Social Choice Ethics in Artificial Intelligence

Seth D. Baum
Global Catastrophic Risk Institute
http://sethbaum.com * http://gcrinstitute.org

Forthcoming, AI & Society, DOI 10.1007/s00146-017-0760-1. This version 2 October 2017.

Abstract
A major approach to the ethics of artificial intelligence (AI) is to use social choice, in which the AI is designed to act according to the aggregate views of society. This is found in the AI ethics of "coherent extrapolated volition" and "bottom-up ethics". This paper shows that the normative basis of AI social choice ethics is weak due to the fact that there is no one single aggregate ethical view of society. Instead, the design of social choice AI faces three sets of decisions: standing, concerning whose ethics views are included; measurement, concerning how their views are identified; and aggregation, concerning how individual views are combined to a single view that will guide AI behavior. These decisions must be made up front in the initial AI design—designers cannot "let the AI figure it out". Each set of decisions poses difficult ethical dilemmas with major consequences for AI behavior, with some decision options yielding pathological or even catastrophic results. Furthermore, non-social choice ethics face similar issues, such as whether to count future generations or the AI itself. These issues can be more important than the question of whether or not to use social choice ethics. Attention should focus on these issues, not on social choice.

Introduction
As artificial intelligences (AIs) become more powerful and more embedded in important social systems, it becomes more important to program them to have ethical frameworks built in. In the words of Picard (1997:134), "The greater the freedom of a machine, the more it will need moral standards." To take one of many examples, the development of autonomous vehicles puts AIs in life-and-death situations, including ethically difficult situations such as when a collision is inevitable and the AI driver must decide which dangerous collision to have (Lin 2016). Outcomes of such situations can depend on AI design decisions, making it incumbent upon AI developers to carefully choose the ethics that are built in.

This paper discusses an approach to AI ethics in which the AI is designed to act according to the aggregate ethical views of society. This approach underlies at least two significant lines of thinking in AI ethics. One is "coherent extrapolated volition" or CEV (Yudkowsky 2004; Muehlhauser and Helm 2012; Bostrom 2014). CEV was formulated specifically for the ethics of superpowerful, superintelligent AI (ASI) that would or could take over the world. CEV abstains from selecting an ethical view for the initial programming and instead seeks to have the AI derive its values from the values of other ethical agents. CEV specifically seeks to extrapolate beyond agents' existing ethical views, essentially to figure out the views that the agents would ideally have if they were as smart as the ASI.

The other line of thinking is the concept of "bottom-up" ethics (Allen et al. 2000; 2005; Wallach et al. 2008; Wallach and Allen 2008). An AI with bottom-up ethics is designed to learn ethics as it interacts with its environment and with other ethical agents, similar to how human

children learn ethics as they grow up.[1] While bottom-up ethics can be consistent with a range of ethical frameworks, some instantiations of it attempt "to train or evolve agents whose behavior emulates morally praiseworthy human behavior" (Allen et al. 2005:149). Bottom-up ethics is contrasted with "top-down" ethics in which the AI is programmed to have a specific ethical view from the start and thus does not seek to identify the views of society or any of its members. Where bottom-up ethics is based on learning from other ethical agents, the ethics views it ends up with will be some aggregation of the agents it learns from.

Though it is not explicitly identified in their respective literatures, CEV and bottom-up ethics have the essential structure of what is known in economics, ethics, and political science as social choice: the procedure for deriving group decisions from individual ethical views.[2] The ethics of social choice is rooted in certain notions of procedural justice, and it underlies both democracy, in which individual preferences are expressed through voting, and capitalism, in which individual preferences are expressed through market behaviors ("voting with dollars"). The study of social choice dates to the early work by political theorist Marquis de Condorcet (1785); modern scholarship largely traces to Arrow (1951) and today is a robust field as featured in dedicated journals such as *Social Choice and Welfare*.

Work on social choice often falls under the term "social choice theory". This paper uses the term "social choice ethics" to refer to the view that the right ethical framework to use is that which corresponds to the aggregate ethical views of society. Social choice theory concerns theoretical issues that arise in the context forming aggregate views. Social choice ethics also includes empirical issues regarding what ethical views individuals hold. In addition, whereas social choice theory often focuses on the aggregation of individual preferences, social choice ethics can include ethical views that are not readily described in terms of preference.

This paper also uses the term "predetermined ethical views" to refer to top-down, non-social choice ethics. The views are predetermined in that they are set prior to any observations of society's views and are used regardless of what society's aggregate views turn out to be. For example, one could attempt to program an AI to maximize the welfare of all sentient beings, as in welfarist utilitarianism (e.g., Ng 1990), or to maximize the fitness of ecosystems, as in ecocentric consequentialism (e.g., Holbrook 1997). While this paper is not concerned with the merits of various predetermined ethical views, the reader should note their availability as alternatives to social choice ethics.

CEV and bottom-up ethics have the structure of social choice in that they identify ethical views held by various ethical agents and combining these into some aggregate view to use for decision making. In CEV, the agents are those whose volition is to be extrapolated. In bottom-up ethics—or at least in certain instantiations of it—the agents are those whose behavior is to be emulated. It is thus remarkable that the literatures on CEV and bottom-up ethics have not discussed issues of social choice in any length. This paper seeks to address this gap. As we will see, some nuances of social choice make CEV and bottom-up ethics more difficult to implement in AI and less ethically desirable than they might initially appear, which, in turn, makes predetermined ethical views more desirable. Furthermore, some issues faced in social choice

[1] Note that while consciousness may play a role in ethics learning among human children, it is not essential for AI. The essential feature is that ethics is learned via interaction with the environment, regardless of whether that interaction involves consciousness.

[2] One exception, in which social choice is (briefly) discussed in the context of CEV, is Tarleton (2010). Keyword searches in Google Scholar identified no other discussions of social choice in CEV or bottom-up ethics. There is a more extensive study of "computational social choice" relating aspects of social choice theory and computer science (Brandt et al. 2015).

ethics parallel those faced in predetermined ethical views; how these issues are resolved can be more important than whether AI is designed with social choice or predetermined view ethics.

