if they knew little. But if they knew much they would be very different people. Is it then good to give people what they would prefer if they had very much knowledge when the fact is that they do not have very much knowledge? Maybe having very much knowledge would make you love spending a year studying French deconstructionism, which you might hate with less knowledge. If we knew more, thought faster, etc., our volitions would be different, but we would also be different, so how should ethics balance our real volitions against our ideal volitions?

Here is a third problem with the idea of coherent extrapolation: making volitions coherent is not enough to say which volitions have more worth when you have to choose between them. How should we value different goals by different numbers of people with different probability of succeeding with different time scales of reaching the goals when we have to weigh them against each other?[7] Is it possible to compare at all? It thus seems that many traditional problems of weight in ethics have no solution in this proposal, since it seems that quite different alternatives could be equally coherent.

As an alternative to Yudkowsky's proposal, Bostrom suggests two other possible ethical principles for a superintelligent AI to follow. The first is to say to the AI that it should do what is morally right, then make it find out itself what is morally right (Bostrom 2016, pp. 266–267). This suggestion is meant to solve the problems for Yudkowsky's model, since the machine would then have to find out whose volitions to include, avoid extrapolating evil preferences, and find the right balance between different considerations.

However, Bostrom also points out weaknesses with this proposal. It does depend on there being a clear meaning to the term "morally right", which may well not be the case since it has been a contested term throughout history (Bostrom 2016, p. 267). I believe that there is no correct definition of the term "morally right"; rather it can mean many different things and have the meaning we choose to give it. This means that if you were to tell a very intelligent AI to do what is morally right, I think it would answer, "'Morally right' can mean many things, so what do you mean?".

It would not help to refer to what most people mean with "morally right" since they mean very different things, from the will of God to what gives humans pleasure. Nor would it help to be very abstract or minimal since too little information would follow. Nor would it help to tell the machine to choose the most coherent interpretation since the basis for making a coherent interpretation would be too small. While I am trying myself to suggest the most coherent interpretation of morally good, I am also adding information and making choices concerning what we should choose to be the basic meaning of "morally good".

Another problem Bostrom points out is that this proposal may well have as a consequence something that humans do not want. For example, the AI could find that it is morally right to do what is morally best, and further that "morally best" is determined by a kind of hedonistic utilitarianism, saying that there should be a maximal amount of pleasure in the universe as soon as possible, which it actualizes by removing all humans and creating beings in a state of pleasure (Bostrom 2016, pp. 268–269). Given that this was actually what was morally best, it would not be a *moral* problem, but still it would be a problem for us if we want a superintelligent machine not to exterminate us.

Bostrom's second proposal is meant to solve the problem that what "morally right" means is unclear while also avoiding a machine that could exterminate all humans. This proposal, called "do what I mean" (DWIM), is to say to the AI that it should do what we would have best reason to ask it to do (Bostrom 2016, p. 270). It should then find the definition of "morally best" that we have best reason to give, and it should create a world where we are not all exterminated (presuming that we do not have best reason to want to be exterminated, which seems plausible).

The problem here is what to mean by "best reason". Since I believe that any reason is relative to a goal, I think the machine would reply that what the best reason is is relative to what the goal is, and then it will ask what the goal is. If the goal is what is morally best, the proposal turns into the "do what is morally right" proposal. If the goal is the most

[7] This point is made by Allen et al. that a major problem for top-down approaches to ethics in machines is that the rules we give to machines will contain conflicting rules (Allen, Smit, & Wallach 2005, p. 149).

coherent combination of what everybody wants, the proposal turns into the coherent extrapolated volition proposal by Yudkowsky. Bostrom also concludes that this proposal turns into one of the other proposals and inherits their problems (Bostrom 2016, p. 271).

To sum up, the CEV model has the problems of 1) whose volitions to include, 2) avoiding acting on evil preferences, 3) determining how to extrapolate volitions, and 4) difficulty balancing volitions. The morally right (MR) proposal has the problem of what "morally right" means and that it may exterminate all humans. Depending on how we interpret the DWIM proposal, it inherits either the first or the second set of these problems. We thus want a principle that can avoid these objections, and I shall suggest one in the next section.

3 A proposal for an ethical principle for an intelligent AI to follow

In this section I shall present a suggestion for what overarching principle a superintelligent AI should follow and show how it can avoid the problems with the previous suggestions. I shall then answer objections to this suggestion in Sects. 4, 5. I start now by presenting the suggested principle.

The principle I suggest is as follows: "Actualize the best way to the best world (BWBW) through evaluated small steps of improvement". By "best" I mean what would be valuated the most by the most, and by "most" I mean anyone probably having a conscious self (or whatever is necessary to be able to valuate something) who exist now or come into being in the future.

To "valuate" something is a broad term which means that an individual experiences something as good – either individually good for him- or herself or ethically good for all and either instrumentally good for reaching another goal or just good for its own sake. This implies that valuation does not have to give a sense of pleasure, but rather just be something that an individual prefers instead of something else. Valuation does not have to be an intellectual procedure, it just requires the capacity for conscious experience of something as good or preferable. Nor is it merely a matter of what gives most pleasure, but a matter of having goals and selecting between them. I use the term "valuate" instead of value or evaluate to specify this particular meaning.[8]

I add "through evaluated small steps of improvement" since, even if you are the most superb intelligence we can imagine, it is always possible that there is something that you do not know that you do not know. Reality may have unknown deeper levels or outer areas or latent laws or hidden indeterministic possibilities that make the world today and that of the future different from what we, or the most intelligent mind possible, thought. Even a possibly superintelligent machine cannot be sure that it knows what it is to be like others and experience what they do. Because of this uncertainty, I think that ethics should evolve through small steps of improvement that are evaluated as actually being an improvement before moving on, such that it becomes a gradual exploration of what is the best way to the best world. Since our desires will change gradually over time and our discovery of what is the best world will happen gradually, ethical decisions should be made through a gradual process of discovery as well.[9]

While this may seem like classical utilitarianism, adding *the best way* to the best world is meant to avoid the traditional objections to utilitarianism. Traditional objections point out unfair situations where, for example, 90% decides to hold 10% as slaves and the pleasure of the 90% is said to outweigh the pain of the 10%, or maybe all organs are taken from a healthy person to save lives of more people in need of different organs.

