Article

From an Ethics of Carefulness to an Ethics of Desirability: Going Beyond Current Ethics Approaches to Sustainable AI

Larissa Bolte *, Tijs Vandemeulebroucke and Aimee van Wynsberghe

The Sustainable AI Lab, Institute for Science and Ethics, University of Bonn, Bonner Talweg 57, 53113 Bonn, Germany; tvandeme@uni-bonn.de (T.V.); aimee@uni-bonn.de (A.v.W.)
* Correspondence: bolte@iwe.uni-bonn.de

Abstract: ‘Sustainable AI’ sets itself apart from other AI ethics frameworks by its inherent regard for the ecological costs of AI, a concern that has so far been woefully overlooked in the policy space. Recently, two German-based research and advocacy institutions have published a joint report on Sustainability Criteria for Artificial Intelligence. This is, to our knowledge, the first AI ethics document in the policy space that puts sustainability at the center of its considerations. We take this as an opportunity to highlight the foundational problems we see in current debates about AI ethics guidelines. Although we do believe the concept of sustainability has the potential to introduce a paradigm shift, we question whether the suggestions and conceptual grounding found in this report have the strength to usher it in. We show this by presenting this new report as an example of current approaches to AI ethics and identify the problems of this approach, which we will describe as ‘checklist ethics’ and ‘ethics of carefulness’. We argue to opt for an ‘ethics of desirability’ approach. This can be completed, we suggest, by reconceptualizing sustainability as a property of complex systems. Finally, we offer a set of indications for further research.

Keywords: sustainable AI; artificial intelligence; AI ethics; checklist ethics; ethics of carefulness; ethics of desirability

1. Introduction

The ethics of Artificial Intelligence (AI) and Machine Learning (ML) have grown in significance within academia, industry, and policy making. Academics continue to point towards the negative consequences associated with the design and development of AI/ML, such as risks to fairness when biased data are used to train the AI model [1]. Although there exists over a hundred different sets of AI ethics principles and guidelines to steer the ethical (trustworthy, responsible) design and implementation of the technology [2], little has been said, in these guidelines, about the environmental sustainability of AI/ML. To be sure, there are a range of environmental consequences associated with the training and usage of AI/ML; carbon emissions and electricity consumption from running the algorithm [3,4], mining of precious minerals [5,6] for the development of different technologies making up the AI creating further environmental damage, land usage and water usage [6], as well as the resulting electronic waste when parts of the infrastructure are no longer needed [7]. Given that it is often marginalized and vulnerable demographics who bear the consequences of these impacts on climate [8], we insist that this is not strictly a technical issue—generating numbers about these consequences—but a moral issue.

In the academic space, the field of Sustainable AI has recently gained traction with Special Issues (such as the one this paper is submitted to), conferences, summer schools, and publications. In 2021, van Wynsberghe suggested that we understand Sustainable AI along two dimensions: AI for sustainability (e.g., to achieve the sustainable development goals) and sustainability of AI (i.e., the impact of making and using AI on the environment) [9]. Although there is a growing body of academic research in this space, there
is still a limited amount of work performed in the policy making space. Recently, two German-based research and advocacy institutions, namely the Institut für ökologische Wirtschaftsforschung and AlgorithmWatch’s SustAIn project, have published a joint report on sustainability criteria for artificial intelligence (henceforth “the report”) [10]. This report is novel in its focus on sustainability as the grounding value through which AI should be evaluated. It is the purpose of this paper to unpack the findings in this report as well as to draw links to what this means for the future of Sustainable AI in the policy making space.

In the following paper, we argue that a new approach to AI ethics is needed and that ‘sustainability’, properly construed, can provide that new approach. We begin by presenting current approaches to AI ethics guidelines and their commonalities. We claim that, despite the abundance of existing AI policy frameworks, ‘Sustainable AI’ deserves consideration. We, furthermore, present problems with AI ethics guideline documents that have been identified in current literature and that, we hold, can be addressed by a proper notion of sustainability. We present the report as one of the first policy documents to address the issue of sustainable AI. We show how the concept of sustainability is understood and ultimately used to structure the report’s arguments. We find that the report could not overcome the problems we have presented and identify specific AI ethics paradigms as problematic, namely a push towards checklist ethics, and an ethics of carefulness. An isolationist view of technology underlies both. We see this view perpetuated in the report. We then present an AI ethics of desirability as an alternative and conclude by sketching a notion of sustainability as a property of complex systems, which both addresses pressing environmental issues in the context of AI and is open to be complemented by such ethics of desirability.

