The Ethics of Sustainability for Artificial Intelligence

Andrea Owe[1], Seth D. Baum[2]

{andrea.owe@gcrinstitute.org[1], seth@gcrinstitute.org[2]}

The Global Catastrophic Risk Institute, PO Box 40364, Washington, DC 20016, USA[1,2]

Abstract. Sustainability is widely considered a good thing and is therefore a matter of ethical significance. This paper analyzes the ethical dimensions of existing work on AI and sustainability, finding that most of it is focused on sustaining the environment for human benefit. The paper calls for sustainability that is not human-centric and that extends into the distant future, especially for advanced future AI as a technology that can advance expansion beyond Earth.

Keywords: Artificial Intelligence, Sustainability, Ethics, Anthropocentrism, Ecocentrism, Long-term, Optimization

1 Introduction

A basic attribute of modern human civilization is that the stock of natural resources steadily decreases, whereas the stock of artificial resources steadily increases. For example, artificial intelligence (AI) research is commonly powered by the burning of fossil fuels, and in the process produces new technologies that civilization can benefit from. Will the increases in artificial resources be sufficient to offset the loss of natural resources, such that civilization can be sustained into the future? That is one important perspective on the ethics of sustainability as it relates to AI, though, as this paper discusses, it is not the only one.

Sustainability is not an inherently ethical concept. In its essence, "sustainability" refers to a particular characteristic of systems as they change over time. The term can be used in many ways that do not have any particular ethical significance. For example, sprinters run at an unsustainable speed; eventually, their muscles will fatigue and they will be unable to continue. This is a basic characteristic of human physiology and not a matter of ethical significance.

In common usage, however, sustainability takes on ethical significance. Sustainability is widely treated as a good thing and something worth pursuing [1]. It is in that spirit that there have been initiatives on AI and sustainability, including conferences such as "Sustainable AI",[1] "Towards Sustainable AI",[2] and "AI for the Planet",[3] as well as groups such as AI4Good that work on AI in support of the United Nations Sustainable Development Goals (SDGs).[4] It is also in that spirit that sustainability is one of the principles found in some AI ethics guidelines [2].

1 https://www.uni-bonn.de/en/news/120-2021

2 http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=140164&copyownerid=158619

3 https://aifortheplanet.org/

4 https://ai4good.org/

CAIP 2021, November 20-24, Bologna, Italy

DOI 10.4108/eai.20-11-2021.2314105

The ethical dimensions of sustainability warrant ethical analysis. Important ethics questions include: What exactly should be sustained, or rather, be able to be sustained? Why should it be able to be sustained? For how long? How much emphasis should be placed on sustainability relative to other goals? These ethics questions can be answered in a variety of ways. How they are answered has important implications for ongoing human activity, including for the development, use, and governance of AI technology.

Prior literature on AI and sustainability (reviewed below) has not considered the ethical dimensions to any significant extent. Therefore, this paper analyzes the ethics of sustainability as it relates to AI. The paper proceeds in three parts. First, we explain the ethics of sustainability as a general concept (Section 2). Second, we describe the current usage of the concept of sustainability within work on AI (Section 3). Third, we present an argument for a long-term, non-anthropocentric conception of sustainability and explain the implications of this for AI (Section 4). Section 5 concludes.

Much of the discussion in this paper covers ethical dimensions of sustainability that are not specific to AI. This is a feature, not a bug. To a large extent, the ethical dimensions of AI and sustainability are the same as those for sustainability in general. Furthermore, aspects of the topic that are specific to AI build on more general sustainability concepts. Therefore, to understand the ethics of sustainability for AI, it is essential to first understand the ethics of sustainability.

The paper contributes to the growing literature on AI and sustainability. Most of this literature is on AI in relation to the sustainability of human civilization and its environmental underpinnings; this includes reviews by Nishant et al. [3] and Liao and Wang [4], the environmental politics of AI [5], AI in relation to systemic risk and sustainability [6], the environmental footprint of AI systems [7], and the role of AI in meeting the SDGs [8-10]. Some literature focuses on the sustainability of the AI systems themselves [11], including for consumer autonomy [12], in global health initiatives [13], in certain economic mechanisms [14], and in decision-making applications [15]. Additionally, there is a broader field of computational sustainability that applies computer science methods to advance environmental and social sustainability [16-18]. Overall, the literature on AI and sustainability offers a variety of important contributions, but it provides limited discussion of the ethics of sustainability.

The paper additionally contributes to some adjacent literatures. Prior studies have considered the application of AI for environmental protection [19-20], for climate change mitigation [21-22], and the energy consumption of AI systems [23]; these are relevant for environmental conceptions of sustainability. Also relevant are debates on the relative importance of near-term, medium-term, and long-term AI [24-27]; as this paper discusses, the future-orientation of sustainability can imply an emphasis on long-term AI. Finally, of more general relevance is prior work on the ethics of sustainability [1], especially sustainability over long time scales [28], and the ethics of AI [2,29-30], especially regarding AI and the future [31] and AI and nonhumans [32].

2 The Ethics of Sustainability

The word “sustainability” is commonly traced to the German word *Nachhaltigkeit*, and specifically to Hans Carl von Carlowitz’s 1713 treatise on sustainable yield forestry [33]. The tension between using resources now and preserving or cultivating them for future use is of

course much older. Modern analysis traces to the environmental economics work of Hotelling [34], which remains relevant today [35]. Hotelling's work did not use the term "sustainability". Use of the term primarily traces to the 1987 report *Our Common Future*, which was led by former Norwegian Prime Minister Gro Harlem Brundtland and is commonly known as the Brundtland Report. Like von Carlowitz's treatise, the Brundtland Report specifically conceptualizes sustainability in socio-environmental terms. The report's definition of sustainable development, "Development that meets the needs of the present without compromising the ability of future generations to meet their needs", has been widely influential and serves as a foundation for the UN SDGs.