One can think of many scenarios of a majority exploiting a minority to use as argument against classical utilitarianism, but they do not work against this model. The reason is that exploiting a minority is not the best way to the best world: there is a better way to a better world where the minority is not exploited and where the majority prefers not to exploit minorities. This is a better way since it is more valuated by more people. While traditional act utilitarianism just asks what action actualizes the best world in the moment of choice, this model considers different ways to the best world, which includes the possibility that the majority can and should change their preferences.[10]

This is a similar move as made by rule utilitarianism, which asks which rules give the best consequences. And it has overlapping similarities with the preference utilitarianism of Richard Hare, emphasizing changeable preferences over happiness or pleasure, etc. (Hare 1952). Using broad

[8] Daniel Hausman distinguishes between different concepts of preference: enjoyment, comparative evaluation, favoring and choice ranking (Hausman 2012). By "valuation" I mean a comparative evaluation, which in the brain is based on enjoyment/desire even if that enjoyment is not always consciously felt, and which comes (fallibly) to expression through choices (see more on this below). In order to use valuation as a coherent foundation for ethics, it is important to include both that it expresses appreciation and that it enables us to compare alternatives.

[9] Gerald Gaus argues that a theory of how to improve society often must make a difficult choice between making an improvement relative to earlier or going in the direction of the highest goal (Gaus 2016, p. 142). These two alternatives coincide in my proposal, since the way to reach the highest goal is to make small improvements (the way thus being part of the goal, while the concrete content of the goal is unknown).

[10] With this addition, there is an aspect of virtue ethics to the utilitarianism I am proposing.

categories, the theory suggested here is a preference utilitarianism, but its distinctness lies in how I develop the details to avoid objections, as I do below.

There are of course many practical problems of how such a principle should be implemented and how the machine should attain the information it needs. I shall discuss this below. I start now by considering how this principle avoids the problems of the MR model and the CEV model, starting with the MR model. Here we assume that the machine under discussion is superintelligent, as Yudkowsky and Bostrom do. Below I shall discuss what is implied by the principle if the machine is less intelligent, since it would be good to implement such a principle in less-than-superintelligent machines on their way to superintelligence.

The problems of the MR model are easily solved. The BWBW principle seems to make it very probable that an intelligent machine will not exterminate all humans since it is hard to imagine that that would be the best way to the best world. What the most would valuate the most seems not to include all people being dead. Also, the term "morally right" has been given a definition, namely the best way to the best world, which has also been further defined, although I shall continue defining it further yet below.

If one disagrees that this is the correct definition of what is morally good or the best world, one could say instead that my proposal is about how to find out which possible world (given today as the point of departure) would be maximally preferred, and how to make intelligent machines actualize the world that would be maximally preferred (including preferred by us). A critic should then argue why it would be better to have another principle, or a vague conception of "morally good in the unknown correct sense".

Why should "what is valuated the most by the most" be the goal we give to a superintelligent machine? Why is that the best justified goal? It is the best justified goal since it is the goal that integrates the most goals: That which is valuated the most by the most is that which makes the most individuals reach most of their goals. It is thus the best justified goal in the sense of being the most goal-inclusive goal. All the individual reasons have been summed to a best overall reason.[11]

The BWBW model is quite similar to the CEV model, so what is the difference between "what most would valuate the most" and "coherent extrapolated volition"? And how does the BWBW model deal with the problems of the CEV model? To recap, the four problems were (1) whose volitions to include, (2) avoiding acting on evil preferences, (3) determining how to extrapolate volitions, and (4) difficult balancing of volitions.

The main difference is probably the specification that the basis of ethics is what most would valuate (as defined above) the most in different alternative scenarios as opposed to having as a basis the desires of today extrapolated into a coherent volition. Maybe the most coherent extrapolated volition is what people would valuate the most, in which case the alternatives are similar. But while extrapolated volition is vague and allows for extrapolation in many different directions, I offer a specification of the goal, namely that individuals should experience to the maximal degree that they valuate the circumstances they are in.

This then helps to solve problem 3 in determining how to extrapolate volitions, and it helps to solve problem 2 on avoiding acting on evil preferences. Obviously there is a possible world where people do not act on evil preferences that would be valuated more by most than a possible world where they *do* act on evil preferences. For example, even if the world had been such today that 90% of people would find it fantastic that the remaining 10% were killed, there is an alternative world where 90% do not valuate killing the remaining 10%, which is overall more valuated by more.

Again, if Yudkowsky could show that the most coherent extrapolated volition is one where such evil preferences are not carried out, the objection by Bostrom would not be a strong one. But in fact, Yudkowsky seems to admit that the objection is a problem which he does not have a good answer to. Showing that what matters is valuation shows why an evil extrapolation would not be preferred over a good extrapolation, since a good extrapolation by definition will be preferred more in total than an evil one (it would not be evil if everybody loved it).