Some Preliminaries

Why might one favor AIs with social choice ethics? Several justifications can be found:

1. The *procedural justification*. This is the procedural justice intuition that individuals should have a say in decisions that affect them, as epitomized by the classic American dictum "no taxation without representation". Per this, it would be unfair for AI designers to impose their own ethics views on everyone else by programming AIs with their choice of predetermined, top-down views. Yudkowsky (2004:17) makes this rationale explicit in stating that AI programmers "do not deserve, a priori, to cast a vote larger than anyone else" in how an AI is designed.
2. The *abstention justification*. Some individuals may rather not wrestle with ethics for themselves, so they abstain from choosing ethics and "pass the buck" onto society as a whole. Per this, AI designers choose a bottom-up ethics design in hopes of producing an ethical AI without themselves thinking about ethics. This justification is consistent with AI veteran Stuart Russell's observation that the AI field has shown little interest in social impacts (Bohannon 2015), and the more general observation that research ethics has concentrated on process issues like plagiarism instead of ethics embedded in the research itself and ethics of the social impacts of research (Schienke et al. 2009; 2011). Similarly, Muehlhauser and Helm (2012) observe that all moral theories thus far developed by humans raise significant objections and propose that an ASI may be more capable at determining a better moral theory.
3. The *wise crowd justification*. It is sometimes posited that better results are achieved when using the views of many individuals, as in the maxim "wisdom of the crowd". Per this, an AI's ethics is likely to be "better" according to some neutral standard if its ethics come from many individuals instead of from one. Thus, a market democracy could outperform a communist dictatorship because it empowers many people to contribute their unique insights; ditto for an open, interactive scientific community vs. individual scientists working in isolation. Likewise, having an AI aggregate across the ethical views of many individuals could "smooth out the rough edges" of humanity—that is, unless only humanity's edges are smooth, i.e. unless "large segments of humanity have base or evil preferences" (Bostrom 2014:217), in which case social choice approaches could yield worse results.

There is an inherent tension between the procedural and wise crowd justifications. The wise crowd justification depends on having some notion of "better results". This notion cannot come from the aggregate views of many individuals—that would be circular logic. ("The aggregate views of many individuals achieve better results according to the aggregate views of many individuals.") Instead, it must come from some predetermined ethical view, which is precisely what the procedural justification seeks to avoid. The wise crowd justification is essentially an empirical claim about the instrumental value of social choice: given a predetermined ethical view, querying large crowds will tend to deliver better results per that view. The truth of this empirical claim is beyond the scope of this paper. If it is true, it still leaves open the question of which predetermined ethical view to design AIs with.

The rest of this paper is mainly concerned with the procedural and abstention justifications. Specifically, the paper argues that these justifications are severely limited, because it is impossible for AI designers to avoid embedding certain ethics views into an AI. This is because there is no one single aggregate ethical view of society. Instead, there are many aggregate views depending on how the views are aggregated. These different aggregations can have very different consequences, some of which could be considered pathological or even catastrophic. Therefore, choosing a social choice ethics for AI does not absolve the AI designer from thinking about ethics, or from making decisions about ethics similar to those required for designing AI with predetermined, top-down ethical views. This weakens the case for social choice ethics and likewise strengthens the case for using predetermined ethical views.

Implementations of social choice ethics must make three types of choices, each of which create their own set of ethical dilemmas (Baum 2009):

1. *Standing*: Who or what is included in the group to have its values factored into the AI?
2. *Measurement*: What procedure is used to obtain values from each member of the selected group?
3. *Aggregation*: How are the values of individual group members combined to form the aggregated group values?

The AI ethics literature has focused mainly on measurement, and the social choice theory literature traditionally focuses on aggregation. However, all three pose serious challenges.

The issues of standing, measurement, and aggregation arise in all implementations of social choices, not just those involving AI. Likewise, many of the ethical issues of standing, measurement, and aggregation that are faced by AI designers are also faced by people designing social choice systems in other contexts. The paper thus draws on scholarship and experience from other contexts (including, but not limited to, social choice theory) and relates them to AI. Meanwhile, AI poses some novel issues for social choice. First, with AI, the social choice process is conducted by machines, not by humans. With AI, a machine can be sent out into society to figure out what society wants it to do and then attempt to do it. Human designers of AI must make decisions about standing, measurement, and aggregation and translate this into the technology. Second, there is the question of whether the AI itself should have standing. Arguably, a sufficiently advanced AI should. This raises distinct ethical questions. Thus, another aim of this paper is show how AI expands the scope of social choice ethics.

Before diving into the details, it should be explained that questions of standing, measurement, and aggregation must be answered by AI designers at the beginning of the social choice process—the questions cannot be delegated to the process. Social choice processes can consider issues of standing, measurement, and aggregation. However, how these issues are considered is determined by how the process was initially set up. In democracies, this is the issue of who gets to vote on who gets to vote, and how that vote is held. For example, in the United States and other countries, women were first enfranchised when men voted to enfranchise them. For this reason, AI designers cannot simply "let the AI figure it out". AI designers who use social choice ethics must make choices about standing, measurement, and aggregation.

Standing

The term "standing" comes from the legal context in which to have standing is to have the right to bring a case to court or otherwise participate in the case in some legally significant way. To

have legal standing, one must be a legal person—generally a human citizen of the jurisdiction in question—and one must have sufficient connection to the case to justify participation. While this is the origin of the term "standing" as it is used in this paper, the paper uses it in a slightly different way.

In this paper, to have standing is to have one's ethics included in a social choice process used to determine the ethics of an AI.[3] In social choice processes, those who have standing are those whose values are factored in. Alternatively, if a predetermined ethical view is used instead of a social choice ethics, then the individuals with standing are whoever decides which view to build into the AI. For example, if a person builds an AI on her or his own, and that person makes a unilateral, top-down decision on which ethics to build in, then that person is the only one who has standing for that AI. Biases can be introduced, all the more so because the demographics of AI tilt heavily towards males of certain backgrounds (Clark 2016). Taken in this context, the first major decision related to social choice ethics is the decision to use social choice ethics in the first place. Assuming this decision has indeed been made, a series of standing decisions must then be resolved.