2. Current Approaches to AI Ethics Guidelines: Why ‘Sustainable AI’?

Becoming a leader in AI/ML technology is by many companies and nation states perceived as a major strategic advantage, so much so that the rhetoric of an AI race has been well established [11]. The rapid acceleration of both research into and implementation of AI technologies and their transformative power underline the urgency of proper ethical guidance. Major public and private stakeholders, such as governments, NGOs, and AI development companies tackle this concern by issuing AI ethics guideline documents. The AI ethics guideline space is ever-growing and already saturated with frameworks, for example Responsible AI, Explainable AI, Trustworthy AI, and AI4Good. The German research and advocacy group AlgorithmWatch currently lists 167 documents in their AI Ethics Guidelines Global Inventory [2]. Despite this overwhelming number of contributions, similarities seem to emerge. Fjeld et al. find that principles found in AI principle documents can be grouped according to the themes of ‘privacy’, ‘accountability’, ‘safety and security’, ‘transparency and explainability’, ‘fairness and non-discrimination’, ‘human control of technology’, ‘professional responsibility’, and ‘promotion of human values’ [12]. Mittelstadt, on the other hand, notes that public–private initiatives seem to converge in their reports on the familiar principles of biomedical ethics: ‘autonomy’, ‘beneficence’, ‘non-maleficence’, and ‘justice’ [13,14]. It is so typical of AI ethics guideline documents to present their propositions in the form of (mutually disjoint) principles that, to our knowledge, all scoping reviews on the topic either group principles or assess the prominence of certain principles [12,13,15].

Given the abundance of existing AI ethics frameworks, one may ask why Sustainable AI, yet another framework, constitutes a justified addition to public debate. What sets Sustainable AI apart from other approaches is its inherent regard for both the needs of future generations and, consequently, the ecological costs of AI, which again have social costs. The latter ethical concern has so far been woefully overlooked by previous approaches. As Jobin et al. reveal, out of 84 reviewed AI ethics policy documents, only 14 address sustainability at all [15]. Only one briefly mentions a “[…] priority of environmental protection and sustainability” [16] (p. 19) and yet another single one refers to AI’s ecological footprint [17]. Although the potential harm of the development and the use of AI is generally
acknowledged when it comes to concerns of social sustainability (e.g., issues concerning bias, explainability, or cultural sensitivity), AI’s impact on ecological sustainability is rarely discussed [9]. This disregard in policy may partially be due to the lack of research that has been conducted on the topic so far. However, this does not indicate that the sustainability of AI is a negligible concern.

Although Vinuesa et al. find that, for the three environmental Sustainable Development Goals (SDGs), namely SDG 13, 14, and 15, AI may prove beneficial for almost all goals and targets and has inhibitory potential for only 30% at most [18], there is good reason to believe that this positive outlook is deceiving. For one, Vinuesa et al. base their analysis on a literature review of existing research, which they themselves note may be biased towards more positive reporting of AI impacts. What is more, Sætra argues that Vinuesa et al.’s analysis is unable to properly account for AI technology’s environmental impact since they do not consider the scale, import, or possible indirect effects of them [19]. To be sure, there are many environmental concerns raised by AI that are yet to be properly researched and quantified. First estimates of carbon emissions produced by training just one single neural network model for Natural Language Processing suggest that higher regard must be paid, for example, to prioritizing more energy efficient hardware and models [3]. Other concerns pertain to the underlying technological infrastructure required for AI development and use, such as water scarcity due to mining for components such as lithium [6], the accumulation of toxic electronic waste [7], or pollution due to waste-water emitted from data centers [7].

What is more, we hold that the sustainability notion has the potential to address problems with current AI ethics guidelines approaches that have been identified in the academic literature. Hagendorff observes that major AI ethics guidelines conceptualize technical artefacts as “[…] isolated entities that can be optimized by experts so as to find technical solutions for technical problems” [20] (p. 103). By contrast, some philosophers of technology urge that technical artefacts must be understood as embedded parts of a socio-political system [20–23]. As a consequence, ethics tends to be perceived by developers and businesses as a hindrance to technological progress [20,24] instead of a chance to define what ‘progress’ means. Another consequence of viewing technological artifacts, AI-systems or models in particular, in isolation is that focus tends to be on their direct impacts while more indirect impacts and ripple effects, such as ecological implications, are overlooked [9,19,20]. Finally, the principle approach of AI ethics guidelines itself has been scrutinized. Mittelstadt argues that while a principlist approach has been working well in biomedical ethics, the case is different for the ethics of AI [13]. Since general AI ethics guideline documents propose principles that are high-level, they are in need of interpretation for use in a particular context [13], a concern that has also been raised against principlism in the context of biomedical ethics under the slogan “thick in status, thin in content” [25]. This is remedied in the medical context, Mittelstadt argues, by the presence of both a well-defined goal, the patient’s well-being, and a historical track record of ethical decision making, an ethos. Both are lacking in the context of AI. It is the aim of this paper to give an outlook on how these problems can be addressed by ‘Sustainable AI’, properly construed.