My model is thus quite similar to Yudkowsky's, but I believe that emphasizing what would be valuated the most makes it easier to answer also the remaining unanswered questions of whose volitions to include (and why) and how to make balancing less difficult. It is also an important part of my model that I emphasize exploring what the best way is towards an unknown goal, where changing what we valuate the most along the way is part of the option. It is an ethics where we explore what is the best ethics. This lets us avoid the objection that human volition today is not a good guide to the best ethics since some human volitions need to be changed.

It was above objected against Yudkowsky that he was unclear on whose volitions to include and why, so how do my suggestions deal with the problem of whose preferences to include in determining what is valuated most by the most? Concerning whose preferences to include, it is a good start to say that it is those who have a preference. Since we do not know which animals have a conscious preference and a self considering something to be good or bad, it is better to be

[11] The do-what-I-mean (DWIM) proposal suggested that the machine should do what we have best reason to ask it to do. If "best reason" is interpreted as here, meaning the most preferred way to the most preferred goal, my proposal could be seen as a version of the DWIM proposal.

safe than sorry, and so we should assume that most animals with brains can have a conscious feeling of pain and pleasure. More intelligent machines in the future may know better, including which robots in the future have conscious preferences. Even if a superintelligent machine would not know exactly what it is like to be someone else (like having indexical knowledge), maybe a future AI with a deeper understanding of consciousness and the possibility of neuron cables between brains (or something with the same function) can let us know much more about what it is like to be someone else.

This still leaves open what to think of the potential preferences of fetuses, persons with damaged brains, and future generations. The best way to the best world must include future generations, including fetuses, persons with destroyed brains who might become healed, and unborn people of the future. Certainly, we would have wanted earlier generations to think of us in their treatment of the planet and the use of resources by the state, and we should do the same. So far, it seems that it would have been good advice to tell Archimedes to find the best way to the best world, meaning the world most valuated by most and including future generations.

These reflections were my answer to the problem of whose volitions to include. The last problem was the problem of difficultly weighing preferences. This is a question where intelligence over the human level would be extremely helpful in finding out what most would prefer the most. In the following, I shall try to make some suggestions on how to weigh preferences just to indicate that it is, in principle, possible even if very difficult in practice. If it is possible in principle but difficult in practice, a future very intelligent machine can do a good job, unlike if it were in principle impossible. What I say about weighing preferences in the following are relevant both for machines with human-level intelligence and those with superhuman-level intelligence. I start by discussing weighing preferences in general and end by discussing specifically machines that are limited in their intelligence.

Since the problem of weighing preferences is a set of related problems, I have collected them in a section of their own. Allen et al. say that the two major problems for top-down approaches to machine ethics is that rules are in conflict and it is difficult to make it work in practice. I deal with the first objection in Sect. 4 by looking at different conflicts, then I deal with the practical objection in Sect. 5.

4 Objections concerning weighing of preferences

In this section I shall offer reflection on the following general questions: can valuations be compared? How can valuations be measured? How should we choose if two valuations are equal? How should one decide between a low value with high probability of occurring and a high value with low probability of occurring? These questions will be discussed in the order mentioned. I discuss ideal solutions to the questions first, then end by discussing how to deal with the questions when resources are limited. In other words, if a machine is not very intelligent or does not have very many resources (like most humans), how should it deal with these problems?

The first two questions are *Can valuations be compared?* and *How can valuations be measured?* Some argue that everything we value and dislike can be placed on a scale of and measured in degrees of pleasure or pain (Moen 2012, p. 37). Others argue that values are too different to be comparable, such as when different people valuate and prioritize finding truth, self-sacrificing in helping others, collecting stamps, whistling, having sex, etc. J.S. Mill famously (and to me, convincingly) argued that the things that people valuate are so different that they cannot all be considered on one scale of pleasure and pain. This is one of the reasons for me to follow the lead of Richard Hare and choose the wider concept of valuation, since people have different preferences and may, for example, prefer things which are painful. But can preferences be compared and measured?

Let us start by considering one individual making choices in situations that are similar to each other. We all compare our own preferences all the time by choosing one thing over the other. Of course, we may sometimes choose A over B without knowing what we would actually have valuated the most. But sometimes we have experienced both A and B and know what we will choose if we have to choose again. We often know what we prefer the most, but sometimes we are unable to choose, since our preferences are equally strong. These choices that we do all the time seem to reflect that there is some *scale of desire* in our brains, which is of course context dependent and coarse-grained.[12] Even if I do not know how to understand in any detail this scale of desire and its units, it seems very fitting to use in ethics, which is presumably one of the reasons that Richard Hare used preferences as the basis for his ethics.

That people make choices seem to express that they can valuate one alternative over another. But maybe they made a stupid choice. Valuating one alternative over another would most appropriately be judged by comparing how an individual would valuate her life resulting from choosing one alternative compared with how she would valuate her life had she chosen the other alternative. Peter Baumann argues that such comparisons cannot be made and points out how different choices turn us into different people with different preferences, making it impossible for an individual to make a choice based on preferences (Baumann 2018).

[12] For evidence that our choices happen by a degree of desire reaching a certain threshold, see (Roskies 2014).

It is true that the individual cannot know which life she would have valuated the most, for example whether she would have valuated most a life with or without children, since she will only live the one life. But we can imagine the person living both lives and grading both lives even if the person would be different and have different preferences. And it is a true answer to the question of what would most probably (in the sense of epistemic probability) be the most valuated life by that person given maximal knowledge of the situation, even if nobody knows the answer. What will probably be most valuated is the best basis for making the choice, even if no certainty can be achieved. Ethical choices include uncertainty about consequences. It makes sense to compare two possible futures according to which would most probably be most valuated, which of course is something we do when making choices all the time. While Baumann says that there is no plausible metastandard for comparing possible futures, I argue that the metastandard for an AI to use is what most probably would have been valuated the most.