Since *Our Common Future*, there has been a proliferation of definitions of sustainability and sustainable development [36-38]. Widespread and imprecise usage of the term "sustainability" has watered it down. Critics argue that "sustainability" has become "a concept that is equivalent to 'good' and thus devoid of any specific meaning—a blanket concept to assure stakeholders of the policy's good intentions" [39, p.3439]. Likewise, the term is said to have been appropriated by self-interested actors to continue "business as usual" activities that drive environmental destruction and social inequity [37].

One notable point of criticism is the so-called "three pillars" of sustainability: the social, the economic, and the environmental—or "people, profit, and planet". The pillars construct is an attempt to represent the major components of socio-environmental sustainability. However, these categories have been criticized for being fuzzy and overlapping, for excluding of other relevant categories such as the cultural and the political, and for not being essential to the core matter of whether civilization can be sustained over time [39]. Indeed, economic transactions are an inherently social activity, and all human activity is inherently environmental because humans are part of nature. Furthermore, sustainability is commonly associated with environmental protection, but some environmental disturbances, such as many types of air or water pollution, rapidly dissipate and have a negligible effect on future generations' ability to sustain themselves. Likewise, some matters that could classify as social and/or economic, such as social justice within the contemporary population, are often of limited relevance to the ability of future generations to sustain themselves, even if these matters may be important for other reasons. Other socio-economic matters, such as education and investment in future economic growth, may be of greater importance to the ability of future generations to sustain themselves. For these reasons, the "three pillars" provide a weak foundation for sustainability. Work on AI and sustainability that incorporates the "three pillars", such as the "AI for People: Towards Sustainable AI" conference,[5] should take note.

We now turn to three fundamental questions for the ethics of sustainability (Sections 2.1-2.3) followed by a comparison between sustainability and optimization (Section 2.4).

2.1 What should be able to be sustained, and why?

The reasons something should be able to be sustained typically derive from what are, in moral philosophy, referred to as intrinsic value and instrumental value. Roughly speaking, something is intrinsically valuable if it is valuable for its own sake, or valuable as an ultimate end in itself; something is instrumentally valuable if it is valuable because it promotes something else that is valuable [40]. For example, we might suppose that sunlight is valuable because it can

5 https://aiforpeople.org/conference/

(among other things) be converted into electricity, and that electricity is valuable because it can (among other things) be used to power AI systems, and that AI systems are valuable because they can (among other things) enable humans to have enjoyable lives, and that enjoyable human lives are just good things on their own. In that case, sunlight, electricity, and AI systems are instrumentally valuable, while enjoyable human lives are intrinsically valuable.

In many conceptions of ethics, what should be done—including what should be sustained—is defined with reference to some conception of what is intrinsically valuable. Conceptions of sustainability can vary in terms of what they intrinsically value and what sorts of instrumental values they focus on. One can seek to sustain X either because X is intrinsically valuable or because X is instrumentally valuable. For example, one can seek to sustain natural ecosystems either because one considers them to be intrinsically valuable or because one considers them to be instrumentally valuable for other things, such as for human welfare.

Common conceptions of sustainability are anthropocentric, meaning that they only intrinsically value humans [41]. For example, although the Brundtland Report's emphasis on future generations could conceivably be interpreted to mean future generations of something other than humans, the Report clearly focuses on humans. Likewise, sustainable management of natural resources generally treats the resources as instrumental values for human benefit. In contrast, some conceptions of sustainability are ecocentric, meaning that they intrinsically value ecosystems. Examples include the Earth Charter[6] and the Earth Manifesto [42]. Though less common, the concept of sustainability can also be used with other notions of intrinsic value, such as the idea that there is intrinsic value in the welfare of sentient nonhuman animals or (if possible) sentient AI systems. The distinction between intrinsic and instrumental value does not always matter, but it often is important. For example, reducing greenhouse gas emissions is generally good on both anthropocentric and ecocentric grounds. However, some biodiversity protection is worth pursuing mainly on ecocentric grounds, because, for better or worse, certain species could go extinct with little impact on human welfare. Therefore, it is important for discussions of sustainability to explicitly specify what they intrinsically value.

2.2 For how long should something be able to be sustained?

There is a big difference between sustaining something for a few days and sustaining it for decades, centuries, or even indefinitely into the distant future. Unfortunately, discussions of sustainability are often not precise in their consideration of time scales. For example, the 1987 Brundtland Report's emphasis on future generations implies a time scale of at least decades, assuming the generations are of humans. But how many future generations? The actions needed to enable the next few generations to be sustained often differ significantly from the actions needed to enable the same for every generation that could ever exist.

2.3 How much effort should be made for sustainability?

The sustainability of intrinsic or instrumental values may be a good thing, but how good? The world has many competing values and opportunities. Indeed, the Brundtland definition was specifically crafted to acknowledge the competing values of present and future generations; while the report aspires to promote actions that support both the present and future

6 https://earthcharter.org/

generations, many choices involve tradeoffs between them. For example, depleting natural resources often benefits present generations at the expense of future generations. Basic research often benefits future generations at the expense of present generations, especially where the same resources could instead be used for applied research. The relative importance of the present and future can be operationalized in a variety of ways, such as through discount rates or other weighting functions [43-44]. This sort of intergenerational evaluation is generally made within the context of anthropocentric conceptions of sustainability, but similar approaches can be taken with other conceptions. One way or another, moral guidance about sustainability must consider how to evaluate tradeoffs between sustainability and other moral goals. This point fits within the broader issue of tensions between different AI ethics principles [45].

Also relevant are fundamental questions about the appropriate degree of effort to take to achieve ethical goals. In moral philosophy, the term "supererogation" refers to actions that "go beyond the call of duty", meaning that they are good but not strictly required [46]. One common question in moral philosophy is whether some moral frameworks are too demanding. This question is especially acute for consequentialist moral frameworks that call for moral agents to maximize some conception of intrinsic value, because maximization is a demanding task [47]. These are important questions for any moral debate, certainly including those involving sustainability. Setting aside tradeoffs between sustainability and other moral goals, one can ask: How much effort should a person or an organization make to advance sustainability? Should they "give it everything they've got"? Or would just a little effort be acceptable? Conversely, is it enough to work to advance sustainability? Or is it important to also work toward more ambitious goals, such as intertemporal optimization of intrinsic value?