3. ‘Sustainable AI’ in the Policy Making Space: The SustAIn Report

Given the pressing ecological issues raised by the rapid adoption of AI technologies, it is encouraging to see that research and advocacy groups are picking up on ‘Sustainable AI’. We present here the SustAIn report as a step in the right direction, but ultimately as one that does not yet realize the full potential of the notion of sustainability. The authors of the report define ‘sustainability’ as the “[…] process that is concerned with the question of just distribution between humans living today and future generations, and the question of just behavior of humans towards one another as well as towards nature” [10] (p. 27) (original quote in Appendix A [A1]). They further specify this definition by adopting a version of the so-called Three-Pillar-Model of sustainability. On this view,
'sustainability' comes in three kinds: the ecological, the social, and the economic. The authors define 'ecological sustainability' as 'safeguarding the scope of action for humanity present and future', i.e., staying within our planetary boundaries. The normative goal is to secure equal chances to a good life for future generations. 'Social sustainability' is characterized by the fulfilment of basic human needs, a regard for living conditions, access to social infrastructure, and the security of social integrity and cohesion and, thus, by the protection of vulnerable groups, intra- and intergenerational justice, and the value of diversity. Ecological and social sustainability, the authors hold, are ‘[…] two sides of the same coin’ [10] (p. 27) (Appendix A [A2]). They contend, on the one hand, that social cohesion is a necessary condition for effective environmental protection, and, on the other hand, that a healthy environment is a necessary condition for human flourishing. Economic sustainability, finally, is defined as servicing these two dimensions. Economic activities are sustainable when they respect planetary boundaries and fulfill the needs of current and future generations. It is the task of a sustainable economy to ‘harmonize’ [10] (p. 27) (Appendix A [A3]) ecological and social concerns.

When AI-systems are developed and used, the authors write, they form part of our economic activities. As such, and according to the above definitions, sustainable AI-systems stay within the normative limits set by the concepts of ecological and social sustainability. Accordingly, the authors define ‘Sustainable AI’ as a ‘[…] system whose development and use respects planetary boundaries, does not exacerbate problematic economic dynamics and does not endanger social cohesion’ [10] (p. 30) (Appendix A [A4]). These are supposed to be minimal conditions. In short then, sustainable AI-systems are at least neutral, if not beneficial, with respect to the achievement of Three-Pillar sustainability.

Based on these definitions, the authors propose a set of 13 sustainability criteria for AI, to which they attribute several indicators and sub-indicators. The criteria are grouped according to the Three-Pillar-Model of sustainability. There is, furthermore, a set of indicators grouped under a “cross-sectional” criterion. These indicators cannot be neatly attributed to only one pillar [10] (p. 60ff). Table 1 lists the criteria, their grouping, and their original German title.

Table 1. Overview of the SustAIn report’s sustainability criteria for AI.

| Grouping            | Criteria                                                                 | German Original                                                                 |
|---------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| Social Sustainability| Transparency and Assumption of Responsibility                           | Transparenz und Verantwortungsübernahme                                        |
|                     | Non-Discrimination and Fairness                                         | Selbstbestimmung und Datenschutz                                                |
|                     | Technical Reliability and Human Oversight                                | Technische Verlässlichkeit und Menschliche Aufsicht                             |
|                     | Autonomy and Data Protection                                             | Selbstdetermination und Datenschutz                                              |
|                     | Inclusive and Participatory Design                                       | Inklusives und Partizipatives Design                                             |
|                     | Cultural Sensibility                                                    | Kulturelle Sensibilität                                                         |
| Economic Sustainability| Market Diversity and Exhaustion of Innovative Potential                | Marktwielfalt und Ausschöpfung des Innovationspotenzials                       |
|                     | Distribution Effect in Target Markets                                   | Verteilungswirkung in Zielmärkten                                               |
|                     | Working Conditions and Jobs                                             | Arbeitsbedingungen und Arbeitsplätze                                           |
| Ecological Sustainability| Energy Consumption                                                      | Energieverbrauch                                                                |
|                     | CO₂ and Greenhouse Gas Emissions                                        | CO₂- und Treibhausgasemission                                                    |
|                     | Sustainability Potentials in Application                                 | Nachhaltigkeitspotenziale in der Anwendung                                     |
|                     | Indirect Resource Consumption                                           | Indirekter Ressourcenverbrauch                                                   |
| All                 | Cross-Sectional Criterion                                              | Querschnittskriterium                                                          |

All criteria pertain to AI-systems, i.e., concrete ML models together with their particular training data. Criteria also pertain to the organizational level. They identify possible points of intervention in the development and/or use of AI-systems for organizations developing and/or using these systems. The criteria sometimes address properties of AI-systems (e.g., explainability or their energy consumption) and sometimes the practices of the developing or using organization (e.g., adoption of a code of conduct or stakeholder participation in the design process). The criteria are derived from current debates, analyses, and evaluation tools, which are not further specified in the report. The criteria can, thus, be
read as a continuation of current debates, while the notion of ‘sustainability’ adds a lens through which criteria can be reinterpreted, clustered, and ultimately applied. Although the final aim is to provide a systematic sustainability evaluation tool for AI-systems, the current report is to be conceived as a first stepping-stone inducing debate and raising awareness for traceable but not-yet-traced sustainability metrics of AI-systems.

In short, this report is an important step for the field of Sustainable AI as it goes deeper into a discussion of what the definition of Sustainable AI ought to consider and how this is situated against other traditional conceptions of sustainable development (e.g., the Three-Pillars approach). Furthermore, we argue that the report’s focus on sustainability must not be understood as yet another added concern for AI ethics, the only additions being future-oriented and a focus on ecological consequences. Instead, it appears that, through the arguments made in the report, previously raised issues can be grouped under the label ‘sustainability’. Criteria such as ‘transparency and assumption of responsibility’, ‘non-discrimination and fairness’, ‘autonomy and data protection’, or ‘working conditions and jobs’ tackle widely embraced AI ethics principles, such as transparency, justice and fairness, non-maleficence, responsibility, and privacy [26]. If these social issues can be grouped as sustainability issues, it seems fair to ask: Does the notion of sustainability have the potential to figure as a unifying umbrella concept for AI ethics as it is currently practiced and beyond? Moreover, can this notion then answer to the problems that have been raised against current AI ethics guideline documents as detailed in Section 2?