However, the two previous paragraphs only considered comparing preferences within one person. Can preferences be compared from person to person? If an intelligent machine has to choose between letting person A actualize his preference for getting a rare stamp for his collection, person B actualizing her preference for meeting a friend, and person C and a lot of other persons with different preferences, how can it decide among them?

This seems like a big problem also for those who think that all value can be placed on one scale of pleasure. For even if every individual says that they experience something as, for example, 7 on a pleasure scale of 1–10, we cannot compare one person's 7 to another person's 7. Maybe person A has a wide emotional register and person B has a very narrow register, so that a 2 for A would have felt like a 7–B. This would then be a problem if the goal is just to achieve the biggest amount of pleasure in the world.

On the other hand, one could argue that experienced value should be considered relative to each person. One could say that it is irrelevant that a 2 for A feels like a 7 for B; instead it only matters what each one feels to be, for example, a 7. As long as we cannot compare experiences between persons, considering these experiences relative to each person seems to be the right move. It is still relevant to talk about amount of value in terms of number of people experiencing a certain amount of value over different periods of time, but the amount of value that each person experiences at a time must be considered relative to that person.

Rachael Briggs offers some arguments against the idea of comparing alternatives relative to each person like this. With reference to Peter Hammond, she offers the example of a greedy person who needs a lot to experience, for example, a value of 7 versus an undemanding person who just needs a little to experience a value of 7. Is it then the best ethical option to give a lot to the greedy person ((Briggs 2015) referring to (Hammond 1991, p. 216))? My answer to this is that, as a starting point, we must measure and compare experienced value relative to each person. But note that the ethical principle I suggest is not that we should do in each situation that which satisfies the most preferences in that situation, but instead that we want to actualize the best way to the best world. In this case, that would be that the greedy person was less greedy, in which case the resources could be shared to let more people experience more valuation. I say more about this below, pointing out the importance of social equality for making everybody (poor and rich) experience the most valuation.

The conclusion in the previous paragraphs is that you cannot compare one person's 7 with another person's 7 to find which one is best, but must instead say that one person's 7 has the same value as another person's 7. What then about comparing one person's 7 with another person's 2? If you have to choose between letting one person experience a pleasure that she values to be a 7 or another person experiencing a pleasure that he values to be 2, does that mean that we should choose to let the first experience a 7 since that will bring more valuation into the world (in each case the other person will experience something of pleasure value 0)?

It seems that this might be a good choice to make in individual cases, but not on a permanent basis. If a machine has to choose between letting one person get 10,000 dollars or another person getting 100,000 dollars, it seems good to choose 100,000 dollars if the persons are otherwise similar and no other negative consequences follow. But the same reasoning does not seem to be a good choice on a permanent basis. It does not seem ethically right to prioritize person A experiencing pleasure 7 day after day instead of person B experiencing pleasure 2 day after day with the argument that prioritizing person A brings more valuation into the world.

Here it looks like this utilitarian ethic needs a criterion of justice in order to distribute the good fairly. Does this mean that teleological ethics do not work on their own? I think the justice aspect can be brought in by seeing that when we consider human lives, there is greater value in raising somebody from a general life quality of 2 to a general life quality of 3 than raising somebody from a general life quality of 8 to a general life quality of 9. To exemplify, it is better that somebody who does not have food and education can get food and education rather than that somebody rich gets even more money and holidays.

Why is that right? To explain this with the teleological approach defended here, I point to three interconnected lines of reasoning. First, it is a fact that somebody who is hungry appreciates getting food more than somebody who is full appreciates getting another ice cream. Raising someone from 2 to 3 thus involves more valuation than raising someone from 8 to 9. We can see this from a thought experiment: If

we had to choose, from behind a veil of ignorance, whether to prioritize people getting from 2 to 3 or from 8 to 9, and afterwards we would ourselves either become someone at 2 or someone at 8, without knowing before where we would end, most people would prioritize that persons should be raised from 2 to 3. This is how John Rawls has argued in favor of securing basic needs (Rawls 1971), but here I gave it a utilitarian justification: we valuate more a person being raised from 2 to 3 than from 8 to 9.

Secondly, it is a fact that how much people valuate something depends on who they compare themselves with. Comparing the wealth of people and their happiness shows that they do not get happier by becoming richer, because it makes people need more in order to be happy (Harari 2017, pp. 38–40). But it is required for happiness to get above the basic requirements for survival – in other words, it is more valuated to go from 1 to 2 than from 8 to 9 in general life quality. This insight is also expressed in Maslow's hierarchy of needs.

Rachael Briggs uses the insight that happiness depends on comparing as an objection to an ethics based on comparing how people valuate different alternatives. She argues that it implies that people can change their welfare merely based on the alternatives they consider in their mind, which seems absurd (Briggs 2015). Instead of finding it absurd, I find it obviously true, and researchers on happiness say that the most efficient way to improve your happiness is to remind yourself of what you are grateful of, including being grateful for the problems you are not having (Emmons and McCullough 2003). However, what you actually see around you clearly has a much stronger effect on your comparison than merely what you choose to think about. Seeing the luxury of your neighbor influences you more than thinking about poor people in another country. For this reason, I think that it is important that societies are actually characterized by economic equality, while merely thinking about hypothetical goods and evils influence our happiness much less.