To help clarify the issue of standing, let us take the concrete example of autonomous vehicles. Some driving decisions pose tradeoffs between vehicle occupants and other individuals. For example, should the vehicle travel faster, so as to minimize travel times (good for the occupants), or slower, so as to minimize energy consumption and environmental harms (good for almost everyone else)? One study found a 13% variation in per-mile energy efficiency of autonomous vehicles depending on how the vehicle is programmed (Mersky and Samaras 2016). Aggregated across a global fleet of vehicles, this is an enormous difference for energy and the environment—a difference that can depend on how the AI handles standing.

Autonomous vehicles may be designed to give their occupants the choice of how the vehicle should drive, especially if the occupants own the vehicle. This gives standing to the occupants but not to everyone else. Furthermore, sometimes the vehicle itself will need to make the choice, because sometimes occupants will neglect to choose: the vehicle needs default drive settings. An AI with social choice ethics would learn from the tendencies of whoever is setting its drive mode and make its own driving choices accordingly. Designing an AI to learn its driving values from its occupants denies standing to everyone else. This could lead the vehicle to travel faster and pollute the environment more, causing the rest of the world considerable harm.

With this example in mind, we turn now to the general questions of who or what gets standing.

Discussions of social choice AI ethics typically propose giving standing to some (often unspecified) portion of humanity. An example of this appears above in the quote "[Bottom-up ethics] attempts to train or evolve agents whose behavior emulates morally praiseworthy *human* behavior" (Allen et al. 2005:149; emphasis added). Similarly, Yudkowsky (2004:5) argues that "the initial dynamic should implement the coherent extrapolated volition *of humankind*" (emphasis added). But which humans are to be included? Martin (2017) considers whether AI ethics should be set by its designers, its users, or by human society as a whole.[4]

Should the antisocial psychopaths be included—the Hitlers and bin Ladens and serial killers and rapists and other "very bad people" of the world? These are people who might, for example, train autonomous vehicles to drive in especially dangerous ways. One could make a case for excluding them as unfit, just as convicted felons are disenfranchised in some democracies.

[3] This is similar to the "boundary problem" in democracy (Arrhenius 2005).

[4] Martin (2017) also considers having AIs set their own ethics or the ethics of other AIs; more on this below.

Indeed, if "large segments of humanity have base or evil preferences" (Bostrom 2014:217), then including them could foul up the entire social choice process, causing AIs to conduct horrific and unspeakable acts. But excluding the antisocial requires defining who they are. As Foucault (1961) documents, conceptions of madness and mental illness have changed dramatically over time. Current conceptions are thus historically contingent and not necessarily correct. The AI designer must choose which antisocial people, if any, to exclude from the social choice process.

Another challenging case is children. Children are routinely excluded from elections on grounds that they lack the intellectual and moral capacity to make reasoned judgments about whom to vote for. In the United States, for example, the minimum age for voting is 18. But about a quarter of the current human population is under age 15, which is a lot to exclude. Excluding them is arguably inherently unfair, and it could bias an AI's values in certain directions, for example because young people generally more accepting of lesbian, gay, bisexual, and transgender (LGBT) people (e.g., Pew 2017). Views about LGBT people may be insignificant for the AI inside autonomous vehicles, but they could be important for, say, AIs that chat with people on social media, such as Microsoft's notorious chatbot Tay (Gibbs 2016), or AIs in robots that enforce proper behavior in public, such as the Knightscope K5 security robot (Metz 2014).[5]

Excluding children may not be hugely consequential as long as their parents have standing. Parents invest heavily in the success of their children, both privately and publicly (i.e., through governments and community groups). However, future generations are at greater risk. The moral psychology of time preference and temporal discounting typically finds that people value future events and future generations less than the present (Frederick et al. 2002). Likewise, philosophers sometimes consider that future generations may have no value at all (Arrhenius and Rabinowicz 2015). For example, Marglin (1963:97) writes "I consider it axiomatic that a democratic government reflects only the preferences of the individuals who are presently members of the body politic". Thus, unless future generations are given standing at the outset of a social choice process, there is a strong likelihood that their interests would be given little consideration by the resulting AI. (There is also the question of how an AI can learn the values of future generations, but this is a matter of measurement, not standing.)

It is difficult to overstate how much is at stake with standing for future generations. The population of future generations could be extremely large, vastly dwarfing the present population. Earth will remain habitable for on the order of a few billion more years (O'Malley-James et al. 2014), and the rest of the universe will remain habitable for much longer (Adams 2008). If AI designers only give standing to the present generation, they risk biasing outcomes against this astronomically large future. For example, this could mean gluttonous consumption of nonrenewable natural resources and pollution of global ecosystems by the present generation (such as in autonomous vehicles designed to travel quickly instead of energy efficiently), with future generations left to struggle through the ensuing mess. An AI that is programmed to only give standing to the present generation could deliver an extreme intergenerational injustice.

These are the ethical issues faced if standing is given to some portion of humanity, as is commonly prescribed in the AI ethics literature. However, this presupposes that standing should be given *only* to humanity. But humans are not the only individuals that could merit standing. Standing has been considered for non-human animals (Sunstein 2000), plants and ecosystems (Stone 1972; Hannon 1998), abiotic environments (Rolston 1986), extraterrestrials (Cockell

[5] Tay was programmed to learn from (and thus give standing to) Twitter users who interact with it, which quickly devolved into deviance and obscenity as Twitter users taught it to misbehave. Microsoft has since been wrestling with the question of how to give standing to a more appropriate mix of people.

2007), technologically enhanced "posthumans" (Buchanan 2009), and AIs (Hubbard 2011). Each of these entities poses its own set of ethical challenges for the issue of standing in an AI social choice process.

Consider nonhuman animals. The more advanced nonhuman animals (e.g., other primates) have cognitive abilities approaching those of humans, a fact that has been the basis of attempts to grant them some legal standing, such as the Great Ape Project. Sentience appears to decline gradually across the animal kingdom—for example, research on the sentience of fish is inconclusive (Rose et al. 2014), while some research suggests the sentience of invertebrate crayfish (Fossat et al. 2014). Failing to give standing to sentient animals would risk mass atrocities committed against them—atrocities on par with or worse than the widespread and horrific abuse of livestock animals in factory farms today (e.g., Anomaly 2015).