4. ‘Sustainable AI’: Perpetuating Problems with the Current AI Guidelines Paradigm

If ‘sustainability’ is understood correctly, we argue, it does indeed have the potential to induce a paradigm shift in how we regiment AI development and use. Although we do not believe that the currently analyzed report realizes this potential, we believe it has provided a necessary stepping-stone towards a deeper understanding. In the following two sections, we show how the problems that have been raised with AI ethics guidelines in general relate to two connected ethics paradigms, namely a ‘checklist approach’ and an ‘ethics of care’ that are both perpetuated by the report. We argue that these paradigms lie at the root of these problems, namely the perception of ethics as a hindrance, the disregard for indirect impacts of technology implementation, and the lack of unequivocal guidance. We then sketch a notion of ‘sustainability’ that avoids these paradigms.

4.1. Checklist Ethics

In almost all AI ethics guidelines that have been presented to date we find a common approach to ethics, namely a checklist of ethical requirements to be fulfilled. Take, for example, the European Commission High Level Expert Group (AI HLEG) who provided the Guidelines for Trustworthy AI [27]. This report starts from high level principles based on European values that must be protected throughout the design and use of AI. From this, principles are derived and, finally, an assessment list of how to operationalize these principles. As such, a move is made from abstract principles to operationalized values. This is not a new phenomenon in the ethics of technology, in fact many authors argue for such approaches [28–30]. In the case of the SustAIN report, there are 13 interconnected but ultimately stand-alone, potentially competing criteria [10].

By design, ethics checklists dissect complex situations of ethical decision making into a multitude of disconnected aspects without transparent procedure for resolution in case of conflict, necessity for prioritization or mixed performance on different criteria. A number of questions abound: How are criteria to be weighed in relation to one another? Can good performance in one criterion offset bad performance in another? How many criteria have to be fulfilled in order to merit the label ‘ethical’ (or ‘trustworthy’, ‘responsible’, ‘sustainable’, etc.)? Without a clear, unifying normative framework to support them, checklists risk being ultimately uninformative. Hence, Mittelstadt’s assessment holds: AI ethics guidelines are lacking a supporting goal, ethos, or both [13]. To be sure, specific approaches, such as value sensitive design, have been struggling with such questions for decades now [31]. In the
case of the report, we understand that the authors propose their criteria and indicator set as a basis for further discussion, not yet as a systematic evaluation tool. It is, however, important to point out that the tendency towards disconnected principles is ingrained within their approach.

Moreover, the concept of sustainability itself appears disunified. The task for Sustainable AI must, therefore, be to offer a unifying alternative to current approaches, rather than introduce even more conceptual scattering. Admittedly, the report’s authors note this problem. They try to remedy it by choosing a conceptualization of ‘sustainability’ that introduces a normative order. This normative order consists of a clear prioritization of two pillars over the other. Although we consider this a viable approach, unfortunately, we are still unclear about the ultimate normative foundation that the authors have in mind. Their definition of ‘ecological sustainability’ implies that the ultimate duty of sustainability is to safeguard human action potential, or, in other words, to create the conditions in which humanity can flourish. This is a distinctly anthropocentric view of sustainability. Here, it is social sustainability that takes precedence over both the ecological and the economic dimension.

In this iteration, it is unclear how ‘Sustainable AI’ sets itself apart from other approaches to AI ethics in the policy space at all. Take, for example, ‘Trustworthy AI’ as embraced by the AI HLEG: The Commission argues that the principles of ‘Fairness’ and ‘Prevention of Harm’ comprise the demand to encourage sustainability and ecological responsibility of AI-systems, as well as the duty to use AI systems to the benefit of all human beings, including future generations. They even consider non-human living beings [27] (p. 19). Under this interpretation then, the center-stage role of sustainability in AI ethics seems ill-suited or at least redundant.

The authors of the SustAIn report do, however, seem to insist that ecological sustainability has normative force beyond its connection to social sustainability. After all, ecological and social concerns need to be “harmonized” by the economy. If ecological concerns were simply part of social sustainability, as the definition suggests, there would be nothing to harmonize. To further underline this point, the authors’ definition of ‘sustainability’ as “just behavior of humans towards one another as well as towards nature” is noteworthy. This formulation, again, begs a couple of questions: What exactly do we owe to nature? Do the authors follow the opinion that nature, the environment, or the preservation of ecosystems and biodiversity are ends in themselves? More conceptual clarity is needed in order to assess the merit of the sustainability concept for AI ethics. In other words, does social sustainability, properly construed with ecological limits in mind, trump all other concerns? Or are there two equally compelling sustainability demands? Do they go for a weak or strong sustainability approach? If these questions cannot be answered, the sustainability concept will remain disunified and a checklist approach, at least towards ‘Sustainable AI’, appears unavoidable.