2.4 Sustainability vs. Optimization

Sustainability can be an optimization criterion—that would mean seeking to optimize the ability for something to be sustained over time. However, this is distinct from optimizing intrinsic value. Sustainability means enabling something to be sustained in at least some minimal form; optimization means making something be the best that it can be. Ensuring sustainability is perhaps best understood as a basic minimum standard of intertemporal conduct, whereas the intertemporal optimization of intrinsic value may be understood as a loftier ideal to aspire for.

This distinction can be seen, for example, in the Brundtland Report call for the present generation to act "without compromising the ability of future generations to meet their needs". The basic needs of human life are, to an approximation, food, clothing, and shelter. Following the Report's call could result in "a society living forever at a minimum subsistence level of consumption" [48, p.327]. Future society would be able to meet its needs, but it may not be able to do anything more. If the present generation does not act so as to enable future generations to do much better than meeting their needs, then the present generation will, quite arguably, have squandered a massive opportunity. Of course, if the present generation fails to enable future generations to meet their needs, that would be, quite arguably, a massive loss.

AI is advanced technology. AI research and development is often oriented toward enabling higher standards of living instead of enabling basic future needs. Human lives do not strictly need, for example, AI systems to steer vehicles or search the internet. Such work generally falls outside the scope of sustainability, but could fall within the scope of optimizing intrinsic value.

3 Prior Work on AI and Sustainability

With the ethics of sustainability in mind, we now survey prior work on AI and sustainability. Sections 3.1 and 3.2 present quantitative analysis of trends in the ethics of sustainability found in AI ethics principles and AI research. Both analyses characterize sustainability in terms of the three ethics dimensions presented in Section 2: intrinsic value, time scale, and degree of effort. Section 3.3 presents overarching trends across Sections 3.1-3.2.

3.1 AI Ethics Principles

Jobin et al. [2] compiles 84 sets of AI ethics principles. 11 of these sets of principles include some reference to sustainability.[7] This indicates that sustainability is a small but nonzero priority in AI ethics. We examined the ethical basis of the 11 sets of principles. We found that 7 refer to some form of environmental sustainability, 3 refer to sustainability of the AI system itself, and 1 refers to sustainable social development. We further found that 3 intrinsically value humans, 5 intrinsically value humans and nonhumans including ecosystems, all life, biodiversity, and the planet, and 5 are ambiguous in terms of what is intrinsically valued. Regarding time scales, 2 refer to "future generations" and 9 do not specify time scales. Finally, none of the principles specify degree of effort. Some examples of the AI ethics principles are presented in Appendix A.

3.2 AI Sustainability Research

We performed a systematic mapping review [49] of research at the intersection of AI and sustainability. Our review maps this literature in terms of the ethical attributes presented in Section 2 and also used in Section 3.1, with some additional nuances to catch the diversity of the research literature. Specifically, we analyzed results from a Google Scholar search for ["artificial intelligence" "sustainability"] and ["AI" "sustainability"] conducted between Sept 16 and 21, 2021. The Google Scholar search engine was selected over other academic databases due to its inclusivity. Whereas databases such as Web of Science concentrate on peer-reviewed journals, artificial intelligence research is often published in other spaces such as arXiv.

The searches returned 229,000 total results for ["artificial intelligence" "sustainability"] and 1,490,000 total results for ["AI" "sustainability"], respectively. We examined the first ten pages of each of the two searches. We observed that after ten pages of each search, the search results became repetitive and less relevant. These two sets of ten pages contained 200 total results, or 153 results after duplicates were extracted. Of these 153 publications, we were unable to access 11. For the remaining 142 publications, we examined the text in the degree of detail needed to categorize its treatment of sustainability. For most of the publications, this involved looking at the abstract and introduction, and skimming the text for discussion of sustainability. In some cases, we examined the entire publication in more detail. Out of the

7 Table 3 of Jobin et al. [2] states that 14 sets if principles include sustainability, but only 12 are referenced in the text and we were unable to identify the other 2. Of the 12 referenced sets of principles, we found that 1 did not cover sustainability, leaving a total of 11 for our analysis.

142 publications, 60 were found not to be relevant because they were not sufficiently on the nexus of AI and sustainability, leaving a data set of 82 publications.

66 publications were on environmental sustainability, 7 were on the sustainability of the AI system itself, and 9 were on the sustainability of something else, including 2 on the sustainability of organizations, 2 on social sustainability in human-robot/AI interactions, 1 on the sustainability of group decision-making processes, 1 on sustainable curriculum planning, 1 on sustainable healthcare systems, 1 on the social sustainability of AI, and 1 on sustainable industrial development. Of the 66 environmental sustainability publications, 43 were on environmental and social sustainability and 23 were exclusively on environmental sustainability. 29 of the 66 environmental sustainability publications referred to the Brundtland definition and/or the SDGs.

56 publications intrinsically valued humans only, 10 intrinsically valued humans and nonhumans including nonhuman species, life on Earth, biodiversity, ecosystems, the biosphere, and the planet. 16 were too ambiguous to interpret any notion of intrinsic value. Regarding time scales, 7 refer to "future generations" and 1 refers to a time frame from 1990 to 2028. The other 74 do not specify time scales. None of the publications specify degree of effort. Some examples of the AI sustainability publications are presented in Appendix B.

3.3 Overarching Trends

In consideration of the work presented in the two preceding subsections, the following overarching trends in the ethics of existing work on AI and sustainability can be identified.