Robots are already managing livestock in some locations (Klein 2016), and it may only be a matter of time before livestock is managed primarily by robots. Working in factory farms is dangerous and unpleasant for humans and also somewhat repetitive, making it an ideal candidate for automation. However, if the robots are programmed to learn their values from humans only, then they could perpetuate the violence against livestock animals. Indeed, robots could even make the violence worse if the empathy of human farmhands is replaced by the callousness of farm management. Alternatively, if the robots learn values from humans and livestock alike, then they may find ways to treat the livestock animals better. In the best-case scenario, this could drastically reduce or even eliminate the mass abuse of livestock animals in factory farms while maintaining or even improving the supply of livestock products to humans. However, the robots may find some tradeoffs unavoidable, in which case giving standing to livestock could worsen the supply of livestock products—in the extreme case, the robots could even conclude that the livestock should be set free.

Similar logic applies to plants, ecosystems, and abiotic environments: AIs designed to learn values from humans only could end up acting for human benefit at the expense of anyone or anything else. The AI designer must choose sides on these issues when choosing who or what gets standing in the social choice process.

A different logic applies to posthumans because these entities do not currently exist. Consider the extreme case of a superpowerful ASI that has taken over the world and is managing it according to the CEV of humanity. It is possible—indeed it may be probable—that humanity ultimately prefers remaining human and not being "enhanced" into posthumans. Indeed, a recent public opinion poll found a majority of Americans indicating opposition to a range of human "enhancement" technologies (Funk et al. 2016). In this case, whether posthumans ever come into existence could depend on whether the ASI (or its precursor seed AI) is designed to give standing to posthumans. If posthumans have standing, then the desire of posthumans to come into existence could be balanced against the desire of humans to stay human.[6]

Finally, there is the AI itself. Hubbard (2011) argues that if AIs possess the attributes that are considered essential for legal standing, then they should have it. Hubbard proposes three criteria for an AI meriting legal standing: complex interaction skills, self-consciousness, and the ability to pursue group benefits. If anything, standards for moral standing in a social choice process should be lower than this, because legal standing implies a certain ability to function in courts and other legal contexts, whereas moral standing does not. Martin (2017) considers that sufficiently advanced AIs could gain capacity for ethical reasoning that matches or even exceeds

[6] There is a certain irony that some proponents of CEV speak in terms of giving standing only to humanity but also favor a transition to posthumanity (e.g., Bostrom 2008).

that of humans, making them independent ethical agents. In contrast, Yampolskiy (2013:393) argues that AIs that merit standing should not be built in the first place, and instead that "machines should be inferior by design; they should have no rights and should be expendable as needed, making their use as tools much more beneficial for their creators".

Once again, much is at stake. To abstain from building certain types of AIs could significantly limit the extent to which the benefits of AI could be realized. AIs that would merit standing may be some of the most sophisticated and capable AIs; abstaining from building them could be an especially large loss. Alternatively, if humans deny standing for sophisticated AIs, this puts core human values at risk. As Hubbard (2011) argues, to deny standing to entities that are at least as deserving of it as we are cuts against the liberal, secular, rationalist, and egalitarian foundations of democratic human society.

But AIs pose a novel and sizable complication. As Yampolskiy (2012) explains, AIs can be readily copied in large numbers. If AIs have standing, they could easily drown out the human population or any other biological population. Some existing nonhuman entities raise the same complication, for example, if standing were given to bacteria, which massively outnumber humans and other large organisms. However, only AI can combine massive population sizes with cognitive sophistication equal or greater to humans. It would arguably be unjust to disenfranchise AIs simply on the grounds of getting outvoted—such a situation would be analogous to the apartheid of South Africa, in which the white minority disenfranchised the black majority. However, to grant AIs standing could mean letting them control the social choice process. Ultimately, whether they would control it depends on the details of measurement and aggregation.

Measurement

Measurement in this context refers to the process through which an individual's ethical views are identified for inclusion in the social choice process. In a typical democracy, measurement is done via voting. In capitalism, measurement is done via buying and selling. Measurement can also be done via observing behavior, as in the economics concept of "revealed preference", via surveys and interviews, such as in public opinion polling, or even via brain imaging, such as in neuropsychology research.

If human beings—or whatever else was being measured—had one single, consistent set of ethical views, then measurement would be a relatively simple issue. Each measurement procedure would yield more or less the same results. The only challenge would be obtaining results in sufficient detail to inform the issue in question. This is a nontrivial challenge, but it pales in comparison to the actual challenge of measurement. Humans do not have a single, consistent set of ethical views, and different measurement procedures can yield different answers to the same ethical question.

To take one example, consider the ethical issue of time preference. This concerns how people value present gains and losses relative to future gains and losses. Frederick et al. (2002, Table 1) review 41 different studies that measure human time preference using either behavior observation or survey methods. The studies show that humans discount future gains and losses at rates ranging from -6% to +∞%. Some of this variation is attributable to differences in the nature of the gains and losses being evaluated (e.g., money vs. health), but much of it appears to be due to differences in the procedures used to measure time preference.

Discount rates from -6% to +∞% are an extremely wide range. Negative discount rates imply that future gains and losses are more valuable than present ones. A discount rate of +∞% implies

that future gains and losses hold zero value. Thus, depending on the choice of measurement procedure, an AI could end up valuing future gains and losses a lot or not at all. For example, this could be the difference between an autonomous vehicle AI traveling faster (if future gains and losses are not valued at all) or instead more energy efficiently (if future gains and losses are valued a lot). Or, for a superpowerful ASI, this could be the difference between lavishing resources on the present generation, future be damned, or preparing for the distant future, even if that requires present sacrifice.

The wide range of time preferences that have been measured in humans is one instance of a broader tendency for humans to show inconsistent and often incoherent ethics views. Even moral philosophers struggle with inconsistency, as evidenced by their various "impossibility theorems" in which they find it impossible to craft a moral theory that meets a series of seemingly plausible moral intuitions (e.g., Arrhenius 2011). For this reason, an effort to measure an individual's ethics views is likely to yield only one particular variation on his or her views.

This diversity of an individual's values poses a thorny challenge for social choice AI ethics. As an example, consider an environmentalist who also happens to be a fast driver. This may seem hypocritical, since driving fast is worse for the environment—perhaps this person just thinks that *other people* should drive slower. However, it is often observed that people disagree *with their own behavior* on these sorts of matters (Stone and Fernandez 2008). In such a situation, what is a social choice AI to do? The AI in an autonomous vehicle could measure this person's ethics by observing her driving behavior, in which case the AI would end up traveling quickly, or by asking her how she thinks she should drive, in which case the AI would end up traveling energy efficiently. Both measurement approaches are technologically simple to implement: one involves recording driving data and the other involves a simple questionnaire. Either approach could lay some claim to measuring the person's "true" view.