4.2. Ethics of Carefulness

Not only does the report adhere to a checklist approach towards ethics; it also seems to adhere to what can be understood as an ‘ethics of carefulness’. Vandemeulebroucke et al. define this approach as an “[...] ethics which works from inside the technological paradigm” [32] (p.34)[33] and, as such, one that is looking to render (inevitable) technology design, development, and use careful and safe. Adopting an ethics of carefulness for AI implies the premise that these AI-systems are unquestionable and inevitable givens. The best one can hope for is to establish ethical criteria which guarantee a careful design, development and use of AI, in order to avoid its sharp edges. From this vantage point, society and natural environments have to adapt to AI instead of the other way around. Ethics has to rein in what is evident in the continuous expansion of the focus of ethics now. What used to be a sole focus on the use of a technology, nowadays also includes a focus on design and development. AI, as a technology, has a functional essence independent of the, in reference to the three pillars of sustainability, eco–socio–economic context of
its application. Ethical considerations are not at the essence of the technology, which makes them costly, limiting add-ons [34]. An ethics of carefulness is, thus, a direct result of a certain conceptualization of technology as isolated artefacts. Moreover, an ethics of carefulness is the underlying scheme of many checklist approaches to ethics guidelines on AI.

We hold that the report subscribes to an ethics of carefulness. This is evident for two reasons. The first comes from a closer inspection of the report’s definition of ‘Sustainable AI’: Sustainable AI, in the authors’ view, shall, at least, not be harmful to the environment, society, and the economy. The focus on carefulness and especially safety is evident. Additionally, it must be noted that this definition is a negative one; it does not offer any attempt to shape technology in line with positive values or a greater vision for the future design, development, and use of the technology [35,36]. As such, AI, as a technology, is posited as-is and only considered in its immediate, potentially harmful effects. As was the case with the checklist approach, this view proves pervasive throughout the AI policy sphere. Jobin et al. point out: “Because references to non-maleficence outnumber those related to beneficence, it appears that issuers of [AI ethics] guidelines are preoccupied with the moral obligation to prevent harm” [15] (p. 396).

Our second reason is more implicit and stems from the level of analysis that a majority of AI ethics guidelines and also the report’s authors have chosen. They decide to focus their attention on and the application of their principles to the sustainability of particular AI-systems (e.g., [16,37,38]). This focus comes naturally if one views technologies as static givens, ultimately uninfluenced by the eco–socio–economic context of their development and use. Particular instances of the technology then are viewed as instantiations that carry the same essence and, thus, the same properties and ultimately effects. Furthermore, if one believes that the essence of a technology cannot be changed, it is only natural to assume that the workings of its instances are where one needs to intervene.

Admittedly, this level of analysis has one clear advantage, as the authors themselves point out: It allows for identifying very clear points of intervention in the concrete process of developing and/or using AI. It, thus, puts the focus on those agents in whose hands the technology lies primarily and who are consequently the agents responsible to change their procedures for the better [39]. Still, this perspective also has a clear disadvantage: It fails to consider the broader ‘AI-ecosystem’. AI-systems are never developed or used in a vacuum. Just as farming and clothing production become much more problematic when they happen en masse, ‘mass AI’ also comes with its very own problems. One major concern on the ecological side is the excessive energy consumption that is to be expected when AI is implemented on a global scale [9]. A set of sustainability criteria that focuses on singular AI-systems is, by virtue of its approach, unable to address proportionality concerns and is, thus, blind to certain indirect impacts and ripple effects. If we want to attend to the sustainability of AI, and ultimately AI ethics, on a broader scale, we need to ask whether the development and use of a particular AI-system is justifiable in the first place, given the current AI landscape. In order to make this assessment, however, it is not enough to scrutinize the AI-system under consideration in isolation.

5. Towards an Ethics of Desirability

Although both the checklist ethics and the ethics of carefulness perspective certainly have their merits in specific contexts, we believe a different approach is necessary that can more naturally account for both the interconnectedness of ethical concerns of AI and the broader eco–socio–economic context of its development and use. We see this alternative in an ‘ethics of desirability’ approach. Vandemeulebroucke et al. define this approach as one that “[…] stands outside of the technological paradigm and critically questions it by taking into account the socio-political determinants that have led to the paradigm” [32], (pp. 34–35). It, thus, conceptualizes technologies as embedded in a socio-political context. This approach operates from the assumption that the technology and its context of development and use interact and co-create each other. Technologies are shaped
by social demands and social demands are shaped by technology and its functioning [40]. In order to arrive at an ethics of desirability, one first has to reconceptualize technology itself.