Thirdly, as seen above and closely connected to the previous point, it is not just relevant what people actually value, but what they potentially *could* value as well. While people who are rich today have certain needs today in order to feel happy or valuate something, it seems clear that, since happiness depends on comparing yourself with others, potentially most people would valuate life the most as a total sum if most people in the world had a similar life quality. Great divisions between poor and rich cause envy and conflict, while great similarities cause stability and general contentment – very broadly speaking, of course.[13]

These things cannot be measured exactly, but if one wanted to be mathematical about it, one could make a scale with higher values at the bottom. For example, one could say that going from 1 to 2 in general life quality is worth a million points, going from 2 to 3 is worth half a million points, 3 to 4 is worth 250,000 points, 4 to 5 is worth 125,000 points, etc. Or it could be negative points for suffering, where it would be much worse to go from suffering 8 to 9 than from suffering 1 to 2. This is a way of including an element of justice in a consequentialist ethical theory by letting the theory give greater value to help those with greater need.

This logic is the one economists use when they speak of marginal utility, where a typical curve shows that the utility per unit is high at first, then lower as you add more units. One could be a kind of traditional utilitarian who says that all that matters is raising someone up the scale but that there is no difference between raising someone from 2 to 3 or from 8 to 9. Or one could be what Parfit calls a prioritarian, like me, saying that it is more important to help those who are worse off. If the utilitarian agrees that going from 2 to 3 has a higher total value than going from 8 to 9, there is no disagreement between the utilitarian and the prioritarian.[14] I believe that differentiating the weight of climbing at different places of the ladder is the right way to think about this.[15]

Here I will end this discussion by pointing out that ensuring that people get their basic needs met could be thought of as securing basic human rights. Securing basic human rights could be thought of as first ensuring that everybody gets lifted to general life quality 1, then life quality 2, etc. The list of basic human rights could be prioritized and extended as basic rights are put in place: all people should have food, clothes and shelter, but also clean water and clean air, then it could be extended to more and more health, education, money, etc. I will return to this issue, as well, at the end of this section. This line of reasoning would imply that if superhumans or superrobots that are considered persons in the sense of having a self-conscious mind–and whose mental life is far more complex than that of humans – should evolve or be developed, these should then also secure human rights (i.e., rights of humans such

[13] Data show a very clear correlation between the number of social problems and the amount of economic inequality in a society, see (Bregman 2017, pp. 54–55).

[14] Derek Parfit makes this point, saying that if we give benefits different weight, there need be no disagreement between utilitarians and prioritarians (Parfit 2012). It will also include the point from the egalitarians that increased equality is good.

[15] Many have argued that utilitarianism needs to be supplied with deontological ethics to secure the right of individual not to be used merely as means. I suggest this alternative way of thinking since I believe that sometimes it would be ethically right to use people merely as means, for example if, in a specific scenario, that was the only way to save the world.

as the kind that live in 2020) instead of prioritizing their own pleasures.[16]

The next problem to consider is the following: What if two values are equally good? How does one choose if both options seem to lead to the same amount of valuation? This may happen, and, ethically speaking, in such a case both options are equally good and a random choice is the right choice.

The case is more difficult if there is one scenario which has a high value but a low probability of becoming actualized versus a scenario which has a low value but a high probability of becoming actualized. The problem is well known in ethics and typically turns into a question of what is most realistic. War versus pacifism, different questions in climate politics, or revolution versus revision are examples where some will argue for drastic means to reach high goals while others argue that walking with smaller steps will be a more realistic way to reach the goal.

In general, it seems that a "better safe than sorry" strategy will produce the best results seen in total, but not even a superintelligent machine can know for sure in advance what will actually become the best result. It is a known debate whether one should be more idealistic and revolutionary, going for big changes, or emphasize context more strongly and work for revision, and although revolution may sometimes be best, the revisionary strategy seems to have the empirically best support as a general strategy since stability and trust are so important for economic and other factors. In other words, given uncertainty it is generally better to prioritize smaller goals that have a higher probability of success rather than prioritizing higher goals with a lower probability, although the world is too complex to offer an exact definition of where to draw the line.

Can we offer any guidelines when it comes to choosing between different ways of exploring what the best way to the best world is? As mentioned before, it is not possible to give an exact recipe for how to compare alternative actions. As Derek Parfit says, ethics cannot conclude that one action is 2,36 times better than another (Parfit 2011, p. 132). But the lack of fine-grained truths does not exclude the existence of coarse-grained truths (it is true that there are bald people even if we cannot define baldness to an exact number of hairs). I will suggest that the reasoning above can help us somewhat, especially what has been said about prioritizing preferences and prioritizing safe choices. Here are some rough guidelines:

We have already seen that if the one alternative will raise one person from life quality 2 to life quality 3 and the other alternative will raise another person from 7 to 8, we should choose the one going from 2 to 3. If two alternatives will raise one person from life quality 2 to life quality 3 but one alternative has a higher probability of occurring, we should choose the one with the higher probability. Choices with high probability of success should be preferred, and choices preventing suffering or raising those with low life quality higher should be preferred. The tricky part comes when one has to balance number of people, number of life quality or suffering, probability of success, and weighing these against each other.

Even as real life situations can be extremely complex, I will make some general suggestions where we assume the same context for all examples. Imagine just coming to a large group of people who you can help, but it has to be done in a certain order, and sometimes you have to choose one alternative over another. Helping people in this example is not about concrete instances of happiness or suffering, but about considering the general life quality of people, where 1 is a life of suffering and misery and 10 is a perfectly happy life. Here is how I suggest one should prioritize:

If there are people that, with a high probability, can be helped, these should be helped first, starting with those suffering the most and moving to those with less suffering, up to higher and higher life quality. Mathematically put, start with those at 1 and move up the scale to 10. If we say, as above, that moving from 1 to 2 is worth many more points than moving from 7 to 8, we could use a mathematical principle to guide us as long as we are aware that it is very inexact and influenced by many other factors as well. The principle would be to take the number of points times the number of people involved times the probability of succeeding, and opt for the alternative with the highest score.

