Manipulating Aggregate Societal Values to Bias AI Social Choice Ethics

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

AI and Ethics, forthcoming, DOI 10.1007/s43681-024-00495-6. This version 13 June 2024.

Abstract
Work on AI ethics often calls for AI systems to employ social choice ethics, in which the values of the AI are matched to the aggregate values of society. Such work includes the concepts of bottom-up ethics, coherent extrapolated volition, human compatibility, and value alignment. This paper describes a major challenge that has previously gone overlooked: the potential for aggregate societal values to be manipulated in ways that bias the values held by the AI systems. The paper uses a "red teaming" approach to identify the various ways in which AI social choice systems can be manipulated. Potential manipulations include redefining which individuals count as members of society, altering the values that individuals hold, and changing how individual values are aggregated into an overall social choice. Experience from human society, especially democratic government, shows that manipulations often occur, such as in voter suppression, disinformation, gerrymandering, sham elections, and various forms of genocide. Similar manipulations could also affect AI social choice systems, as could other means such as adversarial input and the social engineering of AI system designers. In some cases, AI social choice manipulation could have catastrophic results. The design and governance of AI social choice systems needs a separate ethical standard to address manipulations, including to distinguish between good and bad manipulations; such a standard affects the nature of aggregate societal values and therefore cannot be derived from aggregate societal values. Alternatively, designers of AI systems could use a non-social choice ethical framework.

Keywords: AI ethics, social choice ethics, bottom-up ethics, coherent extrapolated volition, human compatibility, value alignment

1. Introduction

In ethics, as with engineering and other domains, it is important to study ways in which certain schemes can be abused or otherwise go wrong. This is needed for troubleshooting purposes, or alternatively to inform decisions to pursue alternative schemes. Such work is sometimes called red teaming or, in the context of cybersecurity, "white hat" ethical hacking. This paper follows in that tradition.

The paper specifically serves to troubleshoot a major approach to AI ethics in which the ethical values of AI systems are to be matched to some aggregate of the ethical values held by society as a whole. This approach goes under various names including bottom-up ethics [1], coherent extrapolated volition [2], human compatibility [3], social choice ethics [4], and value alignment [5]. (Alternative interpretations of some of these terms are explained below.) For succinctness, this paper will use the term social choice ethics to indicate all of these concepts, recognizing that each of them has at its core the ethics of social choice as developed in economics, philosophy, and political science. At this time, actual AI social choice systems are uncommon and used mainly for research. Nonetheless, social choice is a common concept for what ethics to build into AI systems, especially (but not exclusively) for hypothetical future AI systems with advanced capabilities. The stakes here can be quite high, and so these AI ethics frameworks merit careful scrutiny, including to identify their potential flaws.

This paper surveys the ways in which the aggregate societal values can be manipulated so as to bias the ethical frameworks used by AI social choice systems. For purposes of this paper, a manipulation is any action that alters aggregate societal values in a way that results in a social choice being made in a way that is more desirable from the perspective of the actor executing the manipulation. Some prior literature has focused on specific forms of manipulation, such as tactical voting, in which individuals intentionally misrepresent their own views to achieve better outcomes [6-7], and actions to alter the values held by other individuals [8]. One contribution of this paper is to present a more comprehensive taxonomy of social choice manipulation; another contribution is to show how this wider range of manipulations is of relevance to AI social choice ethics.

The paper does not address the use of AI systems to manipulate human behavior outside the context of AI social choice. For example, AI is in active use to manipulate consumer behavior through digital marketing. This form of manipulation raises important ethical issues [9]. However, this is not a manipulation of social choice ethics used by AI systems because it does not seek to influence the values used by the AI systems; indeed, the AI systems may not even use social choice ethics, and instead may pursue some other goal such as increasing business sales. It therefore falls outside the scope of the paper.

To help motivate the discussion, here is an illustrative example of AI social choice manipulation. Suppose actor A_1 wants an AI social choice system to follow ethical framework E_1. To bias the AI system into adopting E_1, A_1 could seek to kill everyone who supports ethical frameworks $E_2...E_N$. With only supporters of E_1 remaining, the AI system would find that E_1 is the aggregate values held by society, and the AI system would therefore act according to E_1. Very arguably, the act of A_1 killing supporters of $E_2...E_N$ could be unethical. However, social choice frameworks do not necessarily account for this sort of behavior, such as if the AI system is designed to *follow the aggregate values of everyone currently alive*. Unless an AI social choice framework does account for such behavior, AI systems may be biased by the behavior and indeed may even incentivize it. The potential for AI social choice ethics to be biased by manipulating aggregate societal values is a major challenge that has not gotten significant attention in the literature (as reviewed below).

After providing further background and prior literature on AI social choice ethics (Section 2), the paper surveys ways in which aggregate societal values can be manipulated (Sections 3-4) and how manipulation can be addressed (Section 5).

2. Background and Prior Literature

2.1 Social Choice Ethics

Social choice ethics refers to any ethical framework in which decisions are evaluated according to an ethical framework that is derived from some aggregation of a set of individual ethical frameworks held by some population of individual moral subjects[1]. The study of social choice traces to the political theory of de Condorcet [10] and, in the modern era, Arrow [11]. Modern scholarship works at the interface of economics, political science, and philosophy and is often focused on technical issues in the aggregation of a given set of preferences, often under the label "social choice theory".[2] However, the underlying ethical issues are broader, including issues of which preferences or values are included in the first place.

Some clarification on terminology is warranted. "Society", as the term is used here, is defined as the set of individual moral subjects with standing in a social choice ethics framework. A moral subject is an individual that holds ethical values. The subject can be human or non-human. AI ethics research

[1] This paper uses the terms ethics and morals interchangeably.

[2] See, for example, Prasad [5] or articles published in the journal *Social Choice and Welfare*.

has generally (but not exclusively) focused on human subjects [12]. Which non-humans may qualify as moral subjects is an ongoing topic of scientific and philosophical inquiry, including research on the moral status of nonhuman animals [13-14] and AI systems [15-16]. The ethical values held by a subject can be any form of any type of ethical theory: consequentialism, deontology, virtue ethics, etc., or any hybrid combination of these theories. The ethical values also do not need to be philosophically rigorous or defensible; they can even include obviously problematic views such as "torture puppies every other Saturday" as long as some member of society holds these views. The paper favors the term "values" over "preferences" to account for the idea that the concept of preferences may be overly narrow and inadequate for capturing the full breadth of values held by humans or other moral subjects [17]. Arguably, a social choice framework should include all values held by moral subjects, not only those that can be described as preferences. Social choice ethics is therefore any ethical framework that is based on some aggregation of ethical values held by some set of moral subjects constituting some society.

There is no one single ethical framework that constitutes *the* aggregate values of society. Instead, there are many frameworks corresponding to many ways in which societal values can be aggregated. Which framework emerges from a social choice process depends on three variables: (1) standing, meaning the question of who or what is counted as a member of society for purposes of a social choice ethics framework, (2) measurement, meaning how each individual's ethical values are obtained for use in the social choice framework, and (3) aggregation, meaning how individual values are combined into a single value system to be used for decision making [4]. Initial decisions on standing, measurement, and aggregation must be made by whoever is designing the social choice system, such as the people writing a constitution for a democracy or the people developing an AI social choice system. These decisions cannot be made by appeal to the aggregate values of society because they determine what the aggregate values of societal will end up being; to expect otherwise is circular reasoning [4]. Prasad [5] refers to this as the "democratic imposition problem": the problem of maintaining procedural legitimacy despite the need to impose some sort of values in the design of social choice processes.