This reconceptualization can be found in Andrew Feenberg’s critical theory of technology. Feenberg shows us that each technological artefact needs to be perceived as the concretization of a particular eco–socio–economic context with its inherent power relations. Attributes of technological artefacts, such as ‘working’/‘not working’ or ‘efficient’/’inefficient’, must be understood in terms of social demands and perceptions. In other words, these attributes are assigned according to the set purpose. The question then becomes who decides what these attributes precisely mean and which ones are more important than others. Hence, there is a specific framing of a given technology’s problems and solutions which will heavily influence its development towards further concretization [34]. The social environment with its needs and demands within which a technology is developed thus shapes the technology’s further development. In view of the fact that our current technology has been designed in isolation from the needs, demands and values of weaker political actors, Feenberg concludes that these newly formed demands appear as a push towards more technological abstraction, meaning they make for more complicated, scattering, and expensive add-ons [34]. If we instead view technology not as static, but as developing towards more concretization, we are able to conceptualize ethics as a choice between several possible development paths.

Feenberg’s energy-efficient house serves as an example. The house performs the three separate functions of shelter, warmth, and lighting. Yet, in the energy-efficient house, these three functions are realized by the unifying structural element of being oriented towards the sun as essential design feature [34]. The demand for energy-efficiency thus steered technological progress towards more structural integration. Different development paths could have been deemed desirable. Arguably, had the demand been that the house be as cheap as possible, different unifying design features would have been chosen.

Against this backdrop, checklist ethics and an ethics of carefulness appear undesirable. These approaches to AI ethics demand that several contradictory functions be integrated under one system: the output of AI-systems shall be precise and reliable, requiring a massive amount of training data, but the system must also be maximally energy-efficient. Harmful effects of the technology must be hedged, and benefits harnessed. Moreover, as these ethics approaches work within the current technological paradigm they are unable to clearly account for the power differences between the different actors involved in the concretization of AI. They then reify the current technological paradigm and its underlying politics [41], which is evident from the current multitude of AI ethics guidelines. Hence, Mittelstadt is again correct when he asserts that current AI ethics and policy making approaches “[…] fail to address fundamental normative and political tensions embedded in key concepts (for example, fairness, privacy)” and in AI development and use itself and that “[d]eclarations by AI companies and developers committing themselves to high-level ethical principles and self-regulatory codes nonetheless provide policy-makers with a reason not to pursue new regulation.” [13] (p. 501).

An ethics of desirability, however, offers an alternative. It accentuates a pathway along which technological problem solving shall be oriented. Because it works outside the established technological paradigm, an ethics of desirability analyses current technological development paths, i.e., chains of problems and solutions, and evaluates their ethical tenability. It avoids the checklist approach and makes transparent how technological development is never a process with a pre-determined end, but rather an ethico-political choice between a multitude of possible paths. It, hence, reveals the possibility to intervene on the current path of technological development and relies on the fact that the way a technology is essentially structured and used is determined by the goals we set as technology developers and users. Hence, an ethics of desirability opens up a possible multitude for AI development and use and, as such, dereifies the current technological paradigm. One way it does this is by giving voice to those actors that are often not heard in AI ethics policies (natural environments, local human populations affected by the
development or waste management of AI technologies, etc.) to express what is desirable for
them instead of merely minimizing harm for them [41]. In an ethics of desirability, we then
need to find a suitable, ethical pathfinder to guide us through the multitude of possible
trajectories of AI development and use. If ‘sustainability’ is to serve as that pathfinder in
AI development and use, it is of utmost importance that the concept is as well defined,
unanimous, and normatively unequivocal, as demand for energy-efficiency in houses.

6. ‘Sustainability’ Understood as a Property of Complex Systems

The sustainability concept discussed in the policy space does not seem suitable to
offer a new paradigm for AI ethics that steers towards an ethics of desirability which views
technology as embedded in eco–socio–economic contexts and technological progress as
value-oriented puzzle-solving. We can now finally propose a different conceptualization
of sustainability that, we believe, holds more promise in this regard. We have argued
above that an ethics of carefulness approach invites the belief that particular instances of
a technology are where intervention needs to take place. The emphasis on this low level
of analysis obscures the role of the eco–socio–economic context within which a particular
instance of a technology operates. As such, it betrays a reductionist way of thinking that
neglects the fact that a holistic system is more than the sum of its parts. As Jørgensen et al.
lament: “If we cannot understand systems except by separating them into parts, and if
by doing this we sacrifice important properties of the whole, then we cannot understand
systems” [42] (p. 3). As an alternative, we urge the conceptualization of sustainability as a
property of complex systems and, as such, a guiding principle in AI development and use.

Inspired by Crojethovich and Rescia, we define a complex system as composed of
a great number of single elements (e.g., organisms, natural environments, technologies)
and actors (e.g., individuals, organizations, industries, political institutions) in interaction,
capable of exchanging information between each other and with their environment, and
capable of adapting their internal structure in reaction to these interactions. Sustainability
is then a measure to maintain the organization and the structure of a system with multiple
pathways of evolution [43].

This complements the described AI ethics of desirability which posits that there are a
multitude of possible trajectories for the concretization of AI technology. Hence, a complex
system analysis allows for the modelling of these trajectories (see Figure 1).