One option available to AI designers is to select the measurement approach that they believe delivers the better result. This is analogous to the concept of libertarian paternalism or "nudges" (Thaler and Sunstein 2008; in the context of robotics, see Borenstein and Arkin 2016). In this context, to nudge someone is to structure their decisions to make it more likely that they make the "better" decision. For example, states can make it the default option that people are organ donors, which increases the rate of organ donors. This is libertarian because anyone is still free to opt out of organ donation, and it is paternalistic because it structures the decision so that people tend to make the "better" decision of being an organ donor.

Analogously, AI designers could select the measurement option that they think delivers the better result. Indeed, designers could even give people the choice of measurement option, in which case designers could influence results by choosing which measurement option is the default. But which to choose? Is it better for people to travel faster or more energy efficiently? Traveling faster may be better for that person, while traveling more energy efficiently may be better for society as a whole. This is a difficult ethics question in its own right, and it is exactly the sort of question that AI designers seek to avoid by choosing social choice ethics. Because humans display multiple ethics views, the choice of measurement approach for social choice ethics can require AI designers to make unilateral ethical choices.

In response to the seeming incoherence of human ethics, some AI ethicists call for measuring an idealized version of human ethics instead of what is observed in actual humans. For example, Muehlhauser and Helm (2012:114) posit that an advantage of CEV for ASI is that it would deliver "what a person would want after reaching reflective equilibrium with respect to his or her values, rather than merely what each person happens to want right now", and that this "may

dissolve the contradictions within each person's current preferences". This follows calls in moral philosophy for the use of "idealized preferences" (e.g., Harsanyi 1996; Ng 1999; and references in Muehlhauser and Helm 2012). The hope is that an ASI would be able to figure out which ethics any given human would really want if he or she was able to think the matter through as carefully as an ASI could, and that following this idealized ethics would yield better results.

It should indeed be expected that a human's idealized ethics would differ from his or her original ethics. This is seen, for example, in differences between the ethics views held by moral philosophers and those held by the lay public. However, it may be a matter of opinion whether the process of idealization tends to deliver ethics that are in some neutral sense "better". Moral philosophers are known to hold a range of idiosyncratic and potentially catastrophic views, as follows from their inclination to following specific moral intuitions to extreme logical conclusions, wherever that may lead. In some cases, perhaps many cases, it may appear that their pre-moral philosophy, pre-idealization, common sense moral views would be better.

As an illustrative example, consider the view of Benatar (2006) that it is wrong for humans to procreate. Benatar's view is the logical conclusion of his carefully considered belief that it is bad to harm people by bringing them into existence but it is not good to benefit people by bringing them into existence. Put differently, procreation can be bad but it cannot be good. The extreme logical conclusion of this view is that humanity should die out after the present generation. Benatar's views may well constitute a minority position among moral philosophers,[7] but this is beside the point. The point is that at least in this one case, the process of idealization would appear to deliver worse results. This presumes that, prior to putting more thought into his moral philosophy, Benatar did not wish for humanity's extinction, which seems likely: few people with "common sense" moral views would wish such a thing. Therefore, for cases like Benatar, measuring idealized values instead of "common sense" initial values can deliver worse results. Whether these cases are outliers or the norm is an empirical question for which an answer may not yet exist. Regardless, it remains the case that the decision to measure ethics using idealized human values is not an ethical "no brainer" but instead is a complex and contentious decision on which much depends.

If measurement is to use an idealization, there remains the question of which idealization process to use. One variable is whether idealization proceeds via solitary reflection or group deliberation. Yudkowsky (2004:7) calls for group deliberation for an idealization of the ethics of humans if they "had grown up farther together" instead of the ethics of "the person you'd become if you made your decisions alone in a padded cell". Yudkowsky posits that the social interactions of group deliberation would provide "social forces contributing to niceness".

But there is reason to doubt that group deliberation would tend to yield better results. Groups can suffer from such pathologies as groupthink and in-group favoritism (Baron 2005; Balliet et al. 2014). Meanwhile, solitary reflection can do well. Indeed, much of the most distinguished moral philosophy produced across human history was produced by individuals working largely in solitude. AI designers who choose idealized measurement must further make the nontrivial choice between idealization via solitary reflection or group deliberation. Other idealization choices include how long to let the idealization process play out and how to handle instances in which idealization leaves some contradictory incoherences intact.

The various measurement issues discussed thus far pertain mainly to present-generation humans. While the issues are substantial, they pale in comparison to the issues presented by other beings.

[7] For an argument against Benatar's views, see Baum (2008).

For starters, how would a social choice AI measure the values of sentient nonhuman animals? Some standard human measurement techniques could not be used—for example, cows and frogs cannot meaningfully respond to survey questionnaires. Other techniques could be used—nonhuman animals' behavior can be observed, and their brains can be scanned with imaging technologies—but interpreting the results is a major challenge. Human AI designers do not know what it is like to be a cow or a frog, so they must make sweeping assumptions in programming an AI to infer cow or frog ethics from any given measurement technique.

What about nonsentient animals, plants, and other living organisms, and nonliving, abiotic matter? The case for giving them standing is weaker, because they (presumably) cannot suffer or experience any sort of pleasure. However, if they are given standing, this presents the challenge of measuring their ethics views. Yet there is no clear procedure for extrapolating the values of nonsentient beings. They have no brains to scan. Some of them (e.g., rocks) have no (or barely any) behavior to observe. The social choice AI is left without the standard empirical techniques of measurement. The same holds for future generations of humans or posthumans. These beings do not yet exist, and so they cannot be surveyed, observed, scanned, or measured via any standard empirical technique.

When empirical techniques are inadequate or unavailable, an alternative (perhaps the only alternative) is to use measurement by proxy. The essence of proxy measurement is to have an available being attempt to describe the views that an unavailable being would have if he, she, or it was available to be measured. For example, in some human voting systems, a person who cannot attend a vote can give his or her proxy vote to a trusted associate. The associate then gets to vote on behalf of the absentee person. Similarly, proxy measurement is at the heart of proposals to build representation of future generations into present democracies (e.g., Tonn 1996). Likewise, attempts to measure the interests of nonsentient beings ultimately come down to humans seeking to infer what those beings really "want".

