sustainability

Towards the Sustainability of AI

Multi-Disciplinary Approaches to Investigate the Hidden Costs of AI

Edited by
Aimee van Wynsberghe, Tijs Vandemeulebroucke, Larissa Bolte and Jamila Nachid

Printed Edition of the Special Issue Published in *Sustainability*

www.mdpi.com/journal/sustainability

MDPI

Towards the Sustainability of AI; Multi-Disciplinary Approaches to Investigate the Hidden Costs of AI

Towards the Sustainability of AI; Multi-Disciplinary Approaches to Investigate the Hidden Costs of AI

Editors

Aimee van Wynsberghe
Tijs Vandemeulebroucke
Larissa Bolte
Jamila Nachid

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin

Editors

Aimee van Wynsberghe
Institute for Science and Ethics
University of Bonn
Germany

Tijs Vandemeulebroucke
Institute for Science and Ethics
University of Bonn
Germany

Larissa Bolte
Institute for Science and Ethics
University of Bonn
Germany

Jamila Nachid
Institute for Science and Ethics
University of Bonn
Germany

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Sustainability* (ISSN 2071-1050) (available at: https://www.mdpi.com/journal/sustainability/special_issues/sustainability_AI).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-0365-6600-9 (Hbk)
ISBN 978-3-0365-6601-6 (PDF)

Contents

About the Editors

Aimee van Wynsberghe

Aimee van Wynsberghe is the Alexander von Humboldt Professor for Applied Ethics of Artificial Intelligence at the University of Bonn in Germany. Aimee is director of the Institute for Science and Ethics and the Bonn Sustainable AI Lab. She is co-director of the Foundation for Responsible Robotics and a member of the European Commission's High-Level Expert Group on AI. She is a founding editor of the international peer-reviewed journal *AI & Ethics* and member of the World Economic Forum's Global Futures Council on Artificial Intelligence and Humanity. She is author of the book Healthcare Robots: Ethics, Design, and Implementation and is regularly interviewed by media outlets. In each of her roles, Aimee works to uncover the ethical risks associated with emerging robotics and AI. Aimee's current research, funded by the Alexander von Humboldt Foundation, brings attention to the sustainability of AI by studying the hidden environmental costs of developing and using AI.

Tijs Vandemeulebroucke

Tijs Vandemeulebroucke is a postdoctoral researcher at the Bonn Sustainable AI Lab of the Institute for Science and Ethics of the University of Bonn, Germany. He holds a Ph.D. in Biomedical Sciences (KU Leuven, 2019) and MA degrees in theology and religious studies (KU Leuven, 2013) and philosophy (KU Leuven, 2015). He researches the ethical tension between the use of AI in healthcare settings in countries in the global north and the environmental, health, and social impact of the development and recycling of AI on local communities across the world. His research relies on philosophical–ethical approaches such as care ethics, bioethics, global bioethics, critical theory of technology, phenomenology, and deconstruction and empirical–ethical approaches inspired by grounded theory. Tijs won the 2020 Doctoral Dissertation Award on Artificial Intelligence & Ethics jointly given by the Microsoft Corporation and the Pontifical Academy for Life. His work is published in journals as *Science & Engineering Ethics, American Journal of Bioethics, Journal of Medical Ethics*.

Larissa Bolte

Larissa Bolte is a Ph.D. student at the Bonn Sustainable AI Lab of the Institute for Science and Ethics at the University of Bonn, Germany, under the supervision of Prof. Dr. Aimee van Wynsberghe. She holds a MA degree in philosophy (University of Bonn, 2021) and a BA degree in philosophy and psychology (University of Bonn, 2019). She spent part of her BA studies at Université Paris 1 Panthéon-Sorbonne, France. For her Ph.D. thesis, Larissa investigates the conceptual foundations of sustainability, considering both the notion's central characteristics and its normative implications. Within the Bonn Sustainable AI Lab, she will examine how sustainability, construed as a theoretical lens, can inform AI ethics. Larissa has previously worked as a research assistant at the Institute for Science and Ethics and at the German Reference Centre for Ethics in the Life Sciences. She has also been a tutor for both moral philosophy and logic and basic research at the philosophy department of the University of Bonn. Larissa was a co-organiser of the Sustainable AI Conference 2021, which is associated with this Special Issue, and has published in the journal *Sustainability*.