This reasoning could then be used also when deciding whether to help group A or group B. The alternative that gets the highest score is the one to choose. If there is a choice where you can help many but a few get it worse, that counts as a negative score that it takes much to make up for. Raising a lot of people from 8 to 9 is not worth it if the price is to take a few down from 5 to 2, such as by exploiting workers. But it may well be worth raising a lot of people from 2 to 5 even if it means taking somebody down from 9 to 8, perhaps by adding taxes for the rich.

While this may seem like an absurd mix of ethics and mathematics, it does help us explain some ethical intuitions that many share. For example, it seems that the life quality of many people in North Korea is very low, meaning that helping them achieve a better life would be worth more than helping a group of people in a country suffering less. But

[16] If robots are not conscious (but still highly intelligent). They do not have a unified conscious self, only a representation of themselves as the whole robot, similar to how our brain has a representation of our body. Then they also cannot have goals for their own sake, in the sense of something they just valuate because it consciously feels good. In practice, they could nevertheless have something very comparable to human goals, but just as dispositions for acting in certain ways.

if the probability of succeeding with a specific attempt of spending time and money in North Korea is very low while the same attempt of spending time and money somewhere else has a high probability of succeeding, it may nevertheless be ethically right to prioritize the other place.

So far I have focused solely on short-term goals and not how to balance short-term goals with long-term goals. Again, long-term goals will usually have a lower probability of success, which lends support to prioritizing short-term goals with a higher probability. This follows from the suggestion of gradually exploring what the best way to the best world is. Again, a guiding principle would be like the one suggested above: to take the number of points times the number of people involved times the probability of succeeding and opt for the alternative with the highest score.

Even if answers like these are very imprecise and full of exceptions, many other ethical theories have no answers at all when it comes to what to do when you have to choose between different alternatives with different numbers of people with different needs and different probabilities of success. Unless they are clearly wrong, rough guidelines are better than no guidelines. A good critique of this alternative should offer better guidelines given that we often actually have to make a choice (not acting or choosing at all being a bad alternative).

I have now spelled out in more detail the principle suggested, especially the part of weighing consequences, by answering relevant possible questions and objections. However, the discussion has focused on what is the best moral choice, which is something a superintelligent machine should aim for, but what is morally good enough for a machine with less intelligence or fewer resources? Humans are not superintelligent, and while it would also be good for us to bring about the best way to best world as best we can, it is a very demanding ethics if one says that this is what all that humans (or all human-level AIs) should do. So far I have focused on what is morally best, but we should also consider what the minimum requirements are, which I will do in the following.

What if you have a machine that either cannot do what is morally best because of limited resources, or you have a machine but only want it to do what is morally good, not necessarily what is morally best. What is the minimum requirement for a machine to act in a way which is morally good? I shall refer to such actions as doing what a (moral) machine (morally speaking) *should* do.

If we ask what a moral machine at least *should* do (as opposed to what would be best to do), we must take as starting points the actual world at the time and place the machine exists and the resources it has. The moral requirement is then that the (human-level intelligent) machine contributes to making the world better, not to making it worse.

Making the world better is good/right/should be done; making it worse is bad/wrong/should not be done. If an action does not make the world better or worse, it is morally neutral. This applies to both humans and to human-level intelligent machines. However, it can still seem both too demanding and too little. Should I always spend time saving dying children in Africa instead of buying a pair of new shoes or spend time with my sick mother? Is it enough that a very rich person gives one dollar to charity since the world then got better? We need to add the qualifier that everyone should make the world better *in light of their resources*. People (and companies and states) with more resources should do more than those with fewer resources.

How do we then decide what agents should do in light of their resources? It is by comparing their actions given their resources with what is the best way to the best world. It seems clear to me that the best way to the best world is a world with room for buying new shoes and taking care of one's sick mother, since the best way to the best world would be a way where nation-states, through taxation rules supported by all, take care of the ones in greatest need. It is also a world where very rich people give more than one dollar to charity.

If we want moral machines to qualify as acting in a way which is morally good (even if not morally best) for that machine, we must consider how that machine can make the world better given its resources and that is what it should do to make a morally good action (remember that I am not discussing what it must to do be a moral agent).

In the final part of this article, I shall answer some other possible objections.

5 Other objections

In this section, I answer five different objections. The first objection is that it is impossible to make it work in practice – how do you make ethical rules computable for machines? Allen et al. point out that machines would need an extreme amount of data to be able to compute what large groups of people would prefer (Allen et al. 2005, p. 150). I am not able to answer the practical questions in detail, and the focus in this article is on the ethics, not the engineering. But I will suggest some main points as a reply, which could hopefully be better developed by others with more AI competence than I have.

I have suggested as a guiding ethical principle to take the number of points times the number of people involved times the probability of succeeding, then to opt for the alternative with the highest score. It is possible to run on computers very advanced simulations of social interaction, with a large number of variables, and predict outcomes of

various events.[17] Through the World Happiness Report, we do have quite detailed data on how different kinds of life conditions make people evaluate their life on a scale from 1 to 10. Machines could learn much by running simulations and matching those with the data from the World Happiness Report.

Machines could also learn from suggesting scenarios and asking people to valuate them, although that is trickier since it is an important point that people's preferences change over time in different contexts. Asking people 50 years ago about their evaluation of gay marriage is not a good indicator on the moral value of gay marriage, and asking people in the US about raising taxes is not a good guide for how they would actually evaluate paying higher taxes in a country working as it does in Norway.