Applying this complex system framework to our current case, we can conceptualize
AI—be that singular AI-systems or broader AI infrastructures—as elements of our eco–
socio–economic system. We can then ask how and under which background conditions
AI development and use maintains or disrupts the organization and structure of different
social, economic, and eco-systems. Under this analysis, the separate but interconnected
Three-Pillars of sustainability suddenly quite naturally converge and aspects of sustainabil-
ity that, on former analyses, had to be studied separately despite their interconnections,
can be viewed holistically. All aspects of sustainability, in AI or elsewhere, work towards
the maintenance of a specific state of a complex system, just as in Feenberg’s house, all
design elements work towards energy-efficiency. Thus, the concern that the sustainability
notion encourages further checklist approaches to AI ethics, instead of unifying the space
under one conceptual umbrella, is at least conceptually averted. Moreover, the system
perspective considers the hierarchical organization of systems and sub-systems. In the
context of AI ethics, it, thus, encourages theorists to consider the broader context of the
technology, its (dynamic) development and use, and discourses a fragmented hyper-focus
on particular instances. Nevertheless, it does not exclude this level of analysis either. Rather,
analyses on a lower level, at least potentially, cumulate towards analyses on a higher level,
thus integrating every level of analysis into a grand whole. We can then speak about
sustainability on a local, organizational, national, and global scale [44].
Applying this complex system framework to our current case, we can conceptualize AI—be that singular AI-systems or broader AI infrastructures—as elements of our eco–socio–economic system. We can then ask how and under which background conditions AI development and use maintains or disrupts the organization and structure of different social, economic, and eco-systems. Under this analysis, the separate but interconnected Three-Pillars of sustainability suddenly quite naturally converge and aspects of sustainability that, on former analyses, had to be studied separately despite their interconnections, can be viewed holistically. All aspects of sustainability, in AI or elsewhere, work towards the maintenance of a complex system, just as in Feenberg’s house, all design elements work towards energy-efficiency. Thus, the concern that the sustainability notion encourages further checklist approaches to AI ethics, instead of unifying the space under one conceptual umbrella, is at least conceptually averted. Moreover, the system perspective considers the hierarchical organization of systems and sub-systems. In the context of AI ethics, it, thus, encourages theorists to consider the broader context of the technology, its (dynamic) development and use, and discourages a fragmented hyper-focus on particular instances. Nevertheless, it does not exclude this level of analysis either. Rather, analyses on a lower level, at least potentially, cumulate towards analyses on a higher level, thus integrating every level of analysis into a grand whole. We can then speak about sustainability on a local, organizational, national, and global scale [44].

Speaking of ‘sustainable AI-systems’, however, turns out uninteresting at best and a misnomer at worst under this interpretation. If AI-systems are objects, not systems in the relevant sense, they cannot be sustainable according to this definition. In this case, it might indeed be best to resort to the notions of responsibility or trustworthiness, as these are notions that pertain to individuals. If AI-systems are systems in the relevant sense, them being sustainable would simply amount to them being capable of maintaining their internal organization and structure. This seems neither interesting nor particularly desirable. What is of interest is the maintenance of systems whose conservation is deemed worthwhile, as well as AI-systems’ contribution to that upkeep. It is thus apparent that the notion of sustainability just sketched is merely descriptive and must be complemented by a normative notion: If ‘sustainability’ is to hold promise for AI ethics, we must be able to derive normative claims from it. In this context, a suitable sustainability notion tells us which systems are worthy of maintenance. Figure 1 summarizes our conceptual framework.

Although we are not able to determine a robust normative framework within the scope of this paper, current discussions on the normative foundations of sustainability at least quite readily translate to the system-perspective. As an example, consider the divide between ‘weak’ and ‘strong’ sustainability proponents. ‘Weak sustainability’ is generally understood as the sustainability of growth-oriented economic systems. It prescribes that economic growth, i.e., the compensation of consumed resources, shall be maintained over time. This growth is indifferent towards the origin of these resources, be they artificially produced, human, or natural [45]. In other words, as long as artificial or human resources compensate for the loss of natural resources, ecological degradation is to be viewed as normatively indifferent [26]. The focus here, thus, lies on the maintenance of the current economic system. The implicit assumption here seems to be that sustained economic growth (or at least overall growth) leads to welfare maximization for humans [46].
a sense then, this can also be seen as an appeal towards social sustainability under the assumption that economic sustainability is a prerequisite for this. ‘Strong sustainability’, by contrast, posits some attributes of nature cannot be replaced by artificial capital [26]. In other words, the integrity of the ecosystem, or at least of specific parts of it, is worthy of maintenance.

7. Conclusion and Further Research Recommendations

AI policy guidelines as they are currently devised tend towards disconnected principles, fragmentation, and isolated ethical assessments. We have argued that a more holistic approach is needed. Although a focus on ‘Sustainable AI’ holds a lot of promise in this regard, we have found that a first conceptualization of the notion in the policy space does not realize its potential. An alternative is offered by an ethics of desirability for AI. In this paper, we point towards a conceptualization of ‘sustainability’ as a property of complex systems that paves the way towards desirable AI ethics. Suitable AI ethics guidelines tell us how AI developers and users can work towards the maintenance of those systems deemed worthwhile, instead of focusing on how AI can be made less destructive. Much more research is necessary before this approach can formulate such guidelines. First, social, economic, and eco-systems must be identified and modelled. What adequate modelling looks like depends on both the level of analysis and the systems and sub-systems deemed relevant. Second, an ethics assessment needs to determine which systems’ functioning is worthy of protection. For this to be possible, a meta-ethical framework needs to be developed which explains how such assessments can be made. Evidently, both fields of research inform each other. Which systems and sub-systems are relevant is to be determined by an ethics assessment while the ethics assessment depends on system analysis outcomes with regard to possible states of a system and system-co-tenabilities. In any case, even if no system approach is adopted, we urge the authors of the report and future ‘Sustainable AI’ theorists to avoid conceptual scattering by clarifying what end, value, or duty they believe justifies the normative force of ‘sustainability’.