First, most work on AI and sustainability is focused on some form of environmental sustainability, with substantial minorities focused on the sustainability of AI systems or on the sustainability of miscellaneous other things. The environmental sustainability work mainly intrinsically values humans and sometimes intrinsically values nonhumans. These trends are consistent with wider usage of sustainability outside the context of AI. Indeed, work on AI and sustainability often explicitly links to broader treatments of sustainability, such as the Brundtland Report and the UN SDGs. Work on the sustainability of AI systems treats AI systems as instrumentally valuable, such as in decision-making applications [15] and in global health initiatives [13]. Outside the context of sustainability, some research considers that AI systems could be intrinsically valuable [50-52]. We find that the sustainability of intrinsically valuable AI systems has not yet been addressed.

Second, work on AI and sustainability is imprecise on its ethical dimensions. As illustrated in Appendices A-B, our classification of AI ethics principles and AI sustainability research involved frequent parsing of ambiguous phrasings. Indeed, outside references to the SDGs or the Brundtland definition, few publications explicitly define sustainability. Within treatments of environmental sustainability, the term "sustainability" was commonly equated with environmental protection, especially efforts to minimize energy and resource consumption, even though environmental issues do not necessarily have sustainability implications. Some work equated "sustainability" with "good for people/and or society" or even just "good", which further drains "sustainability" of its meaning. As discussed in Section 2, these are common problems with sustainability discourse. Our analysis finds that these problems have been reproduced in the AI literature.

4 The Moral Case for Long-Term, Non-Anthropocentric Sustainability and Optimization

In this section, we present our own views on the ethics of sustainability as it relates to AI. Specifically, we make the case for sustainability that is non-anthropocentric and long-term oriented. We further argue for substantial effort for this conception of sustainability, or, preferably, for the optimization of long-term intrinsic value. Finally, we explain the implications of such sustainability for AI.

Each of our arguments depends on certain positions on underlying ethical principles. As with all ethical principles, there is no universal consensus on which position to take. One can disagree with the positions we take, but doing so requires taking a different position on the underlying ethical principles. This is important to bear in mind, especially when considering the implications for AI.

4.1 Non-Anthropocentric Sustainability

First, we call for non-anthropocentric conceptions of sustainability. By non-anthropocentric, we mean that humans are not the only entities that are intrinsically valued. Modern science unambiguously shows that humans are members of the animal kingdom and part of nature. Morally significant attributes such as the innate drive to live and flourish or to experience pleasure and pain are not unique to the human species.[8] For purposes of this paper, we set aside important debates about which nonhuman entities to intrinsically value, but some potential examples include sentient nonhuman animals or (if possible) sentient AI systems, natural ecosystems, and biodiversity. We, the authors of this paper, also happen to disagree among ourselves as to which nonhumans are intrinsically valuable, but we agree that some are. There can be legitimate reasons to sometimes intrinsically value humans more than other entities—for example, a human can have a longer and richer life than a spider. However, we see no morally sound reasons to refuse to intrinsically value any nonhuman lives or entities. This means that sustainability should be defined as to also sustain some nonhumans for their own sake: it is not enough to sustain nonhumans for their instrumental role for humans.[9]

This non-anthropocentrism is at odds with common conceptions of sustainability, including that of the Brundtland Report. In failing to intrinsically value anything other than humans, we believe these conceptions of sustainability are in moral error. It is unfortunate but not surprising that similar anthropocentric tendencies are found within existing work on AI and sustainability (Section 3). Future work on AI and sustainability should be more inclusive of the intrinsic value of nonhumans.

8 Additionally, some conceptions of intrinsic value are rooted in the attributes of systems, such as interdependencies of biotic and abiotic entities within ecosystems. These holistic conceptions of intrinsic value are also not specific to humans [53].

9 For a more detailed argument for non-anthropocentrism advanced within AI ethics, see [32].

4.2 Long-Term Sustainability

Second, we call for sustainability over long time scales. Our motivation for this is an ethical principle of equality across time. In essence, this means that no one or no thing of moral significance should be disadvantaged because of the time in which they happen to exist. A person (for example) is of the same intrinsic value regardless of whether they live in the year 2021 or 2051 or 2151 or even 22021 or any other future time [54-55]. This perspective can be justified, for example, by a "veil of ignorance" thought experiment in which one does not know in advance which time period one would exist in [56]. Under such hypothetical circumstances, it would only be fair to value each time period equally. Taking temporal equality seriously means including attention to all future time periods, including the astronomically distant future. Combined with our call for non-anthropocentrism, this means sustainability should aim to sustain that which is intrinsically valuable into the distant future.

This long-termism is broadly consistent with common conceptions of sustainability, though it is different in emphasis. Common conceptions do not specify precise time scales; this is seen, for example, in the Brundtland Report's emphasis on an unspecified number of future generations. With no clear time limit, these conceptions could include the astronomically distant future, though in practice, they focus on matters that are short-term in comparison. We believe this is a moral error, an unjustified exclusion of distant-future generations and distant-future instances of anyone and anything else of intrinsic value.

4.3 Substantial Effort for Sustainability or Long-Term Optimization

Third, we call for a high degree of effort toward sustainability, or, preferably, for the optimization of long-term intrinsic value. The astronomically distant future offers astronomically large opportunities for advancing intrinsic value. These opportunities are vastly larger than those available for the present time and the near-term future. This point suggests a high degree of priority for actions oriented toward the long-term. That does not mean ignoring the present. As members of the present time period, we have special opportunities to help with present circumstances. The present also sets the stage for the future. Nonetheless, if the principle of equality across time is to be taken seriously, it requires a major focus on long-term outcomes. We further believe that people should make great efforts to advancing moral progress of all types, including sustainability, balanced mainly by the need for reasonable self-care, and that organizations and institutions should likewise be oriented accordingly.

An important perspective comes from the physics of the long-term future. Earth will become uninhabitable in roughly one billion years due to the gradual warming and expanding of the Sun [57]. Survival beyond this time can only occur in outer space. For Earth-originating entities, this will require an advanced technological civilization capable of settling in outer space. Human civilization is already positioned to accomplish this task, given its ongoing space missions and general technological progress. As long as human civilization remains intact, the ability to sustain Earth-originating entities will persist. Long-term sustainability requires resettling in outer space [28]. For the present generation, that means keeping human civilization intact.