However, machines could learn how much people prefer different alternatives and try to look for correctly weighted rules that match people's preferences. This could be done on small and easy tests first, then on big and complicated scenarios. I also said above that one could give different points for going from 1 to 2 than from 8 to 9, and setting reasonable weights here is also something that could also be tested with simulations. Similar simulations could be used to determine what a reasonable threshold is for what counts as gradual steps in improving. While this is in no way a complete solution to the practical problems, I see much potential for training morality to machines with this ethical model as a point of departure. It would be a way of training machines in a combined top-down and bottom-up approach.

The second objection is by Ernest Davis, who disagrees with Bostrom's claim that it is difficult to make a machine understand ethics and to implement a good moral principle in it. Davis argues that a machine intelligent enough to understand and interact with humans must also understand their morals and the concept of morality itself (Davis 2015, pp. 122–123).

It would be easy to give a very superintelligent AI a good moral principle to follow, according to Davis. It is just to specify a collection of admirable people from the past whom the AI will know everything about, and then instruct the AI: "Don't do anything that these people would have mostly seriously disapproved of." (Davis 2015, p. 123).

Davis anticipates the objection by Bostrom that the idea is to make an ethics for the future and not one for the past. But he answers that it feels safer with an ethics based on the past (e.g. 2014 or 1700) rather than an ethics based on future decisions (Davis 2015, p. 123). And in any case, he thinks it is easy to add a safe turn-off switch that humans, not AIs, have control over: "All you need is to place in the internals of the robot, inaccessible to it, a device that, when it receives a specified signal, cuts off the power – or, if you want something more dramatic, triggers a small grenade. This can be done in a way that the computer probably cannot find out the details of how the grenade is placed or triggered, and certainly cannot prevent it." (Davis 2015, p. 123).

Bostrom would certainly disagree that it is easy to make a safe turn-off switch for superintellingent AI (Bostrom 2016, pp. 155–176). And he would certainly be critical of the idea that an ethics based on the past would be good for us, given how dominated earlier thinking has been by racism, religious fundamentalism, etc. Seeing how people act teaches us the descriptive morality of how people actually behave, which is not the same as the normative question of what is actually good and right.

It would be very difficult to specify a collection of admirable people from the past, since people would fight over whether to include Buddha, Epicurus, Jesus, Muhammed, etc. The people in such a base would have quite different and inconsistent views on whether we should quench all our desires, whether we should violently resist Hitler, what sexual ethics are right, etc., etc.

In sum, I find the proposal by Davis to be very problematic, and in any case, the starting point for the discussion in this article is to find the best ethics for the future given that we do not know today what the best ethics are.

The third objection is the mere addition paradox. Derek Parfit has argued that if what matters is the total amount of happiness, it seems better with a population of very many people having a little happiness than a smaller population with great happiness, which he calls a repugnant conclusion (Parfit 1984, p. chapter 17). This could seem even worse in the scenario I just described, since I said that low scores are worth more points. But it seems clearly wrong to say that it would be better to have one billion people alive and experiencing pleasure of value 1 rather than if there were one million people alive experiencing pleasure of value 10, even if the total sum of pleasure is greater in the one billion people scenario.

I agree that it is the wrong conclusion to think it better to have many people with a low score of happiness rather than fewer with a higher score. When considering value of pleasure in the world, this should be divided by the number of individuals experiencing the pleasure, so that the goal is to have this kind of highest average score. Since new humans hopefully will continue to be born century after century after century, it is best in total that they all have a higher average score. This might seem to lead to the conclusion that we should strive for having very few people alive at any given point of time, but this does not follow, since it is good for all who live that there are also many others who live at the

[17] See, for example, the work carried out at the Virginia Modeling, Analysis, and Simulation Center: odu.edu/vmasc.

same time that can specialize in different areas helping each other.[18]

The fourth objection is by Marcus Arvan, who presents a trilemma for those who want to program ethical AI: either it will be too semantically strict, too semantically flexible, or overly unpredictable (Arvan 2018). With the solution presented here, I suggest a middle way between the too strict and the too flexible alternatives. It is semantically strict at a very abstract level (determining that the goal is to find the best way to the best goal, where "best" is what most would prefer the most), but semantically flexible when it comes to the concrete content, i.e. what it is that most would prefer most.

Arvan himself suggests that the solution is for AI to engage in mental time travel in order to consider what would be a fair negotiation for those involved, and Arvan suggests four principles of fairness, which he argues that an AI could do better than humans (Arvan 2018). My suggestion is very similar, but instead of Arvan's four principles, I have suggested one principle which I believe explains and justifies his four principles, although I do not have room for that discussion here. If Arvan's solution avoids his trilemma, then so does my solution.

The fifth objection is by Shamik Dasgupta, who argues that AI may become so different from normal moral agents that it will be difficult to make moral concepts meaningfully apply to them. One AI could have several multiplying and changeable minds in several bodies with all sorts of disconnections and relations to time, all of which would seem to make them so different from humans that it is hard to see how they could fit into our moral schemes at all (Dasgupta, forthcoming).

This objection is certainly a challenge to those who want to apply ethics of intentionality or rule-following or virtue ethics to machines. However, I think the approach of consequentialist ethics that I have chosen here can best solve the challenge, since we are interested in good consequences regardless of how the agent should be understood.