Social choice ethics is in wide use. Democracy is an example. Democracies grant standing to whichever individuals are eligible to vote in elections; they measure individual values via how individuals vote in elections; and they aggregate individuals values into group values via schemes such as first-past-the-post and ranked choice. Market capitalism is another example: individuals "vote with dollars" to orient an economy toward some aggregate of individual values. Whereas democracies generally adhere to a "one person one vote" principle, in markets, wealthier individuals can have their values counted more. This distinction illustrates the point that there is no one single way of designing social choice frameworks and the point that there is no one singular framework that constitutes aggregate societal values.

2.2 AI Social Choice Ethics

Social choice ethics is used in the following lines of work in AI ethics:

Bottom-up ethics, in which AI systems learn values from interaction with other moral subjects and then seek to imitate ethical behavior, in contrast with "top-down ethics" that are specified by AI system designers [1, 18-19]. Top-down ethics would not classify as social choice unless the designers select a social choice framework.

Coherent extrapolated volition, in which AI systems observe moral subjects' existing values and then extrapolate which values they would hold if they were as smart as the AI system [2, 20]. The AI system could defer to actual, un-extrapolated human values obtained through humans voting, but only "when people are grown enough to handle it" [2, p.21], meaning a possible future time in which humans have reached a sufficiently advanced state of moral and intellectual wisdom.

Human compatibility, in which AI systems are designed such that their "only objective is to maximize the realization of human preferences" [3, p.173]. Alternatively, AI systems could be designed to advance human *interests*, meaning what is good for humans. Compatibility with interests does not classify as social choice because it is not rooted in what humans consider to be valuable: humans do not necessarily value what is good for them and they may also value other things.

Value alignment, in which AI systems are designed to align their own values to some other values. Value alignment can be interpreted in terms of social choice, in which AI system values are aligned to the values of some population of moral subjects [5]. Work on value alignment often focuses on aligning an AI system to the values of a single individual [21], in which case it can be interpreted as social choice in which only that one individual has standing. Alternatively, value alignment can be interpreted as the process of aligning AI systems to some predetermined ethical value framework [22], which would not classify as social choice unless the predetermined framework involves social choice.

These various concepts can be applied to a range of AI systems, though they are often discussed for potential advanced future systems, including superintelligence [2-5, 20-21].

There are several reasons why social choice ethics are sometimes favored for AI systems [4]. First, AI system designers may not want to impose their own values on the rest of society, perhaps especially for highly consequential advanced AI systems, so they instead favor an AI system design that accounts for the values of all of society. Second, AI system designers may not want to focus on ethics issues and opt for a social choice framework on grounds that this delegates ethics decisions to society at large. Third, AI system designers may believe that better results will tend to be achieved if a large number of individual values are considered in an AI system. Prior research shows that, despite these reasons, AI system designers must make decisions about standing, measurement, and aggregation [4]. This paper shows that AI social choice design decisions also include the issue of manipulation. Manipulation additionally poses issues for AI governance. If the challenge of designing and governing social choice ethics proves overly difficult, that could constitute a reason to instead favor other ethical frameworks.

2.3 Prior Literature

This paper's focus on value manipulation is distinctive within the literature on AI social choice ethics. Prior literature on AI social choice ethics has considered issues such as the appropriate balance of social choice and non-social choice frameworks [22], how to make decisions about standing, measurement, and aggregation [4], whether to give standing to nonhumans and how to measure and aggregate their values [12, 23-25], how to conceptualize human values for AI systems to measure [17, 26], the prospects for reaching consensus on values [27-28], meta-ethical foundations [29], and how to apply prior social choice theory research to aggregation [5]. Research on implementing social choice ethics in AI systems has thus far focused mainly on the simpler task of designing an AI system to learn the values held by a single moral subject [3, 30], though some work has explored multiple subjects [21, 31-32] and multiple sets of moral values [33-34]. Importantly, this means that there are few actual AI social choice systems in operation today, and those are mainly used for research purposes. Additionally, computer science research on AI has contributed techniques for the general social choice problem of aggregating individual values into group values [35-36]. Finally, AI governance research has studied how AI systems should be governed within human social choice institutions, especially democracies [37] and economic communities [38].[3]

[3] Some of the publications reviewed in this paragraph also contain elements that do not classify as social choice. Daley [23], Owe and Baum [12], and Ziesche [24] consider nonhumans' interests in addition to their values. Hendrycks et al. [33] and Stray [31] consider wellbeing as that which makes an individual's life go well. Hendrycks et al. [33] and Wernaart [34] consider moral values that can be implemented outside the context of social choice ethics. Nonetheless, all of these publications cover social choice ethics in various ways.

The prior literature on AI social choice ethics has two shortcomings that are addressed in this paper. First, the literature takes societal values as fixed, except insofar as they can be altered through decisions about standing, measurement, and aggregation. Little attention has gone to the idea that it is possible to change which individuals hold moral values and which values they hold. An exception is in the brief discussion of "preference engineering" in Russel [3, p.244-245]. Another exception of sorts is the brief discussion of adaptive preferences, such as poor people who have adapted to their poverty by wanting less, in Gabriel [27]. Second, the literature treats aggregate societal values in benign terms. Where decisions on standing, measurement, and aggregation are considered, the literature focuses on which decisions should be made according to certain ethical ideals. The idea that social choice ethics can be manipulated in unethical ways has gotten very little attention. An exception is the brief discussion of self-interested aggregation techniques (gerrymandering) in Baum [4]. Another exception is the brief discussion of tactical voting in Prasad [5]. Aside from these and perhaps a few other minor exceptions, the issue of manipulation has gone overlooked in literature on AI social choice ethics.

2.4 Social Choice Manipulation

A dictionary definition of "manipulate" is to "control or influence (a person or situation) cleverly or unscrupulously".[4] This paper uses "manipulation" similarly to refer actions that control or influence social choice processes. This is a broad definition that includes any action that can alter social choice processes.

Some prior literature has used the term "manipulation" more narrowly. In moral philosophy, the term "manipulation" has been used to refer to actions in which one individual induces another individual to change its values [8]. Emphasis is on devious or inappropriate actions such as deception, social pressure, or emotional appeals. This usage of "manipulation" is consistent with the idea of manipulation as clever or unscrupulous, though it only captures a portion of the actions of relevance to social choice. Furthermore, this usage of "manipulation" is sometimes distinguished from rational persuasion and forceful coercion, though the distinction can be blurry; the broader definition used in this paper avoids the need to make this distinction.