In many contexts, there is no significant distinction between sustainability and intertemporal optimization for the distant future. Both goals require maintaining the basic functionality of civilization, including by sustaining sufficient resources and by handling major threats such as global warming, pandemics, and nuclear warfare. They likewise entail evaluating environmental threats in terms of their implications for the continuity of human

civilization and not in terms of biogeophysical disturbances or smaller-scale human consequences [58-59]. However, looking ahead, the goals point in different directions. Sustaining Earth-originating entities into the distant future only requires some minimal space settlement over very long timescales. In contrast, optimization of long-term intrinsic value entails space expansion sooner and at larger scales, in order to fill the universe with whatever is intrinsically valuable.

The distinction between sustainability and optimization of long-term intrinsic value is also important in terms of what is intrinsically valuable. Long-term sustainability can entail the same course of action for both anthropocentric and non-anthropocentric conceptions of intrinsic value: If humanity fails to settle in outer space, then other Earth-originating entities would also die out in a billion or so years, if not sooner [28,60]. However, long-term optimization generally entails expansion into outer space—but expansion in what way? Anthropocentrism would entail expansion of human populations, whereas non-anthropocentrism would entail expansion of something else.

4.4 Implications for AI

AI has several important roles to play in the story outlined above. First, current and near-term forms of AI can be applied to addressing certain immediate threats to global civilization. For example, AI is in active use for addressing global warming and environmental protection in a variety of ways, and additional ways have been identified and called for [21]. AI is also in active use for addressing the ongoing COVID-19 pandemic by supporting tasks such as medical analysis [61] and robotics to support social distancing [62]. Further work along these lines could be of value for improving the resilience pf human civilization to COVID-19 and future pandemics. Whereas global warming is a traditional environmental sustainability topic, pandemics are not, though pandemics can derive from environmental activities, in particular those that put humans in contact with novel zoonotic pathogens. Nonetheless, both issues threaten the ability of global human civilization to be sustained into the long-term future.

Second, future forms of AI could be particularly consequential. The field of AI has long entertained notions of extreme future AI that could be "the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control" [63, p.33]. Recently, there has been debate on the extent to which people in AI should focus on near-term or long-term AI. To some extent, this debate may be unnecessary, due to the existence of activities that are good to do for both near-term and long-term AI [24, 26-27]. Nonetheless, to the extent that each type of AI merits some distinct attention, a case can be made for attention to long-term AI due to its potential importance for long-term sustainability.

Long-term AI can play three roles of relevance to long-term sustainability. First, it could bolster efforts to address threats such as global warming and pandemics. Second, it could pose a threat of its own, especially for runaway AI scenarios in which the AI effectively takes over the world. Third, it could play an instrumental role in space expansion.

A significant dilemma exists for the dual status of long-term AI as both threat and tool for addressing other threats. Ideally, long-term AI would be designed slowly and carefully to ensure a high standard for safety and ethics. However, delaying the deployment of long-term AI reduces the potential for its use to address other threats. One implication of this is that other work to address other threats can be of value for "buying time" to safely and ethically develop long-term AI [64]. This includes work using near-term AI to address these other threats.

If there is extended time available to slowly and carefully design long-term AI, then that also buys time to reflect on what should be done with respect to space expansion and related opportunities [65]. On the other hand, there is no guarantee that an extended time will be available. Indeed, AI research and development is proceeding at brisk pace, prompting concerns about a race to develop long-term AI [66]. One proposed means of buying some time is to deploy a moderately powerful AI "nanny" who can protect and support humanity while it reflects on what to do next [67]. That possibility is not without its own risks, such as the risk of a poorly designed nanny AI that steers the world in a bad or even catastrophic direction. These are all among the AI issues that can be of profound importance for long-term sustainability.

5 Conclusion

In this paper, we have surveyed the ethics of sustainability, analyzed the ethical basis of existing work on AI and sustainability, and presented an argument for a non-anthropocentric, long-term conception of sustainability and an accompanying argument for favoring optimization over sustainability. Taken together, the paper provides some guidance on how ongoing work on AI and sustainability can and should proceed. First, work on AI and sustainability should precisely specify its ethical basis, in particular on what it seeks to sustain, for how long, and how much effort should be made on sustainability. Second, work on AI and sustainability should consider adopting the non-anthropocentric, long-term conception of sustainability and optimization that this paper argues for. In practice, that entails a focus on applying AI to addressing major global threats such as global warming and pandemics, to ensuring the long-term sustainability of the resource base needed for civilization, and for pursuing opportunities to expand human civilization into outer space.

In closing, we wish to emphasize the ethical principle of equality across time. A fundamental aspect of sustainability is its future-orientation. The principle of equality across time means that all future times should be treated equally, or rather that there should be no bias against something just because of when it exists. This is a compelling moral principle. If it is to be taken seriously, it demands an attention to the big-picture distant future of the universe and to the ways in which near-term actions can affect it. Actions involving AI are among the most significant ways to affect the distant future. The field of AI has special opportunities to make an astronomically large positive difference—to make the universe a better place. It should make pursuit of these opportunities a major priority.

Acknowledgments. Robert de Neufville, three anonymous reviewers, and one anonymous program committee member provided helpful feedback on earlier versions of this paper. Any remaining errors are the authors' alone.

Appendix A: Examples of AI Ethics Principles

Section 3.1 presents data on sustainability in AI ethics principles. This appendix presents some illustrative examples of these data.

The European Commission Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems includes the principle, "Sustainability: AI technology must be in line with the human responsibility to ensure the basic preconditions for life on our planet, continued prospering for mankind and preservation of a good environment for future generations." The reference to "the basic preconditions for life on our planet" implies intrinsic value of nonhumans, specifically a concern for all life on Earth. The reference to the "prospering for mankind" implies intrinsic value of humans (albeit with an unfortunate gendered phrasing). Therefore, this statement is classified as intrinsically valuing both humans and nonhumans. The reference to "future generations" implies time scales of decades or longer, though how many future generations is not specified. Nothing in the statement clarifies the degree of effort to be placed on advancing sustainability.