6 Concluding remarks

In this article, I have argued that the best principle for a superintelligent AI to follow is to make it find out what is the best way to the best world. The best way to the best world is the *way* that would be preferred the most by the most to the *world* that would be preferred the most by the most. I believe that this is the most coherent interpretation of what morality is all about, but it is also possible to leave the question of definition of morality aside, and argue that this is what we should try to make superintelligent machines do. It is what we have best reason to do since it is the best way to the best goal, and the goal is best because it integrates all preferences to the highest degree.

Derek Parfit argues that the three main ethical theories (Kantian ethics, consequentialist ethics and contractualism) interpreted in the most reasonable way are three ways to the same mountaintop: Kant wants to find laws that it is rational for everyone to follow, but Parfit argues that for it to be rational for everyone to follow, it must be best for all, which means that Kantian ethics imply rule utilitarianism. Contractualism says that we should do what nobody can reasonably reject, but that is just the same as laws that are rational for everyone to follow (Parfit 2011, pp. 412–413).

Parfit does not want to define the goal of ethics, but he wants to find rules that take us to the goal (Parfit 2011, p. 418). I want to define the goal and say that it is the best way to the best world. My analysis of these models is that Kantian ethics and contractualism suggest rules we can follow in order to find the best way to the best world, all being versions of the Golden Rule, where the main point is to consider others as yourself. We will not actualize the best way to the best world if everybody gives themselves special treatment, and thus we need some general rules that all people follow in order for the best world to be actualized. I think that the three ethical models are ways to the same mountaintop because the top is the world which is valuated the most by the most, which is by definition what we mean by "the good", and the three ethical models are suggesting general guidelines which are in fact guidelines for how to get to that top – a procedure for getting up which works even when we do not know what it looks like at the mountaintop.

I suggest that it is possible to let machines learn what actions fall under some of these main rules of ethics, first with easy examples and then with tougher examples. Let them play against a set of humans given moral dilemmas, first simple ones then tough ones, and see how long it takes before they beat the humans. The interesting question then becomes what are the rules the machine plays by when it consistently beats humans? It seems plausible that machines can teach us some important moral insights in the near future.

Acknowledgements Thanks to the members of the TechPhil research group, Asle Eikrem, and two anonymous referees for valuable comments to this article.

Authors' contributions Not applicable.

Funding No funding was received.

[18] One could think that it follows that it is better to just have one person alive experiencing something of value 7. This does not follow, since of course there is a much greater potential for valuation over time with many people alive, but at one point there can be too many.

Data availability No data were used.

Code availability No software was developed.

Declarations

Conflicts of interest The author declares that there is no conflict of interests.

References

Allen C, Smit I, Wallach W (2005) Artificial morality: top-down, bottom-up, and hybrid approaches. Ethics Inf Technol 7(3):149–155. https://doi.org/10.1007/s10676-006-0004-4

Arvan M (2018) Mental time-travel, semantic flexibility, and A.I. ethics. AI Soc. https://doi.org/10.1007/s00146-018-0848-2

Baumann P (2018) What will be best for me? Big decisions and the problem of inter-world comparisons. Dialectica 72(2):253–273. https://doi.org/10.1111/1746-8361.12219

Bostrom N (2016) Superintelligence: paths, dangers, strategies. Oxford University Press

Bostrom N, Yudkowsky E (2014) The ethics of artificial intelligence. In: Frankish K, Ramsey WM (eds) The Cambridge handbook of artificial intelligence. Cambridge University Press, pp 316–334

Bregman R (2017) Utopia for realister. Spartacus, Oslo

Briggs R (2015) Transformative experience and interpersonal utility comparisons. Res Philosophica 92(2):189–216

Davis E (2015) Ethical guidelines for a superintelligence. Artif Intell 220:121–124. https://doi.org/10.1016/j.artint.2014.12.003

Emmons RA, McCullough ME (2003) Counting blessings versus burdens: an experimental investigation of gratitude and subjective well-being in daily life. J Pers Soc Psychol 84(2):377–389. https://doi.org/10.1037//0022-3514.84.2.377

Gaus GF (2016) The Tyranny of the Ideal: justice in a diverse society. Princeton University Press

Hammond P (1991). Interpersonal comparisons of utility: why and how they are and should be made. https://doi.org/10.1017/CBO9781139172387.008

Harari YN (2017) Homo deus: a brief history of tomorrow (First U.S edition). Harper, an imprint of HarperCollins Publishers

Hare RM (1952) The language of morals. Clarendon Press

Hausman DM (2012) Preference, value, choice, and welfare. Cambridge University Press

Martin M, Monnier R (2003) The impossibility of god. Prometheus Books

Moen OM (2012) Because it feels good: a hedonistic theory of intrinsic value. University of Oslo

Muldoon R (2016) Social contract theory for a diverse world: beyond tolerance. Routledge

Parfit D (1984) Reasons and persons. Clarendon Press

Parfit D (2011) On what matters. Oxford University Press

Parfit D (2012) Another defence of the priority view. Utilitas 24(3):399–440

Phillips DZ (2004) The problem of evil and the problem of god. SCM Press

Puntel LB (2008) Structure and being: a theoretical framework for a systematic philosophy (A White Trans). Pennsylvania State University Press

Rawls J (1971) A theory of justice. Belknap Press of Harvard University Press

Roskies A (2014) Monkey decision-making as a model system for human decision-making. In: Mele AR (ed) Surrounding Free will: philosophy, psychology, neuroscience. Oxford University Press, pp 231–254

Søvik AO (2011) The problem of evil and the power of god. Brill, Leiden

Tegmark M (2017) Life 3.0: being human in the age of artificial intelligence. Alfred A Knopf

Yudkowsky E (2004). Coherent extrapolated volition. 1–37. https://intelligence.org/files/CEV.pdf