Social choice theory literature typically uses the term "manipulation" to refer to tactical voting [6-7, 39-40]. A voter may decline to vote for its favorite option and instead vote for a lesser option in order to improve the outcome of the election. For example, if voter V1 prefers option O1 but there is a tie between leading options O2 and O3, then V1 could opt to vote for its preference between O2 and O3. In this case, O1 was not going to win the election anyway, so from V1's perspective, it is better for V1 to choose between O2 and O3. Tactical voting is arguably undesirable because it entails V1 presenting a dishonest statement of preferences. However, research dating to Gibbard [6] and Satterthwaite [7] shows that under common conditions, the incentive for tactical voting cannot be avoided. Tactical voting counts as a manipulation under this paper's definition, but it is only one of many. The emphasis on tactical voting as manipulation in the social choice literature is consistent with this literature's unfortunate general orientation toward theoretical issues in preference aggregation and not toward the wider scope of issues affecting social choice.

The term "manipulation" has a negative connotation. This paper likewise has an emphasis on apparently unethical manipulations of social choice processes. However, under the broad definition of this paper, not all manipulations are unethical. For example, ethics research and education have the effect of changing the values held by moral philosophers and philosophy students. Very arguably, this is a good thing. Indeed, this paper is itself a work that may change the values that some people hold. This sort of work is vital for making progress on moral philosophy and its application to AI; it should

[4] As per the Oxford English Dictionary.

be encouraged, not suppressed. Therefore, one challenge for AI social choice ethics is to distinguish between which manipulations are good, which are bad, and which are neutral. This is another value judgment that AI social choice system designers must make: it is a design decision, not something that can be left for the AI system to do. Ethics principles for evaluating the goodness or badness of manipulations are discussed in Section 5.

The manipulation of aggregate social values relates AI social choice ethics to the concept of algorithmic bias. Algorithmic bias can be defined as a situation in which "the outputs of an algorithm benefit or disadvantage certain individuals or groups more than others without a justified reason for such unequal impacts" [41]. Individuals or groups may be unjustly disadvantaged in ways that affect aggregate social values. For example, it can be readily claimed that the genocide of Indigenous peoples [42-43] has unjustly disadvantaged them in a variety of ways. One of these ways is that the values held by Indigenous peoples are underrepresented in social choice ethics. Any AI system using social choice ethics may be biased against Indigenous values due to the genocide. More generally, AI implementations of social choice ethics that do not account for these sorts of manipulations may exhibit algorithmic bias.

Sections 3 and 4 categorize manipulations of aggregate social values in terms of design and implementation. Design refers to how decisions on standing/measurement/aggregation are built into the AI social choice system, whereas implementation refers to what happens when the AI system is used to measure and aggregate individual values. In other words, design is on the development side, whereas implementation is on the deployment side. Design and implementation are interrelated. For example, democracies are sometimes designed to make it easier for certain portions of the population to vote, which has the effect of manipulating which people actually show up to vote when elections are held. Nonetheless, the design/implementation binary offers a useful organizing structure for the paper. This binary, combined with the standing/measurement/aggregation distinction, forms a taxonomy of social choice manipulations (Table 1).

	Design	**Implementation**
Standing	Only give standing to allies	Kill rivals
Measurement	Ask loaded questions	Forcibly assimilate rivals
Aggregation	Gerrymander	N/A

Table 1. Taxonomy of social choice manipulations with select examples. Details are in Sections 3-4.

3. Manipulating AI Social Choice Design

3.1 Design Manipulation By Authorized Designers

Authorized AI system designers are those who are supposed to be involved in decisions of how to design the AI system. Authorized designers can include the engineers who build the systems, the organization(s) hosting the engineers, and certain outside parties, such as in multistakeholder governance processes.[5] Whoever they are, the group of authorized designers faces decisions on standing, measurement, and aggregation. The process of making these decisions is prone to manipulation. For brevity, here and throughout the paper, the unqualified term "designers" is used to refer to authorized designers.

AI system designers could seek to follow certain ethics principles in deciding on standing, measurement, and aggregation. Standing could be set, for example, according to the principle that all moral subjects should have standing, following some set of criteria for what classifies as a moral

[5] In multistakeholder AI governance, decisions about AI system design (among other things) are made by a variety of stakeholders such as representatives of industry, governments, civil society, and academia [44].

subject. Or, standing could be set according to the principle that standing should go to all moral subjects that currently exist (excluding past and future generations), or to all moral subjects that meet certain standards of morality (excluding social deviants of one sort or another). Measurement could be set, for example, according to the principle that moral subjects should be measured in their current form, or according to the principle that they should be measured after they have reached reflective equilibrium,[6] or according to the principle that the AI system should extrapolate which values they would hold if they were able to reflect on values as well as the AI system (as in coherent extrapolated volition). Aggregation could be set, for example, according to the principle that each individual should have the same say (as in "one person, one vote"), or that more advanced individuals should have more say (for example, human values should count more than the values of less intelligent animals), or according to the principle that the values of individuals with more strongly held positions should count more (which is roughly analogous to "voting with dollars"[7]).

Any of the above approaches would be ethically defensible. However, the fact that a variety of ethically defensible options exists creates opportunity for manipulation. AI system designers could select which approaches they consider to have the strongest ethical case. Or, AI system designers could survey the set of ethically defensible options, assess which option seems most likely to deliver their own personally preferred outcome, and select that. Their personally preferred outcome could itself be determined based on sound ethical principles, or it could be something else, such as the designers' personal self-interest. The designers could select an ethically defensible option for an ethically indefensible reason.

Alternatively, AI system designers could opt to not follow any particular ethics principles in deciding on standing. They could restrict standing to individuals that hold, or are likely to hold, whichever moral values are favored by the designers. They could measure individuals' values in ways that tend to yield designers' preferred values, such as through survey questionnaires full of loaded phrases that commonly evoke certain types of reactions or by restricting which options are available for individuals to select. They could aggregate individual values in ways that amplify the importance of individuals holding certain values, analogous to the gerrymandering of political districts. Such design decisions may all be ethically indefensible, but they are possible. There is no law of nature that requires AI system designers to choose ethical designs.

Unfortunately, there is an extensive history of social choice designers choosing designs for reasons that appear to be biased or otherwise ethically questionable. The frequent use of gerrymandering is one of many examples. Additional examples from the current United States democracy include decisions on voter identification requirements, felon disenfranchisement, statehood for the District of Columbia, and the electoral college (as in the National Popular Vote Interstate Compact). The ethicality of politicians' positions on these issues can vary, and it is not the place of this paper to judge issues of political partisanship. Nonetheless, it is apparent from observation of these issues that they are not always resolved in an unbiased and ethically sound fashion.[8] And that is just from the current United States democracy. Further abuses, including more extreme ones, can be observed in different times and places. This is especially apparent in instances of "electoral authoritarianism", in which elections occur with such a high degree of unfair manipulation (e.g., vote rigging) that they ensure that the existing

[6] Reflective equilibrium is an equilibrium state of moral reflection in which an individual's moral views will not change if provided with further information, argument, or opportunity to reflect further [45].

[7] In a market economy, people can indicate how much they care about something via how much they would be willing to pay for it, causing the market to respond to how much people care. A caveat is that this effect is distorted by wealth inequality: the wealthy may be willing to pay a lot for something they do not care much about. Another caveat is that market spending may not capture everything that people care about, or may not capture it accurately [46].

[8] For example, United States political parties have taken positions on statehood for the District of Columbia that coincide with the parties' own political interest; the same was previously done for statehood for Alaska and Hawaii [47].

leadership is retained [48]. Electoral authoritarianism is a social choice process, but it is designed to achieve a predetermined outcome that is favorable to the designers.