The *Ethics Guidelines for Trustworthy AI* by the High-Level Expert Group on AI includes the two principles, "Beneficence: Promoting Well-Being, Preserving Dignity, and Sustaining the Planet: the prominence of these principles of beneficence firmly underlines the central importance of promoting the well-being of people and the planet" and "Societal and environmental well-being: AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully considered." This principle is also classified as intrinsically valuing both humans and nonhumans. The lines "promoting the wellbeing of people and planet", "societal and environmental wellbeing", and "they should take into account the environment, including other living beings" imply intrinsic value of nonhumans, whereas the line "AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly" indicate intrinsic value of humans and instrumental value of nonhumans. As above, the reference to "future generations" implies an undefined time scale, and nothing in the principles clarifies the degree of effort toward sustainability.

The IEEE *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 1* includes the principle, "Through affordable and universal access to communications networks and the Internet, autonomous and intelligent systems can be made available to and benefit populations anywhere. They can significantly alter institutions and institutional relationships toward more human-centric structures, and they can address humanitarian and sustainable development issues resulting in increased individual, societal and environmental well-being. Such efforts could be facilitated through the recognition of and adherence to established indicators of societal flourishing such as the United Nations Sustainable Development Goals so that human well-being is utilized as a primary success criteria for A/IS development." This principle is classified as intrinsically valuing humans. The end of the line "increased individual, societal and environmental well-being" could indicate intrinsically valuing nonhumans, but the otherwise heavy emphasis on human wellbeing, human benefits, and societal flourishing, and the reference to the SDGs, strongly indicate that the natural environment is valued instrumentally to advance human wellbeing. Nothing in the statement clarifies the time scales of sustainability or degree of effort.

The UNI Global *Top 10 Principles for Ethical AI* includes the principle, "Make AI serve people and planet: This includes codes of ethics for the development, application and use of AI so that throughout their entire operational process, AI systems remain compatible and increase the principles of human dignity, integrity, freedom, privacy and cultural and gender diversity, as well as with fundamental human rights. In addition, AI systems must protect and even improve our planet's ecosystems and biodiversity." This principle is classified as intrinsically valuing humans, ecosystems and biodiversity as the final line calls for not only protection but improvement of nonhuman entities, and the phrasing "make AI serve people and planet" strongly suggest moral consideration for also nonhumans.

Appendix B: Examples of AI Sustainability Publications

Section 3.2 presents data on publications on the nexus of AI and sustainability. This appendix presents some illustrative examples of these data.

Theodorou et al. [68] developed a single-player game designed to simulate a sustainable world based on accurate ecological models and behavior economics principles. In the game, individual people must aim to act toward a sustainable society. The paper describes the game in terms implying that natural resources play an instrumental role to advance individual people's survival and wellbeing. Other indications of what is meant by a sustainable world is discussed in social and economic terms. This publication was, therefore, classified as implying intrinsic value of humans and instrumental value of nonhumans.

Yigitcanlar & Cugurullo [69] define smart and sustainable cities as "an urban locality functioning as a robust system of systems with sustainable practices, supported by community, technology, and policy, to generate desired outcomes and futures for all humans and non-humans". This publication is classified as intrinsically valuing humans and nonhumans as it explicitly refers to their benefits.

van Wysnberghe [7] calls for a "third wave" in AI ethics and states that "This third wave must place *sustainable* development at its core" (emphasis original). Sustainable development is in turn defined as in the Brundtland report. This suggests an anthropocentric conception of sustainability, in which only humans are intrinsically valuable. It further suggests placing a high degree of effort on sustainability, though the emphasis on sustainable development leaves open the question of the relative importance of present vs. future generations.

Gomes et al. [70] argue that "computational sustainability harnesses computing and artificial intelligence for human well-being and the protection of our planet" and that "planning for sustainable development encompasses complex interdisciplinary decisions spanning a range of questions concerning human well-being, infrastructure (…) and the environmental protection of the Earth and its species". All of this is strongly suggestive of intrinsic value of nonhumans as well as humans, so this publication is classified as such.

Zhang et al. [71] studies the contributions of big data analytics capability and artificial intelligence capability to the sustainability of organizational development. The publication does not define sustainability and apply the following uses of "sustainability during the abstract and introduction only: "sustainable innovation and performance", "sustainability development projects", "sustainability design and commercialization processes", "sustainable growth and performance", "sustainable organizational growth", "sustainable competitive advantages", "sustainable investment", "sustainable development goals", "sustainable

positional advantages", and big data as "sustainable resources". This publication is therefore classified as ambiguous.

Larsson et al. [72] by the AI Sustainability Center defines sustainable AI as follows: "The AI Sustainability Center supports an approach in which the positive and negative impacts of AI on people and society are as important as the commercial benefits or efficiency gains. We call it Sustainable AI." This publication is classified as intrinsically valuing humans as it discusses and defines sustainability as pertaining to human and social aspects only, including sustainable AI which is defined by accountability, bias, malicious use, and transparency. The publication is further noteworthy for presenting an extensive literature review of AI ethics as a review of literature on sustainable AI, thus equating AI ethics or ethical AI with sustainable AI, which arguably drains "sustainability" of meaning.