Therefore, if standing is given to nonsentient or future beings, a social choice AI may need some procedure for measuring their ethics via proxy. This raises its own set of questions: Who to give proxy to? How to measure their proxy? These are difficult questions, but they at least resemble the more general questions of standing and measurement and thus are both familiar and tractable. But the stakes are quite large. Nonsentient and/or future beings could vastly outnumber present sentient beings. Therefore, different handlings of proxy could yield massively different results, not just for each individual's ethics measurement but for the entire social choice process. How the entire social choice process would play out would also depend on the technique used for aggregation.

If AIs have standing, they pose different measurement issues—or rather, measurement opportunities. It may be possible to infer an AI's ethical views from its source code. This is not necessarily a straightforward task: the relevant code could be large and complex. But the inference may nonetheless be feasible, perhaps with the assistance of an AI designed for the task. In this case, it may be possible to achieve a high degree of precision in measurement, at least for this measurement technique. Other techniques may still yield other descriptions of AIs' ethical views, just as different techniques can yield different descriptions for humans and other types of entities.

Aggregation

In a social choice process, no matter who has standing and how their ethics views are measured, once the measurements are in, the final step is to aggregate these individual ethics views into a

single societal ethics view. For AI, a single view is needed to determine how the AI is to make decisions. In other words, the aggregate ethics view is what the AI is supposed to use to decide how it should behave.

Classic social choice theory recognizes that how to aggregate is not always straightforward. The simplest case occurs when three individuals (labeled X, Y, and Z) each have a different ranking of three choice options (labeled A, B, and C), as shown in Table 1. When the vote is between A and B, A wins; between A and C, C wins; between B and C, B wins. Thus it would appear that the group prefers A to B, B to C, and C to A. In other words, the preferences are intransitive. This case is known as the voting paradox or Condorcet's paradox, having been documented by Condorcet (1785).

	First Choice	Second Choice	Third Choice
Individual X	A	B	C
Individual Y	B	C	A
Individual Z	C	A	B

Table 1. A hypothetical set of choice rankings from three different individuals.

One solution to Condorcet's paradox is to also consider the strength of individual preferences for each choice option. For example, perhaps X and Y have only a slight preference for B over C, yet Z has a strong preference for C, in which case C might be the best option. This solution is hinted at in Yudkowsky's (2004:8) argument that "A minor, muddled preference of 60% of humanity might be countered by a strong, unmuddled preference of 10% of humanity." However, this imposes a greater burden on measurement by requiring the strength of ethics views and not just ranked orderings. It also makes aggregation more demanding because it requires being able to compare "ethics view strength" between individuals.

The inter-individual comparison of ethics view strength is complicated by the fact that people often hold "protected" or "sacred" values that they believe should not be traded off against other values, much like deontological rules (Ginges et al. 2007; Ritov and Baron 1999). In order to form a single ethics view, an aggregation process must resolve conflicts between different people's protected values. For example, suppose one person believes that people should be categorically prohibited from putting excessive pollution into the environment, while another believes that it should be categorically prohibited to deny anyone the right to drive however they would like. An AI's attempt to measure the strength of these views could find that they have infinite strength pointing in opposite directions. The AI's aggregation process requires some means of resolving such disagreements. One option is to insist that no one's views can have infinite strength, but this imposes a distortion on the views of people who hold protected values.

Another aggregation challenge comes from aggregation procedures that cluster individuals into distinct groups. In modern democracies, voters are commonly clustered into geographic districts. How elections turn out can depend heavily on how district lines are drawn. This is seen in the concept of "gerrymandering", in which unintuitive district lines are drawn so as to maximize advantage for a political party, or more generally in the modifiable areal unit problem, in which different demarcations of spatial units yield different results (Openshaw 1983). Similar quirks of clustering underlie the fact that a US Presidential candidate can win the popular vote but lose the electoral college and hence the overall election,[8] as well the fact that each US state has two Senators regardless of its population, giving proportionately more Senate representation

[8] This happened in 2000 and 2016, when Al Gore and Hillary Clinton, respectively, received more votes from individual voters but George W. Bush and Donald Trump, respectively, received more votes in the electoral college.

to residents of small states. The AI designer may be tempted to reject clustering and stick to a strict "one individual, one vote" principle. However, this can facilitate "tyranny of the majority" at the expense of minority rights and interests. Indeed, it was to protect the interests of small states that US Senate representation was designed with two Senators per state, as per the 1787 Connecticut Compromise (Yazawa 2016).

Many AIs with bottom-up ethics are designed in a rather different fashion, updating their ethics each time someone interacts with it. This is essentially a "vote early and often" principle. Hence, for example, Microsoft's chatbot Tay was trained to speak vulgarities even though many people did not want this—Tay's training was driven mainly by the "vocal minority" who repeatedly engaged with it. In this case, the vocal minority aggregation scheme yielded pathological results. A "one individual, one vote" aggregation procedure would have given everyone equal say in Tay's training and may have yielded better results.[9]

Further complications come when nonhumans have standing. Do their views count the same as humans? A "one individual, one vote" principle may make little sense, for example, if standing is given to the massive population of bacteria. An argument can be made for giving nonhuman animals and other living beings less of a vote than humans because humans are more cognitively sophisticated. However, the same argument would suggest that humans should have less of a vote if standing is granted to more sophisticated beings, such as posthumans or ASIs.

Aggregation for AIs poses an additional complication that derives from AI reproduction. An AI can be copied extensively; if each copy has standing, and each has equal say in the aggregation process, then this can drown out everyone else's vote. An aggregation process could face a dilemma between disenfranchising AIs or disenfranchising everyone else. A similar issue is posed by future generations, given their astronomically large potential population.

It should be noted that these various aggregation challenges only exist to the extent that different individuals have different ethics views. Where measurement yields consensus, aggregation is irrelevant. Some AI theorists posit that superintelligent beings all tend towards the same universal set of values (Goertzel 2016). Whether or not this is the case remains to be seen and cannot a priori be counted on. Without consensus, issues of aggregation must be addressed.