AI systems are being developed in many places around the world. If an AI social choice system is developed in a country with a heavily biased democracy, then perhaps the AI social choice design would be similarly biased, especially if the state is involved in the design process.[9] This is not to say that AI social choice designs would be better when AI systems are developed privately without state involvement. A common view is that corporations should focus exclusively on maximizing shareholder value [49]. Orienting AI corporations less toward their own shareholder value and more toward the public interest (however that is defined) is an ongoing challenge [50]. AI social choice systems designed to maximize a corporation's shareholder value are likely to have extensive biases.

Electoral authoritarians use sham elections to create a veneer of legitimacy. Similarly, AI system designers could use a biased social choice framework as way to claim that they are behaving ethically when in fact they are not. Unqualified support for AI social choice ethics may make it easier for unethical AI system designers to get what they want in the same way that unqualified support for elections can make it easier for authoritarians to get what they want. To the extent that social choice ethics is desirable for AI systems in the first place, it is not unconditionally desirable. It is only desirable if the design decisions of standing, measurement, and aggregation are made in some ethically sound fashion.

3.2 Design Manipulation By Unauthorized Outside Parties

It can be possible for AI social choice design to be manipulated by parties that lack authorization to do so.

Hacking the AI system is one way for unauthorized parties to manipulate the social choice design. Such an activity is analogous to the process of hacking voting machines in electoral democracies. Voting machines have been found to be vulnerable to hacking [51]. Proof of actual hacking can be elusive, but there are at least serious allegations that attempts to hack voting machines have occurred, such as U.S. allegations of Russian efforts to hack voting machines in the 2016 U.S. election [52]. It is likewise possible that outside parties could seek to hack AI social choice systems, especially if the AI systems are used in a high-stakes capacity. Though conducted by outside parties, these hacks affect the system design. They can alter who or what has standing (such as by adding or removing votes), how values are measured (such as by changing votes), and how values are aggregated (such as by changing the tabulation of votes).

Unauthorized outside parties may also be able to alter AI system design by influencing authorized designers in unauthorized ways. Such activity falls within the domain of social engineering [53]. Because it involves participation by authorized designers, it blurs the boundary between authorized and unauthorized design manipulation. It is nonetheless another issue that AI social choice system design must handle.

4. Manipulating AI Social Choice Implementation

Decisions on how to set standing, measurement, and aggregation are not the only ways to manipulate the outcomes of social choice processes. Further opportunities come on the implementation side. Given a fixed set of rules on standing, measurement, and aggregation, there is still a lot of flexibility in what a social choice process can result in. For example, in democracies, setting the design parameters does not necessarily determine who will win an election—the election itself needs to occur, with campaigning, voter turnout, and so on. The same can hold for AI social choice processes.

[9] It is plausible that all democracies are in some way biased, or rather that there is no such thing as an "unbiased" democracy.

The exact scope of opportunities to manipulate the implementation will in general depend on the design. For example, consider the distinction between AI systems that measure individual values via arranging for individuals to vote on specific issues (as in democracy) or via observing individuals' behavior as they go about their lives (as in revealed preference [46]). Both of these can be influenced, and indeed certain actions, such as advertising, can influence both. However, the specifics of how to influence voting differ from how to influence daily behavior. To take a more extreme example, consider the distinction between AI systems that give standing to nonhuman animals and those that only give standing to humans. Manipulating the moral values held by nonhuman animals is a much different type of task; to the extent that it is even possible in the first place, it may require specialized knowledge about nonhuman animal morality and access to relevant populations of nonhuman animals.

This section addresses means of manipulating implementation of standing and measurement. Manipulations of aggregation, such as through hacking (Section 3.2) have the effect of changing how values are aggregated, which, for purposes of this paper, is treated as being on the design side.

There is one important means of manipulating both standing and measurement: the manipulation of the AI system itself. Current AI technology is vulnerable to the use of adversarial input, in which inputs to machine learning algorithms are altered in ways that cause faulty pattern recognition [54]. AI social choice systems may in general require input to determine which individuals have standing and what their moral values are; both of these inputs may be prone to adversarial manipulation.

4.1 Manipulating Implementation: Standing

The paper has already discussed one way to manipulate the implementation of standing: by killing individuals that hold certain moral values. Killing is an especially jarring and important means of manipulating standing, but it is not the only one.

Historical killings may constitute a significant bias of social choice ethical systems. Consider the genocide of Indigenous peoples. For much of history, Indigenous peoples were the only inhabitants of certain regions, such as the Americas, where they currently constitute a small minority. Had the genocide not occurred, Indigenous peoples would presumably constitute a much larger portion of the global population—perhaps a large majority in the Americas, plus a smaller diaspora scattered across other regions via ordinary migration. The demographic calculation is complicated by the important role played by infectious diseases imported from Africa and Eurasia to the Americas, which may have caused extensive depopulation even if European migrants to the Americas had no ill intent. This sort of complication makes it difficult to quantify the exact Indigenous population in a counterfactual no-genocide world. Nonetheless, it is clear that the counterfactual population would be significantly larger than the actual population. Indigenous peoples are notable for tending to hold different moral values than other populations. In particular, they tend to place greater intrinsic moral value on nonhumans and on future generations. This can be seen, for example, in Indigenous research on AI ethics [55-56]. Therefore, had the genocide of Indigenous peoples not occurred, social choice systems may be significantly more oriented toward valuing nonhumans and future generations. Unless AI social choice ethics makes some sort of accounting of this historical political violence, it effectively embraces a principle of "might makes right", in which the spoils of military victory include a higher degree of representation in the ethics embedded in AI systems.

It is possible that future killings could be committed to further bias social choice systems. However, the general prohibition on murder may tend to limit the use of killing to bias AI social choice systems. Furthermore, to cause a significant bias, the killings may need to occur on a large scale. Large-scale killing is generally the domain of states, but they may be disinclined to engage in large-scale killing (war, genocide) in order to bias AI social choice systems. An exception could be for advanced, high-stakes AI systems. The extreme power of an advanced artificial general intelligence or superintelligent

AI system could conceivably motivate states to take extreme measures to manipulate it. One way to manipulate it is to be the one to build it in the first place; hence prior research has often posited that there could be a race to build advanced AI [57]. However, the "losers" of such a race could still manipulate the outcome if the "winner" opts to build a social choice system. Consider a world with two countries C_1 and C_2. Suppose C_1 invests its resources in AI research and C_2 invests its resources in weaponry. C_1 would typically be the first to build an advanced AI system. Suppose C_1 opts to build its AI system with a social choice ethics framework, perhaps inspired by literature on bottom-up ethics, human compatibility, and value alignment.[10] Following certain ethical reasoning, C_1 designs its AI system to give standing to all current humans. C_2 then uses its military to kill everyone in C_1, causing the AI system to follow the values of the people of C_2.

Such an attack may seem abhorrent, but if the alternative is for an advanced AI system to proceed according to some dispreferred set of values, then the attack could, from the attacker's perspective, appear to be the better option. Especially worrisome is the possibility that the attacker could miscalculate, causing catastrophic harm to all parties. For example, countries had considered waging nuclear war for several decades prior to the discovery of nuclear winter [58]. Nuclear winter is a global environmental phenomenon, threatening all countries including the attacker [59]. Even with awareness of the potential effects of the attack, a country may decide that the risk is worth it in exchange for the chance to dominate the social choice process of an advanced AI.