Author Contributions: Conceptualization, L.B., T.V. and A.v.W.; writing—original draft preparation, L.B.; writing—review and editing, T.V. and A.v.W.; supervision T.V. and A.v.W. All authors have read and agreed to the published version of the manuscript.

Funding: Funding for this research was provided by the Alexander von Humboldt Foundation in the framework of the Alexander von Humboldt Professorship for Artificial Intelligence endowed by the Federal Ministry and Research to Prof. Dr. Aimee van Wynsberghe.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

[A1] Prozess, “der sich um die Frage nach der gerechten Verteilung zwischen den heute lebenden Menschen und zukünftigen Generationen dreht und um den gerechten Umgang der Menschen miteinander sowie mit der Natur.”

[A2] “Es wird deutlich, dass ökologische und soziale Nachhaltigkeit im Grunde zwei Seiten einer Medaille sind.”

[A3] “in Einklang bringen”.

[A4] “Eine nachhaltige KI ist aus unserer Perspektive vorhanden, wenn Entwicklung und Einsatz dieser Systeme die planetaren Grenzen respektiert, keine problematischen ökonomischen Dynamiken verstärkt und den gesellschaftlichen Zusammenhalt nicht gefährdet.”
29. Van Wynsberghe, A. Designing robots for care: Care centered value-sensitive design. Sci. Eng. Ethics 2013, 19, 407–433. [CrossRef] [PubMed]

30. Brey, P. Values in technology and disclosive computer ethics. Camb. Handb. Inf. Comput. Ethics 2010, 4, 41–58.

31. Borning, A.; Muller, M. Next steps for value sensitive design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, TX, USA, 5–10 May 2012; ACM: New York NY, USA, 2012; pp. 1125–1134. [CrossRef]

32. Vandemeulebroecke, T.; Cavolo, A.; Gastmans, C. ‘Yes we hear you. Do you hear us?’. A sociopolitical approach to video-based telepsychiatric consultations. J. Med. Ethics 2022, 48, 34–35. [CrossRef] [PubMed]

33. ten Have, H. Ethical perspectives on health technology assessment. Int. J. Technol. Assess. Health Care 2004, 20, 71–76. [CrossRef] [PubMed]

34. Coeckelbergh, M. Green Leviathan or the Poetics of Political Liberty. Navigating Freedom in the Age of Climate Change and Artificial Intelligence; Routledge: New York, NY, USA; London, UK, 2021.

35. ten Have, H. Ethical perspectives on health technology assessment. Int. J. Technol. Assess. Health Care 2004, 20, 71–76. [CrossRef] [PubMed]

36. Coeckelbergh, M. Artificial agents, good care, and modernity. Med. Bioeth. 2015, 36, 265–277. [CrossRef]

37. AI Ethics Impact Group. From Principles to Practice. An Interdisciplinary Framework to Operationalise AI Ethics; Bertelsmann Stiftung: Gütersloh, Germany, 2020.

38. UNESCO. Recommendation on the Ethics of Artificial Intelligence. Online Publication. 2021. Document Code: Document Code: Code: SHS/BIO/REC-AIETHICS/2021. Available online: https://unesdoc.unesco.org/ark:/48223/pf0000380455 (accessed on 29 March 2022).

39. Floridi, L. Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. Philos. Trans. A Math. Phys. Eng. Sci. 2016, 374, 20160112. [CrossRef] [PubMed]

40. Feenberg, A. Questioning Technology; Routledge: New York, NY, USA; London, UK, 1999.

41. Feenberg, A. Lukács’s theory of reification and contemporary social movements. Rethink. Marx. 2015, 27, 490–507. [CrossRef]

42. Jørgensen, S.E.; Patten, B.C.; Straškraba, M. Ecosystems emerging: Toward an ecology of complex systems in a complex future. Ecol Model. 1992, 62, 1–28. [CrossRef]

43. Crojethovich-Martín, A.D.; Perazzo-Rescia, A.J. Organización y sostenibilidad en un sistema urbano socio-ecológico y complejo. Rev. Int. Sostenibilidad Tecnol. Y Humanismo 2006, 1, 103–121.

44. Feenberg, A. Technosystem. The Social Life of Reason; Harvard University Press: Cambridge, MA, USA; London, UK, 2017.

45. Pearce, D.W.; Atkinson, G.D. Are National Economies Sustainable? Measuring Sustainable Development; CSERGE Working Paper GEC 92-11; Centre for Social and Economic Research on the Global Environment: London, UK, 1992.

46. Beckermann, W. ‘Sustainable Development’: Is It a Useful Concept? Environ. Ethics 1994, 3, 191–209. [CrossRef]