Discussion and Conclusion

Having considered issues of standing, measurement, and aggregation in detail, let us now revisit the procedural and abstention justifications for social choice ethics in AI. (Recall that the wise crowd justification requires a predetermined ethical view.)

The procedural justification maintains that individuals should have a say in decisions that affect them instead of AI designers imposing their own views on them. But this ideal leaves open questions about standing, measurement, and aggregation; AI designers cannot avoid imposing their own views on these questions. At most, social choice ethics reduces the extent to which AI designers impose their own views. Furthermore, it is often difficult or even impossible to give everyone a say. When AI has global effects, such as in autonomous vehicles affecting global warming or ASI that takes over the world, a massive number of individuals are affected, most of whom may not have direct interaction with the AI. When AI affects future generations of humans, future entities such as posthumans or new forms of AI, or less intelligent nonhuman entities, giving them a say becomes inherently difficult. Thus, while the procedural justification

[9] There is no indication that Tay was designed with bottom-up ethics in mind, but the net result is the same in that Tay acquired its principles for behavior via input from the people it interacted with.

may capture an admirable ethical sentiment, it provides weak support and limited guidance for AI ethics design.

The abstention justification maintains that by using social choice ethics, AI designers can abstain from making ethics decisions and instead let the AI figure out what society's aggregate views are. AI designers may take comfort in this insofar as they are unaccustomed or disinclined to think about ethics. However, use of social choice ethics requires decisions about standing, measurement, and aggregation, which can in turn substantially affect outcomes of AI decisions. There is simply no way for AI designers to successfully abstain from ethical decision making. Furthermore, the ethical issues in standing, measurement, and aggregation are numerous and profound. Using social choice ethics requires much more than a minimal amount of ethical thought. The abstention justification fails.

AI designers are left with two options. They can use social choice ethics and make the profound ethical decisions of standing, measurement, and aggregation. They may justify their use of social choice ethics with a weakened form of the procedural justification; their decisions regarding standing, measurement, and aggregation would require separate justification. Or, they can abandon social choice ethics in favor of a predetermined ethical view. A predetermined view would require its own justification, but this may not be much different than justifying decisions of standing, measurement, and aggregation.

Indeed, depending on the details, the distinction between social choice ethics and predetermined view ethics may be insignificant. For example, consider the question of standing for future generations. This is broadly similar to the question of whether to count the welfare of future generations in welfarist utilitarianism. An AI may make the same decisions under either social choice ethics or welfarist utilitarianism as long as it does (or does not) count future generations. For example, an autonomous vehicle that counts future generations may drive in an energy efficient fashion to reduce its greenhouse gas emissions, while an ASI may embark on long-term projects at the expense of short-term splurges. Or, consider the question of whether to give standing to AIs. A social choice ethics may give AIs standing if they have the capacity to form ethical views. Similarly, welfarist utilitarianism may count AIs if they have the capacity to experience pleasure and pain. In both cases, the moral calculus could be dominated by the overwhelmingly large number of AIs that can be created by copying AI software ad infinitum.

In these and other cases, the distinction between social choice ethics and predetermined view ethics is unimportant. Thus, proposals such as CEV and bottom-up ethics do not actually do much to resolve the important decisions to be made in the design of AI ethics. These are inherently decisions that must be made by AI designers—one cannot "let the AI figure it out", because the decisions concern how the AI would figure it out. Focus should likewise be on the important decisions, not on whether AI uses some sort of social choice ethics.

Acknowledgements

Anders Sandberg provided helpful discussion for the development of this paper. Tony Barrett and two anonymous reviewers provided helpful feedback on earlier drafts. Any errors or shortcomings in the paper are the author's alone. Work on this paper was funded in part by Future of Life Institute grant number 2015-143911. The views in this paper are the author's and are not necessarily the views of the Future of Life Institute or the Global Catastrophic Risk Institute.

References

Adams FC (2008) Long-term astrophysical processes. In Bostrom N, Ćirković MM (eds), Global Catastrophic Risks. Oxford University Press, Oxford, pp. 33-47.

Allen C, Varner G, Zinser J (2000) Prolegomena to any future artificial moral agent. Journal of Experimental & Theoretical Artificial Intelligence 12:251-261

Allen C, Smit I, Wallach W (2005) Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics & Information Technology 7(3):149-155.

Anomaly J (2015) What's wrong with factory farming? Public Health Ethics 8(3):246-254.

Arrhenius G (2005) The boundary problem in democratic theory. In Tersman F (ed), Democracy Unbound: Basic Explorations I. Filosofiska Institutionen, Stockholm, pp.14-29.

Arrhenius G (2011) The impossibility of a satisfactory population ethics. In Dzhafarov E, Lacey P (eds), Descriptive and Normative Approaches to Human Behavior. World Scientific, Singapore, pp.1-26.

Arrhenius G, Rabinowicz W (2015) The value of existence. In Hirose I, Olson J (eds), The Oxford Handbook of Value Theory. Oxford University Press, Oxford, pp.424-443.

Arrow KJ (1951) Social Choice and Individual Values. Wiley, New York.

Balliet D, Wu J, De Dreu CKW (2014) Ingroup favoritism in cooperation: A meta-analysis. Psychological Bulletin 140(6):1556–1581.

Baron RS (2005) So right it's wrong: Groupthink and the ubiquitous nature of polarized group decision making. Advances in Experimental Social Psychology 37:219-253.

Baum SD (2008) Better to exist: A reply to Benatar. Journal of Medical Ethics 34(12):875-876.

Baum SD (2009) Description, prescription and the choice of discount rates. Ecological Economics 69(1):197-205.

Benatar D (2006) Better Never to Have Been: The Harm of Coming Into Existence. Oxford University Press, Oxford.

Bohannon J (2015) Fears of an AI pioneer. Science 349(6245):252.

Borenstein J, Arkin R (2016) Robotic nudges: The ethics of engineering a more socially just human being. Science and Engineering Ethics 22(1):31-46.

Bostrom N (2008) Why I want to be a posthuman when I grow up. In Gordijn B, Chadwick R (eds), Medical Enhancement and Posthumanity. Springer, Berlin, pp. 107-136.