Other actions can also manipulate standing. Non-lethal violence can prevent individuals from participating in certain social choice systems. For example, violence has long been used by White Americans, including members of the Ku Klux Klan, to prevent or discourage Black Americans from voting [60]. Other tactics include persuasion, such as in advertising campaigns designed to demotivate political rivals, deception, such as by telling political rivals the wrong day to vote, and economic coercion, such as by refusing to give employees time off to vote on election day. Standing can also be manipulated in ways that are arguably more ethical, such as by encouraging people to participate in democracy and facilitating their participation.

Finally, forward-looking manipulation could seek to alter which new moral subjects are created. For example, Israel has long pursued a fertility policy aimed in part at maintaining a Jewish majority [61]. Many countries have engaged in forced sterilization to suppress the size of certain segments of their population [62]. These approaches could be pursued to influence which moral views are more common when future AI systems are built. The creation of new moral subjects may be a much larger issue if certain AI systems are themselves given standing as moral subjects. The prospect of artificial moral subjects is an ongoing topic of inquiry [15-16]. If AI systems are given standing in a social choice framework, then it may be possible to mass produce them so as to dominate the social choice; mass production could be as simple as copying and pasting AI software onto new hardware [63].

4.2 Manipulating Implementation: Measurement

As noted above, ethics research and education is one way of manipulating the moral values that individuals hold. Such work is generally taken to be a good thing. Yes, the work can be used to bias social choice frameworks in certain directions, but this is understood to be desirable: supporters of certain frameworks are supposed to explain why so that others may consider these arguments in their pursuit of reflective equilibrium. Likewise, persuasive rhetoric is generally taken to be an essential practice in a healthy deliberative democracy. At least among humans, individual moral subjects do not come prepackaged with fixed moral views and instead rely on open debate to refine their thinking.[11]

[10] Coherent extrapolate volition by advanced AI is an alternative means of identifying the reflective equilibrium of human moral subjects.

[11] Perhaps the same would not hold for at least some forms of artificial moral subjects.

This manipulation of moral values is a feature, not a bug.

Other manipulation of moral values is of more questionable ethicality. Political discourse can be ripe with disinformation, appeals to spurious values (e.g., tribalism), and propaganda. AI technology is increasingly used for such purposes [64]. A more extreme practice is the forced alteration of moral subjects through subjugation, forced assimilation, "re-education", and similar practices that can fall broadly in the category of cultural genocide [65]. These practices are unfortunate because they favor those with the most financial resources (such as for advertising expenditures), political control (such as for forcing assimilation), and willingness to engage in inappropriate persuasive tactics, instead of favoring those with stronger moral arguments. The result is a world that may not be progressing toward any sort of ethically sound reflective equilibrium.

Artificial moral subjects may also be vulnerable to manipulation, such as via adversarial input or cyber attacks. The prospect of such manipulation is especially worrisome if a certain type of artificial moral subject has been mass produced: a single attack may manipulate the moral values held by a large number of artificial moral subjects. If artificial moral subjects are sufficiently vulnerable to attack, that may even constitute a reason to deny them standing in the first place.

The above examples involve an actor manipulating the measured values of other individuals. This is consistent with the use of the term "manipulation" in moral philosophy [8]. Individuals can also manipulate their own measured values, such as via tactical voting. As social choice theory research explains, this sort of manipulation can, in some circumstances, lead to social choice processes producing results that are better from the perspective of the individual tactical voter [6-7].

5. How to Address AI Social Choice Manipulation

The issue of manipulating AI social choice systems can be addressed in a variety of ways. The best ways of addressing it depend on several factors including the ethics of how to evaluate good and bad manipulations, the design particulars of AI systems, the particular actors who may seek to manipulate the systems, and the resources available to those who would address the manipulations. Therefore, a complete accounting of how to address manipulation is beyond the scope of this paper. What follows is an outline of some possibilities.

5.1 Do Nothing

One option is to do nothing and accept the fact that AI social choice systems can be manipulated. That is effectively the current state of practice in work on AI social choice ethics. However, doing nothing can leave AI social choice systems vulnerable to the wide range of manipulations surveyed in Sections 3-4. The aggregate societal values used in the AI system could be heavily biased, including in apparently unethical ways. Such an outcome would seem to go against the spirit of AI social choice ethics. As long as there is significant potential for manipulation, the do nothing option should be rejected.

5.2 Defining Good and Bad Social Choice Manipulation

In order for AI social choice manipulation to be addressed in an ethically sound manner, it is necessary to have some standard for evaluating the goodness or badness of manipulations, so that the good ones can be promoted and the bad ones can be countered. Thus far, the paper has discussed the goodness and badness of manipulations informally: genocide seems bad, ethics education seems good, etc. Here, some formality is introduced along with a brief review of relevant literature that can be used for future research.

As a starting point, a manipulation may be taken to be good (or bad) if it: (A) is supported (or opposed) by the existing social choice process, if there is one; (B) gives the social choice process a better (or worse) specification of standing, measurement, and/or aggregation; or (C) is good (or bad)

according to an ethical theory that is not social choice.[12] For example, one might suppose that a better specification of standing includes both women and men, and therefore the women's suffrage movement was a good manipulation per (B). Preventing eligible people from voting against their will would be a bad manipulation per (B) because it causes their values to not be measured and likewise go uncounted in the aggregation. However, killing people would be neutral per (B) if standing goes to all living adults, because the deceased no longer has standing and therefore does not factor into the social choice process. Instead, killing people would be bad per (A) as long as it occurs in one of the many places where a society has, through a social choice process, outlawed killing, and it may be bad per (C) following any of the various ethical theories that regard killing to be bad.

Little needs to be said on (A); one need only consult the applicable local laws. Much can be said on (C), which contains the entire universe of non-social choice ethical theories. One notable point is that (C) is a non-social choice criterion that could play a large role in the structures and outcomes of social choice processes. For example, suppose an AI social choice system is developed in an authoritarian country where laws are not produced through a social choice process. In that scenario, killing people to manipulate the AI social choice system would not be wrong per (A) or (B) and would only be wrong per (C). This constitutes a reason why AI ethics should not rely exclusively on social choice.

Much can also be said on (B). This topic is important both for the evaluation of manipulation and for the more general ethics of AI social choice, and yet AI social choice research has neglected it. Therefore, it is worth briefly reviewing some applicable literature. There is a lot to draw on, especially political philosophy literature on the ethics of democracy.

For standing, research has considered whether individuals should have standing if they are subject to the decisions made by a social choice process (the "all subjected principle") or if they are affected by the decisions (the "all affected principle") [66-69]. Women's suffrage follows the all subjected principle [67]; the all affected principle arises in proposals for transnational democracy as an alternative to legacy geographic national and subnational boundaries [68]. These lines of research have also considered questions of standing for children [70], future generations [71-72], the deceased [73], nonhuman animals [74-75], and AI systems [76-77].