References

[1] Becker, C.: Sustainability Ethics and Sustainability Research. Springer (2012)

[2] Jobin, A., Ienca, M., & Vayena, E.: The global landscape of AI ethics guidelines. Nature Machine Intelligence, Vol 1, pp. 389-399 (2019)

[3] Nishant, R. Kennedy, M., & Corbett, J.: Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. International Journal of Information Management, Vol 53, 102104 (2020)

[4] Liao, H-T., & Wang, Z.: Sustainability and artificial intelligence: Necessary, challenging, and promising intersections. 2020 Management Science Informatization and Economic Innovation Development Conference (MSIEID) (2020)

[5] Dauvergne, P.: AI in the Wild: Sustainability in the age of artificial intelligence. MIT Press (2020)

[6] Galaz, V. et al.: Machine intelligence, systemic risks, and sustainability. Beijer Discussion Paper Series No. 274. Beijer Institute of Ecological Economics (2021)

[7] van Wynsberghe, A.: Sustainable AI: AI for sustainability and the sustainability of AI. AI & Ethics, Vol 1, pp. 213-218 (2021)

[8] Sætra, H.S.: AI in context and the Sustainable Development Goals: Factoring in the unsustainability of the sociotechnical system. Sustainability, Vol 13, 1738 (2021a)

[9] Sætra, H.S.: A framework for evaluating and disclosing the ESG related impacts of AI with the SDGs. Sustainability, Vol 13, 8503 (2021b)

[10] Gupta, S. et al.: Assessing whether artificial intelligence is an enabler or an inhibitor of sustainability at indicator level. Transportation Engineering, Vol 4, 100064 (2021)

[11] Kim, J., Sunghae, J., Jang, D., & Park, S.: Sustainable technology analysis of artificial intelligence using Bayesian and Social network models. Sustainability, Vol 10, 115 (2018)

[12] Bjørlo, J., Moen, Ø., & Pasquine, M.: The role of consumer autonomy in developing sustainable AI: A conceptual framework. Sustainability, Vol 13, 2332 (2021)

[13] Hadley, T.D., Pettit, R.W., Malik, T., Khoei, A.A., & Salihu, H.M.: Artificial intelligence in global health – A framework and strategy for adoption and sustainability. International Journal of Maternal and Child Health and AIDS, Vol 9, Issue 1, pp. 121-127 (2020)

[14] O'Connor, J., & Wilson, N.E.: Reduced demand uncertainty and the sustainability of collusion: How AI could affect competition. Information Economics and Policy, Vol 54, 100882 (2021)

[15] Wu, J., & Shang, S.: Managing uncertainty in AI-enabled decision making and achieving sustainability. Sustainability, Vol 12, 8758 (2020)

[16] Lässig, J., et al.: Computational Sustainability. Springer (2016)
[17] Fisher, D.H.: A selected summary of AI for computational sustainability. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17) (2017)
[18] Gomes, C., et al.: Computational sustainability: Computing for a better world and a sustainable future. Communications of the ACM, Vol 62, Issue 9, pp. 56-65 (2019)
[19] Salcedo-Sanz, S., Cuadra, L., & Vermeij, M.J.: A review of computational intelligence techniques in coral reef-related applications. Ecological Informatics, Vol 32, pp. 107–123 (2016)
[20] Lamba, A., Cassey, P., Segaran, R.R., & Koh, L.P.: Deep learning for environmental conservation. Current Biology, Vol 29, Issue 19, pp. R977-R982 (2019)
[21] Rolnick, D., Donti, P.L., Kaack, L.H., Kochanski, K., Lacoste, A., Sankaran, K., et al.: Tackling climate change with machine learning. https://arxiv.org/abs/1906.05433 (2019)
[22] Coecklebergh, M.: AI for climate: Freedom, justice, and other ethical and political challenges. AI and Ethics, Vol 1, pp. 67-72 (2021)
[23] Strubell, E., Ganesh, A., & McCallum, A.: Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3645–3650 (2019)
[24] Baum, S.D.: Reconciliation between factions focused on near-term and long-term artificial intelligence. AI & Society, Vol. 33, No. 4 (November), pp. 565-572 (2018)
[25] Baum, S.D.: Medium-term artificial intelligence and society. Information, Vol. 11, No. 6, pp. 290 (2020)
[26] Cave, S., & ÓhÉigeartaigh, S.S.: Bridging near-and long-term concerns about AI. Nature Machine Intelligence, Vol 1, Issue 1, pp. 5-6 (2019)
[27] Stix, C., & Maas, M.M.: Bridging the gap: the case for an 'Incompletely Theorized Agreement' on AI policy. AI & Ethics, https://doi.org/10.1007/s43681-020-00037-w (2021)
[28] Tonn, B.E.: Futures sustainability. Futures, Vol 39, pp. 1097-1116 (2007)
[29] Dubber, M.D., Pasquale, F., & Das, S. (Eds.): The Oxford Handbook of Ethics of AI. Oxford University Press (2020)
[30] Liao, S.M. (Ed.): Ethics of Artificial Intelligence. Oxford University Press (2020)
[31] Adamson, G., Havens, J.C., & Chatila, R.: Designing a value-driven future for ethical autonomous and intelligent systems. Proceedings of the IEEE, Vol 107, Issue 3, pp. 518-525 (2019)
[32] Owe, A., & Baum, S.D.: Moral consideration of nonhumans in the ethics of artificial intelligence. AI & Ethics, https://doi.org/10.1007/s43681-021-00065-0 (2021)
[33] von Carlowitz, H.C.: Sylvicultura Oeconomica, oder haußwirthliche Nachricht und Naturmäßige Anweisung zur wilden Baum-Zucht. Meißen (1713)
[34] Hotelling, H.: The economics of exhaustible resources. Journal of Political Economy, Vol 39, Issue 2, pp. 137-175 (1931)
[35] Franco, M.P., Gaspard, M., & Mueller, T.: Time discounting in Harold Hotelling's approach to natural resource economics: The unsolved ethical question. Ecological Economics, Vol 163, pp. 52-60 (2019)
[36] Pezzey, J.: Sustainability: An interdisciplinary guide. Environmental Values, Vol 1, Issue 4, pp. 321-362(42) (1992)
[37] Dryzek, J.S.: The Politics of the Earth: Environmental Discourses. 3rd Edition. Oxford University Press (2013)
[38] Mensah, J.: Sustainable development: Meaning, history, principles, pillars, and implications for human action: Literature review. Cogent Social Sciences, Vol 5, Issue 1 (2019)