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

Brandt F, Conitzer V, Endriss U, Lang J, Procaccia AD (2015) Handbook of Computational Social Choice. Cambridge University Press, Cambridge, UK.

Buchanan A (2009) Moral status and human enhancement. Philosophy & Public Affairs 37(4):346-381.

Clark J (2016) Artificial intelligence has a 'sea of dudes' problem. Bloomberg, June 23.

Cockell CS (2007) Originism: Ethics and extraterrestrial life. Journal of the British Interplanetary Society 60:147-153.

Condorcet M de (1785) Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix. L'imprimerie Royale, Paris.

Fossat P, Bacqué-Cazenave J, De Deurwaerdère P, Delbecque JP, Cattaert D (2014). Anxiety-like behavior in crayfish is controlled by serotonin. Science 344(6189):1293-1297.

Foucault M (1961) Folie et Déraison: Histoire de la Folie à l'âge Classique. Plon, Paris.

Frederick S, Loewenstein G, O'donoghue T (2002). Time discounting and time preference: A critical review. Journal of Economic Literature 40(2):351-401.

Funk C, Kennedy B, Podrebarac Sciupac E (2016) U.S. public wary of biomedical technologies to 'enhance' human abilities. Pew Research Center, July 26.
Gibbs S (2016) Microsoft's racist chatbot returns with drug-smoking Twitter meltdown. The Guardian, 30 March.
Ginges J, Atran S, Medin D, Shikaki K (2007) Sacred bounds on rational resolution of violent political conflict. Proceedings of the National Academy of Sciences 104(18):7357-7360.
Goertzel B (2016) Infusing advanced AGIs with human-like value systems: Two theses. Journal of Evolution & Technology 26(1):50-72.
Hannon B (1998) How might nature value man? Ecological Economics 25:265–279.
Harsanyi JC (1996) Utilities, preferences, and substantive goods. Social Choice and Welfare 14(1):129-145.
Holbrook D (1997) The consequentialistic side of environmental ethics. Environmental Values 6:87–96.
Hubbard FP (2011) 'Do androids dream?': Personhood and intelligent artifacts. Temple Law Review 83:405–441.
Klein A (2016) Robot ranchers monitor animals on giant Australian farms. New Scientist, May 20.
Lin P (2016) Why ethics matters for autonomous cars. In Maurer M, Gerdes JC, Lenz B, Winner H (eds), Autonomous Driving: Technical, Legal and Social Aspects. Springer, Berlin, pp.69-85.
Marglin SA (1963) The social rate of discount and the optimal rate of investment. Quarterly Journal of Economics 77(1):95-111.
Martin D (2017) Who should decide how machines make morally laden decisions? Science and Engineering Ethics 23(4):951-967.
Mersky AC, Samaras C (2016) Fuel economy testing of autonomous vehicles. Transportation Research Part C: Emerging Technologies 65:31-48.
Metz R (2014) Startup Knightscope is preparing to roll out human-size robot patrols. MIT Technology Review, November 13.
Muehlhauser L, Helm L (2012) Intelligence explosion and machine ethics. In Eden A, Søraker J, Moor JH, Steinhart E (eds), Singularity Hypotheses: A Scientific and Philosophical Assessment. Springer, Berlin, pp.101-126.
Ng YK (1990) Welfarism and utilitarianism: A rehabilitation. Utilitas 2(2):171-193.
Ng YK (1999) Utility, informed preference, or happiness: Following Harsanyi's argument to its logical conclusion. Social Choice and Welfare 16(2):197-216.
O'Malley-James JT, Cockell CS, Greaves JS, Raven JA (2014) Swansong biospheres II: The final signs of life on terrestrial planets near the end of their habitable lifetimes. International Journal of Astrobiology 13:229-243.
Openshaw S (1983) The modifiable areal unit problem. Geo Books, Norwich.
Pew (2017) Changing attitudes on gay marriage. Pew Research Center, 26 June
Picard R (1997) Affective Computing. MIT Press, Cambridge, MA.
Ritov I, Baron J (1999) Protected values and omission bias. Organizational Behavior & Human Decision Processes 79(2):79-94.
Rolston H III (1986) The preservation of natural value in the solar system. In Hargrove EC (ed), Beyond Spaceship Earth: Environmental Ethics and the Solar System. Sierra Club Books, San Francisco, pp. 140-182.

Rose JD, Arlinghaus R, Cooke SJ, Diggles BK, Sawynok W, Stevens ED, Wynne CDL (2014). Can fish really feel pain? Fish & Fisheries 15(1):97-133.
Schienke EW, Tuana N, Brown DA, Davis KJ, Keller K, Shortle JS, Stickler M, Baum SD (2009) The role of the NSF Broader Impacts Criterion in enhancing research ethics pedagogy. Social Epistemology 23(3-4):317-336.
Schienke EW, Baum SD, Tuana N, Davis KJ, Keller K (2011) Intrinsic ethics regarding integrated assessment models for climate management. Science & Engineering Ethics 17(3):503-523.
Stone C (1972) Should trees have standing? Toward legal rights for natural objects. Southern California Law Review 45:450–501.
Stone J, Fernandez NC (2008) To practice what we preach: The use of hypocrisy and cognitive dissonance to motivate behavior change. Social & Personality Psychology Compass 2(2):1024-1051.
Sunstein CR (2000) Standing for animals. UCLA Law Review 47(5):1333-1368.
Tarleton N (2010) Coherent Extrapolated Volition: A Meta-Level Approach to Machine Ethics. The Singularity Institute.
Thaler R, Sunstein C (2008) Nudge: Improving decisions about health, wealth, and happiness. Yale University Press, New Haven.
Tonn B (1996) A design for future-oriented government. Futures 28(5):413-431.
Wallach W, Allen C (2008) Moral machines: Teaching robots right from wrong. Oxford University Press, Oxford.
Wallach W, Allen C, Smit I (2008) Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. AI & Society 22(4):565-582.
Yampolskiy RV (2013) Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In Müller VC (ed), Philosophy and Theory of Artificial Intelligence. Springer, Berlin, pp. 389-396.
Yazawa M (2016) Contested Conventions: The Struggle to Establish the Constitution and Save the Union, 1787-1789. Johns Hopkins University Press, Baltimore.
Yudkowsky E (2004) Coherent extrapolated volition. The Singularity Institute, San Francisco.