For measurement, research has explored various techniques including surveys [46, 78], observations of behavior [79-80], brain imaging [81-82], and extrapolation of what individuals would favor under idealized conditions [2, 20, 83]. Another debate concerns whether to only measure individuals' preference rankings or to instead measure the strength of their preferences [84-85]; this also has implications for aggregation. Research has also explored methods for measuring the values of nonhumans and future humans [86-87] and developed a theory of wrongful value manipulation [88].

For aggregation, two major principles are "one person, one vote" and the idea that each individual's values should be counted according to how strongly the values are held [89-90]. When nonhumans have standing, it has been proposed to count individuals' values according to their sophistication [4]. A different type of principle covers partisan fairness, the idea that aggregation should not systematically favor any particular set of values, as in gerrymandering [91]. AI social choice design should consider whether to apportion individuals into sub-population groupings such as legislative districts or to combine everyone's values together as in a national popular vote.

5.3 Governance of AI Social Choice Manipulation

Given an ethical framework for good and bad manipulation, AI governance can proceed to advance good manipulations and counter bad manipulations. For example, authorized AI system designers can

[12] (A) and (B) assume that social choice processes can be inherently good if they are implemented in an ethically sound manner. (C) assumes that there can be a non-social choice basis of inherent goodness. Both assumptions are made here only to facilitate discussion; the paper does not insist that either assumption is correct.

pursue good design manipulations and protect against bad ones. Ethicists and civil society can support AI ethics research and education. Journalists and public media outlets can support constructive public deliberation. Law enforcement can crack down on certain bad manipulations. National governments can pursue diplomacy to avoid military conflict. This is not an exhaustive list,[13] but instead is an illustration of the variety of efforts that may be needed to ensure that AI social choice systems are only manipulated in good ways.

How manipulation is handled can depend heavily on the particulars of the AI social choice system. Here are three examples. First, if the system uses something similar to current machine learning algorithms, it may be vulnerable to adversarial input in the way that current machine learning is [54], whereas systems using other algorithms may not be vulnerable to adversarial input, or may be vulnerable in a different way. Second, if the system measures the actual values held by current moral subjects, then it may be vulnerable to efforts to manipulate subjects' values, comparable to the manipulation of the values held by voters in democracies [8, 88], whereas systems using other measurement approaches, such as coherent extrapolated volition, may not face this issue and instead may face other forms of manipulation. Third, if the system is highly capable, such as superintelligence, then the high stakes may prompt more aggressive efforts of manipulation, whereas a less capable system may prompt more limited efforts. These distinctions all point to different approaches to address manipulation, such as coding paradigms to counter adversarial input, educational campaigns to counter the manipulation of moral subjects' values, and more aggressive measures to counter more aggressive manipulation efforts.

Given the many forms that AI social choice systems could take, and likewise the many ways those systems could be manipulated, a prudent course of action would be one of flexibility. To date, implementations of social choice ethics in AI systems have been mainly for research purposes [3, 21, 30-34]. These research systems do not have a large impact on the world, and so, to the extent that they are being manipulated, there is little need to address the manipulations. Instead, an anticipatory governance approach is warranted in which new developments in AI technology are monitored so that governance measures can be implemented if and when they are needed [92]. If and when AI social choice systems have more significant implementation, measures to address their manipulation can be implemented, with the measures customized to the particular implementations and the manipulations they may prompt. Meanwhile, research can develop concepts for addressing manipulation, so that the ideas will be there if and when they are needed.

5.4 Design AI Systems With Non-Social Choice Ethics

Many of the manipulation issues are unique to social choice ethics. Other ethical frameworks do not place weight on the moral values held by populations and thus cannot be manipulated by altering the moral values held by populations. Social choice is not the only defensible type of ethical framework. Indeed, sound arguments can be made in favor of non-social choice frameworks, for example, arguments for utilitarianism that seeks to maximize experienced welfare (e.g., happiness) instead of preference satisfaction [93-94]. Perhaps other frameworks would be better for AI systems, especially after accounting for the potential for manipulation.

Other frameworks can also be vulnerable to manipulation. Indeed, all AI ethics frameworks may be vulnerable to design manipulation, whether by authorized designers or unauthorized outside parties, for the sorts of reasons described in Section 3. Some frameworks may also be vulnerable on the implementation side. For example, a framework oriented toward increasing the quality of subjective experience (e.g., happiness) could conceivably be manipulated by altering the capacity for subjective

[13] A more detailed set of governance measures for AI systems developed by corporations (as many AI systems are, especially the most powerful systems) is provided by Cihon et al. [50].

experiences held by sentient beings. Decisions on which type of framework to use should account for the potential for all frameworks to be manipulated, not just social choice frameworks.

5.5 Design AI Systems With A Hybrid of Social Choice and Non-Social Choice Ethics

The challenge of social choice manipulation, alongside other concerns about social choice, could be resolved by limiting the scope of what social choice is used for, with some other type(s) of ethics used elsewhere. AI systems can be designed with a hybrid of social choice and non-social choice frameworks [19, 22]. For example, an AI system could be designed to focus mainly on maximizing total experienced utility, and to also use social choice wherever doing so does not significantly reduce total experienced utility. That could give society (or societies plural) a significant degree of autonomy and influence, achieving some of the desirable aspects of social choice, while maintaining a separate standard (total experienced utility) to ensure that any social choice manipulations do not induce strongly unethical outcomes. A carefully crafted hybrid may constitute an optimal, best-of-both worlds ethical framework, though there may also be a case for either a pure social choice or pure non-social choice approach.

6. Conclusion

Social choice ethics is not the inherently benign framework that its advocates in the AI literature make it out to be. To the contrary, it is prone to a wide range of manipulations, many of which are very arguably unethical. Calls for AI social choice ethics must take these manipulations into account and explain how to address them. Prior literature has not done this. This paper is an initial step in that direction. The absence of prior attention to manipulations is especially worrisome because unqualified support for AI social choice could lead to ethically inferior outcomes, with sham AI social choice ethics used to justify unethical AI system design in the same way that authoritarians sometimes use sham elections as political cover. This is especially worrisome for potential future highly consequential advanced AI systems, including superintelligence, for which prior research often proposes social choice ethics. The high stakes of the AI systems could motivate extreme manipulations, which could have catastrophic consequences on its own and lead to catastrophically bad values used by the AI system. In this context, attention to social choice manipulation is particularly important.

The importance of manipulation reinforces the need for a critical and interdisciplinary study of AI ethics. Issues of manipulation cannot be handled by computer science alone; they also need moral philosophy, social science, and perhaps also other disciplines. Furthermore, the issues of manipulation cannot be studied through a focus on ethical ideals. Instead, a "red teaming" critical analysis is needed to identify flaws. This paper is unusual as a work of AI ethics research that focuses primarily on unethical behavior, but such research is needed to fully evaluate the merits of different approaches to AI systems.

As AI technology becomes more advanced and more widely deployed, the need to address these issues becomes urgent. Indeed, some discussions of social choice ethics, such as on coherent extrapolated volition, pertain to highly advanced future AI systems in which flawed ethical design can be catastrophic. These ethics issues are certainly also pertinent to systems that are not quite that powerful. Good outcomes from AI systems depend on the successful resolution of these difficult ethics challenges.

Acknowledgments

McKenna Fitzgerald, Jonathan Stray, Iason Gabriel, Dominika Krupocin, Mahendra Prasad, and an anonymous reviewer provided helpful comments on an earlier version of this paper. Any remaining errors are the author's alone.