[39] Kuhlman, T., & Farrington, J.: What is sustainability? Sustainability, Vol 2, Issue 11, pp. 3436-3448 (2010)
[40] Zimmerman, M.J., & Bradley, B.: Intrinsic vs. extrinsic value. In Zalta, E.N. (Ed.) Stanford Encyclopedia of Philosophy, Spring 2019 Edition. https://plato.stanford.edu/archives/spr2019/entries/value-intrinsic-extrinsic (2019)
[41] Washington, W., Taylor, B., Kopnina, H.N., Cryer, P., & Piccolo, J.J.: Why ecocentrism is the key pathway to sustainability. Ecological Citizen, Vol 1, Issue 1, pp. 35-41 (2017)
[42] Mosquin, T., & Rowe, S.: A manifesto for Earth. Biodiversity, Vol 5, Issue 1 (2004)
[43] Lind, R.C. (Ed.): Discounting for time and risk in energy policy. Resources for the Future, US (1982)
[44] Portney, P., & Weyant, J. (Eds.): Discounting and intergenerational equity. Resources For the Future, US (1999)
[45] Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S.: The role and limits of principles in AI ethics: Towards a focus on tensions. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 195-200 (2019)
[46] Heyd, D.: "Supererogation", The Stanford Encyclopedia of Philosophy (Winter 2019 Edition), Zalta, E.N. (Ed.) (2019)
[47] Sinnot-Armstrong, W.: "Consequentialism", The Stanford Encyclopedia of Philosophy (Fall 2021 Edition), Zalta, E.N. (Ed.) (2021)
[48] Guest, R.: The economics of sustainability in the context of climate change: An overview. Journal of World Business, Vol 45, Issue 4, pp. 326-335 (2010)
[49] Grant, M.J., & Booth, A.: A typology of reviews: An analysis of 14 review types and associated methodologies. Health Information & Libraries Journal, Vol 26, Issue 2, pp. 91-108 (2009)
[50] Coeckelbergh, M.: Robot rights? Towards a social-relational justification of moral consideration. Ethics of Information and Technology, Vol 12, pp. 209-221 (2010)
[51] Gunkel, D.J.: Robot Rights? The MIT Press (2018)
[52] Danaher, J.: Welcoming robots into the moral circle: A defence of ethical behaviorism. Science and Engineering Ethics, Vol 26, Issue 4, pp. 2023–2049 (2020)
[53] Rolston, H.,III.: Environmental Ethics: Duties to and Values in the Natural World. Temple University Press (1988)
[54] Cowen, T., & Parfit, D.: Against the social discount rate. In: Laslett, P., & Fishkin, J.S. (Eds.), Justice Between Age Groups and Generations. Yale University Press, pp. 144–161 (1992)
[55] Tonn, B.E.: Philosophical, institutional, and decision making frameworks for meeting obligations to future generations. Futures, Vol 95, pp. 44-57 (2018)
[56] Tremmel, J.C.: The convention of representatives of all generations under the 'veil of ignorance'. Constellations, Vol 20, Issue 3, pp. 483-502 (2013)
[57] Wolf, E.T., & Toon, O.B.: The evolution of habitable climates under the brightening sun. Journal of Geophysical Research: Atmospheres, Vol 120, Issue 12, pp. 5775-94 (2015)
[58] Baum, S.D., & Handoh, I.C.: Integrating the planetary boundaries and global catastrophic risk paradigms. Ecological Economics, Vol 107, pp. 13-21 (2014)
[59] Beard, S.J., Holt, L., Tzachor, A., Kemp, L., Avin, S., Torres, P., & Belfield, H.: Assessing climate change's contribution to global catastrophic risk. Futures, Vol 127, 102673 (2021)
[60] Baum, S.D.: Is humanity doomed? Insights from astrobiology. Sustainability, Vol 2, Issue 2, pp. 591-603 (2010)
[61] Hussain, A.A., Bouachir, O., Al-Turjman, F., & Aloqaily, M.: AI techniques for COVID-19. IEEE Access, Vol 8, pp. 128776-128795 (2020)

[62] Javaid, M., Haleem, A., Vaish, A., Raju, V., & Iyengar, K.P.: Robotics applications in COVID-19: A review. Journal of Industrial Integration and Management (2020)
[63] Good, I.J.: Speculations concerning the first ultraintelligent machine. Advances in Computers, Vol 6, pp. 31-88 (1966)
[64] Baum, S.D.: The great downside dilemma for risky emerging technologies. Physica Scripta, Vol. 89, No. 12 (December), 128004 (2014)
[65] Ord, T.: The Precipice: Existential Risk and the Future of Humanity. Bloomsbury Publishing (2020)
[66] de Neufville, R., & Baum, S.D.: Collective action on artificial intelligence: A primer and review. Technology in Society, Vol 66 (August), 101649 (2021)
[67] Goertzel, B.: Should humanity build a global AI nanny to delay the singularity until it's better understood? Journal of Consciousness Studies, Vol, 19, Issue 1-2, pp. 96-111 (2012)
[68] Theodorou, A., Bandt-Law, B., & Bryson, J.J.: The sustainability game: AI technology as an intervention for public understanding of cooperative investment. 2019 IEEE Conference on Games (CoG), pp. 1-4 (2019)
[69] Yigitcanlar, T., & Cugurello, F.: The sustainability of artificial intelligence: An urbanistic viewpoint from the lens of smart and sustainable cities. Sustainability, Vol 12, 8548 (2020)
[70] Gomes, C.P., Fink, D., van Dover, R.B., & Gregoire, J.M.: Computational sustainability meets materials science. Nature Reviews: Materials, Vol 6, August, pp.645-647 (2021)
[71] Zhang, H., Song, M., & He, H.: Achieving the success of sustainability development projects through big data analytics and artificial intelligence capability. Sustainability, Vol 12, Issue 3, pp. 949 (2020)
[72] Larsson, S., Anneroth, M., Felländer, A., Felländer-Tsai, L., Heintz, F., & Cedering Ångström, R.: Sustainable AI: An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence. AI Sustainability Center (2019)