References

1. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, Oxford (2008)
2. Yudkowsky, E.: Coherent extrapolated volition. The Singularity Institute (2004)
3. Russell, S.J.: Human Compatible: Artificial Intelligence and the Problem of Control. Viking (2019)
4. Baum, S.D. Social choice ethics in artificial intelligence. AI & Society 35(1), 165-176 (2020)
5. Prasad, M.: Social choice and the value alignment problem. In: Yampolskiy, R.V. (ed.), Artificial Intelligence Safety and Security, pp. 291-314. Chapman and Hall/CRC (2018)
6. Gibbard, A.: Manipulation of voting schemes: A general result. Econometrica 41(4), 587-601 (1973)
7. Satterthwaite, M.A.: Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. Journal of Economic Theory 10(2), 187-217 (1975)
8. Noggle, R.: The ethics of manipulation. In: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Summer 2022 Edition (2022). https://plato.stanford.edu/archives/sum2022/entries/ethics-manipulation
9. Botes, M.: Autonomy and the social dilemma of online manipulative behavior. *AI and Ethics 3*(1), 315-323 (2023)
10. De Condorcet, M.: Essai Sur L'Application de L'Analyse a la Probabilite Des Decisions Rendues a la Pluralite Des Voix. L'imprimerie Royale (1785)
11. Arrow, K.J.: Social Choice and Individual Values. Wiley, New York (1951)
12. Owe, A., Baum, S.D.: Moral consideration of nonhumans in the ethics of artificial intelligence. AI and Ethics 1(4), 517-528 (2021)
13. De Waal, F.: Primates and Philosophers: How Morality Evolved. Princeton University Press, Princeton (2006)
14. Monsó, S., Benz-Schwarzburg, J., Bremhorst, A.: Animal morality: What it means and why it matters. The Journal of Ethics 22, 283-310 (2018)
15. Cervantes, J.A., López, S., Rodríguez, L.F., Cervantes, S., Cervantes, F., Ramos, F.: Artificial moral agents: A survey of the current status. Science and Engineering Ethics 26(2), 501-532 (2020)
16. Szentgáli-Tóth, B.A.: Robotic personhood and its potential impact to democracy: Should artificial intelligence be citizens and vested with right to vote? In: The Law of the Future - The Future of Law, pp. 771-807. Páneurópska Vysoká škola (2021)
17. Sotala, K.: Defining human values for value learners. In: Workshops at the Thirtieth AAAI Conference on Artificial Intelligence: AI, Ethics, and Society (2016) https://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12633
18. Allen, C., Smit, I., Wallach, W.: Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics and Information Technology 7(3), 149-155 (2005)
19. Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., Bernstein, A.: Implementations in machine ethics: A survey. ACM Computing Surveys 53(6), article 132 (2020)
20. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford (2014)
21. Critch, A., Krueger, D.: AI research considerations for human existential safety (ARCHES). https://arxiv.org/abs/2006.04948 (2020)
22. Kim, T.W., Hooker, J., Donaldson, T.: Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. Journal of Artificial Intelligence Research 70, 871-890 (2021)
23. Daley, K.: Two arguments against human-friendly AI. AI and Ethics 1(4), 435-444 (2021)
24. Ziesche, S.: AI ethics and value alignment for nonhuman animals. Philosophies 6, article 31 (2021)
25. Moret, A.R.: Taking into account sentient non-humans in AI ambitious value learning: Sentientist

coherent extrapolated volition. Journal of Artificial Intelligence and Consciousness 10(02), 309-334 (2023)
26. Han, S., Kelly, E., Nikou, S., Svee, E.O.: Aligning artificial intelligence with human values: reflections from a phenomenological perspective. AI & Society 37, 1383-1395 (2022). https://doi.org/10.1007/s00146-021-01247-4
27. Gabriel, I.: Artificial intelligence, values, and alignment. Minds and Machines 30(3), 411-437 (2020)
28. Sutrop, M.: Challenges of aligning artificial intelligence with human values. Acta Baltica Historiae et Philosophiae Scientiarum 8(2), 54-72 (2020)
29. Boyles, R. J. M.: Can't bottom-up artificial moral agents make moral judgements? Filosofija. Sociologija 35(1), 14-22 (2024)
30. Rodriguez-Soto, M., Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A.: Instilling moral value alignment by means of multi-objective reinforcement learning. Ethics and Information Technology 24, article 9 (2022)
31. Stray, J.: Aligning AI optimization to community well-being. International Journal of Community Well-Being 3(4), 443-463 (2020)
32. Koster, R., Balaguer, J., Tacchetti, A., Weinstein, A., Zhu, T., Hauser, O., et al.: Human-centred mechanism design with Democratic AI. Nature Human Behaviour 6(10), 1398-1407 (2022)
33. Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., Steinhardt, J.: Aligning AI with shared human values. (2020). https://arxiv.org/abs/2008.02275
34. Wernaart, B.: Developing a roadmap for the moral programming of smart technology. Technology in Society 64, article 101466 (2021)
35. Lackner, M., Skowron, P.: Approval-based multi-winner rules and strategic voting. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), pp. 340-346 (2018). https://www.ijcai.org/proceedings/2018/47.
36. Noothigattu, R., Gaikwad, S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., Procaccia, A.: A voting-based system for ethical decision making. In: Proceedings of the AAAI Conference on Artificial Intelligence 32(1) (2018). https://ojs.aaai.org/index.php/AAAI/article/view/11512
37. Erman, E., Furendal, M.: The global governance of artificial intelligence: Some normative concerns. Moral Philosophy and Politics 9(2), 267-291 (2022). https://doi.org/10.1515/mopp-2020-0046
38. Häußermann, J.J., Lütge, C.: Community-in-the-loop: towards pluralistic value creation in AI, or—why AI needs business ethics. AI and Ethics 2, 341-362 (2022)
39. Campbell, D. E., Kelly, J. S.: Gains from manipulating social choice rules. Economic Theory 40(3), 349-371 (2009)
40. Gori, M.: Manipulation of social choice functions under incomplete information. Games and Economic Behavior 129, 350-369 (2021)
41. Kordzadeh, N., Ghasemaghaei, M.: Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems 31(3), 388-409 (2022). https://doi.org/10.1080/0960085X.2021.1927212
42. Maybury-Lewis, D.: Genocide against Indigenous peoples. In: Annihilating Difference: The Anthropology of Genocide, pp. 43-53. University of California Press (2002)
43. Totten, S., Parsons, W.S., Hitchcock, R.K.: Confronting genocide and ethnocide of Indigenous peoples: An interdisciplinary approach to definition, intervention, prevention, and advocacy. In: Annihilating Difference: The Anthropology of Genocide, pp. 54-94. University of California Press (2002)
44. Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., Floridi, L.: Artificial intelligence and the 'good

society': The US, EU, and UK approach. Science and Engineering Ethics 24(2), 505-528 (2018)
45. Rawls, J.: A Theory of Justice. Belknap Press, Cambridge, MA (1971)
46. Carson, R.T., Hanemann, W.M.: Contingent valuation. Handbook of Environmental Economics 2, 821-936 (2005)
47. Feinberg, A.: Everyone knows why Republicans really oppose DC statehood — even members of their own party. The Independent, 23 April (2021). https://www.independent.co.uk/voices/dc-statehood-republicans-racist-black-voters-b1836504.html
48. Schedler, A.: The Politics of Uncertainty: Sustaining and Subverting Electoral Authoritarianism. Oxford University Press, Oxford (2013)
49. Stout, L.A.: The Shareholder Value Myth: How Putting Shareholders First Harms Investors, Corporations, and the Public. Berrett-Koehler (2012)
50. Cihon, P., Schuett, J., Baum, S.D.: Corporate governance of Artificial Intelligence in the public interest. Information 12(7), article 275 (2021)
51. Hao, F., Ryan, P.Y.A. (Eds.): Real-World Electronic Voting: Design, Analysis and Deployment. CRC Press (2016)
52. Sanger, D.E., Edmondson, C.: Russia targeted election systems in all 50 states, report finds. The New York Times, 25 July (2019). https://www.nytimes.com/2019/07/25/us/politics/russian-hacking-elections.html
53. Hadnagy, C.: Social Engineering: The Science of Human Hacking, 2nd Edition. Wiley (2018)
54. Biggio, B., Roli, F.: Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition 84, 317-331 (2018)
55. Abdilla, A.: Beyond imperial tools: Future-proofing technology through indigenous governance and traditional knowledge systems. In: Harle, J., Abdilla, A., Newman, A. (eds.), Decolonising the Digital Technology as Cultural Practice, pp. 67-81. Tactical Space Lab (2018)
56. Lewis, J.E. (ed.): Indigenous protocol and artificial intelligence position paper. The Initiative for Indigenous Futures and the Canadian Institute for Advanced Research (CIFAR) (2020)
57. De Neufville, R., Baum, S.D.: Collective action on artificial intelligence: A primer and review. Technology in Society 66, article 101649 (2021)
58. Badash, L.: A Nuclear Winter's Tale: Science and Politics in the 1980s. MIT Press, Cambridge, MA (2009)
59. Robock, A., Toon, O.B.: Self-assured destruction: The climate impacts of nuclear war. Bulletin of the Atomic Scientists 68(5), 66-74 (2012)
60. Logan, M.A.: The vote is precious. Indiana Journal of Law and Social Equality 5(1), 105-131 (2016)
61. Steinfeld, R.: War of the Wombs: The History and Politics of Fertility Policies in Israel, 1948-2010. Doctoral dissertation, Oxford University (2011)
62. Patel, P.: Forced sterilization of women as discrimination. Public Health Reviews 38, article 15 (2017)
63. Yampolskiy, R.V.: Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In: Müller, V.C. (ed.), Philosophy and Theory of Artificial Intelligence, pp. 389-396. Springer (2013)
64. Woolley, S.C., Howard, P.N. (eds.): Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media. Oxford University Press, Oxford (2018)
65. Novic, E.: The Concept of Cultural Genocide: An International Law Perspective. Oxford University Press, Oxford (2016)
66. Arrhenius, G.: The boundary problem in democratic theory. In: Democracy Unbound: Basic Explorations I, pp. 14-29. Filosofiska Institutionen (2005)

67. Näsström, S.: The challenge of the all-affected principle. Political Studies 59(1), 116-134 (2011)
68. Schaffer, J.K.: The boundaries of transnational democracy: Alternatives to the all-affected principle. Review of International Studies, 38(2), 321-342 (2012)
69. Andrić, V.: Is the all-subjected principle extensionally adequate? Res Publica, 27, 387-407 (2021)
70. Campos, A. S.: Infant political agency: Redrawing the epistemic boundaries of democratic inclusion. European Journal of Political Theory 21(2), 368-389 (2022)
71. Heyward, C.: Can the all-affected principle include future persons? Green deliberative democracy and the non-identity problem. Environmental Politics 17(4), 625-643 (2008)
72. Schuessler, R., Gillerke, F.: Voice and no votes for future citizens. In: Representing the Absent, pp. 375-392. Nomos (2023)
73. Bengtson, A.: Dead people and the all-affected principle. Journal of Applied Philosophy 37(1), 89-102 (2020)
74. Garner, R.: Animals and democratic theory: Beyond an anthropocentric account. Contemporary Political Theory 16, 459-477 (2017)
75. Magaña, P.: Nonhuman animals and the all affected interests principle. Critical Review of International Social and Political Philosophy, in press, https://doi.org/10.1080/13698230.2022.2100962
76. Beckman, L., Rosenberg, J.H.: The democratic inclusion of artificial intelligence? Exploring the patiency, agency and relational conditions for demos membership. Philosophy & Technology 35(2), article 24 (2022)
77. Akova, F.: Artificially sentient beings: Moral, political, and legal issues. New Techno Humanities, 3(1), 41-48 (2023)
78. Esmer, Y., Pettersson, T.: Measuring and Mapping Cultures: 25 Years of Comparative Value Surveys. Brill (2007)
79. Sen, A.: Behaviour and the concept of preference. Economica 40(159), 241-259 (1973)
80. Crawford, I., De Rock, B. Empirical revealed preference. Annual Review of Economics 6(1), 503-524 (2014)
81. Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D.: An fMRI investigation of emotional engagement in moral judgment. Science 293, 2105-2108 (2001)
82. Liao, S.M.: Moral Brains: The Neuroscience of Morality. Oxford University Press (2016)
83. Broome, J.: Can there be a preference-based utilitarianism. In: Justice, Political Liberalism and Utilitarianism: Themes from Harsanyi and Rawls, pp. 221-238. Cambridge University Press (2008)
84. Drakopoulos, S.A.: The historical perspective of the problem of interpersonal comparisons of utility. Journal of Economic Studies 16(4), 35-51 (1989)
85. Balinski, M., Laraki, R.: A theory of measuring, electing, and ranking. Proceedings of the National Academy of Sciences 104(21), 8720-8725 (2007)
86. Eckersley, R.: Representing nature. In: The Future of Representative Democracy, pp. 236-257. Cambridge University Press (2011)
87. Gonzalez-Ricoy, I., Rey, F.: Enfranchising the future: Climate justice and the representation of future generations. Wiley Interdisciplinary Reviews: Climate Change 10(5), article e598 (2019)
88. Christiano, T.: Algorithms, manipulation, and democracy. Canadian Journal of Philosophy 52(1), 109-124 (2022)
89. Hayden, G.M.: The false promise of one person, one vote. Michigan Law Review 102(2), 213-267 (2003)
90. Gersbach, H.: Why one person one vote? Social Choice and Welfare 23(3), 449-464 (2004)
91. Stephanopoulos, N.O., McGhee, E.M.: Partisan gerrymandering and the efficiency gap. University of Chicago Law Review 82, 831-900 (2015)

92. Cremer, C. Z., Whittlestone, J.: Artificial canaries: Early warning signs for anticipatory and democratic governance of AI. International Journal of Interactive Multimedia and Artificial Intelligence 6(5), 100-109 (2021)
93. Ng, Y.K.: From preference to happiness: Towards a more complete welfare economics. Social Choice and Welfare 20(2), 307-350 (2003)
94. Kahneman, D., Sugden, R.: Experienced utility as a standard of policy evaluation. Environmental and Resource Economics 32(1), 161-181 (2005)