Jamila Nachid

Jamila Nachid is currently a master's student of Applied Ethics at the University of Utrecht in the Netherlands. She holds a BA degree in philosophy, German studies, and educational sciences (University of Bonn, Germany, 2022), having also studied at the University of St Andrews in Scotland. Focussing on topics such as bioethics, AI, and sustainability, Jamila has worked as a research student at the Bonn Sustainable AI Lab, the Institute for Science and Ethics (IWE), and the German Reference Centre for Ethics in the Life Sciences (DRZE). She has also been a tutor for ethics at the philosophy department of the University of Bonn. For two consecutive years, she was granted the Deutschlandstipendium, a merit-based scholarship. Jamila was a co-organiser of the Sustainable AI Conference 2021 that is associated with this Special Issue.

sustainability

MDPI

Editorial

Special Issue "Towards the Sustainability of AI; Multi-Disciplinary Approaches to Investigate the Hidden Costs of AI"

Aimee van Wynsberghe, Tijs Vandemeulebroucke *, Larissa Bolte and Jamila Nachid

The Sustainable AI Lab, Institute for Science and Ethics, University of Bonn, Bonner Talweg 57, 53113 Bonn, Germany

* Correspondence: tvandeme@uni-bonn.de

Artificial Intelligence (AI) applications, i.e., applications of machine learning, deep learning, and other related technologies, are increasing at a rapid pace in our personal and professional lives. AI is used, for example, to predict fraud in the banking sector, benefitting both customer and company; as a decision support tool in healthcare, enhancing the efficiency of healthcare institutions; or to predict and mitigate natural disasters, protecting individuals in the surrounding area. The responsible use of AI may be of great benefit for humanity, from monitoring and repairing climate change destruction to uncovering new forms of disease and their respective treatments. Yet, despite the success and efficiency that AI promises, there are growing societal and ethical concerns that need to be addressed to prevent the design, development, and use of AI from creating new and exacerbating existing social and environmental injustices.

To date, the field of AI ethics has focused on uncovering and raising awareness of a host of issues in relation to AI including the potential loss of jobs, the quality of jobs available, privacy concerns about data collection, the embedding of cultural stereotypes and prejudices into AI models when using historical data to train these models, and the lack of transparency of decision rules generated by AI models, to name a few. While all of these are pressing issues, the issue of the sustainability of AI remains underexplored. Very recently, researchers have begun to uncover the environmental risks related to the materiality of AI [1–4]. These first publications focus strongly on energy consumption and on greenhouse gas emissions produced by training, tuning, and using AI systems. However, the environmental impact of AI does not stop there; the hardware used to run algorithms requires an industrial infrastructure of mining of natural resources (e.g., gold, tungsten), assembly of technical elements, cooling of technical infrastructures, and (electronic) waste disposal. Along this line of production, maintenance, and obsolescence, environmental risks arise that remain obscured and unquantified. To exacerbate these issues, AI is not only itself materially instantiated, it also changes the material conditions of the context in which it is employed, bringing efficiency gains, but potentially also creating rebound and ripple effects [5]. If AI is truly to succeed in making our world a better, more sustainable place, its full environmental and social impact must be uncovered. This is the objective of research on *Sustainable AI*.

Sustainable AI is a term that has recently garnered more and more attention, yet the understanding of what it means is still in development. In recent years, initiatives and articles predominantly addressed AI for sustainability, with few conferences [6,7], articles [2,8], and books [9,10] being dedicated to the sustainability of AI. It is this observation that has inspired the founding of the *Sustainable AI Lab* at the University of Bonn, Germany. Research towards the sustainability of AI is aimed at shining a light on the currently hidden sustainability issues and impacts of AI and, based on this knowledge, providing an ethically sound way forward for academics, policy makers, and industry alike. This requires an exploration of the concept of sustainability itself, its relation to AI, and a clear understanding of where in the AI design, development, implementation, and use

Citation: van Wynsberghe, A.; Vandemeulebroucke, T.; Bolte, L.; Nachid, J. Special Issue "Towards the Sustainability of AI; Multi-Disciplinary Approaches to Investigate the Hidden Costs of AI". *Sustainability* **2022**, *14*, 16352. https://doi.org/10.3390/su142416352

Received: 10 November 2022
Accepted: 1 December 2022
Published: 7 December 2022

process ethical issues can and should be addressed. Thus, *Sustainable AI* has an enormous task ahead to both raise awareness of possible harms as well as assist in their prevention.

A big step towards this goal is to re-assess our collective conception of AI. Generally, the term 'AI' denotes any and all computing technologies that emulate human cognitive abilities. Not only is this general definition ambiguous towards the terms 'emulate' and 'human cognitive abilities', it also invokes connotations of the digital, the virtual, the mental, and the immaterial [11]. If we are to think about AI in the context of sustainability, we ought to broaden our conception of what AI is and how it is embedded in the material world as a material object. The co-editors of this Special Issue, Aimee van Wynsberghe, Tijs Vandemeulebroucke, Larissa Bolte, and Jamila Nachid, hence urge readers to conceptualise AI as a *world object* [12]. Indeed, AI "[...] affects the world as a whole and not just a small corner of it" [13] (p. 5). This encompasses, first, the realisation that software cannot and should not be divorced from hardware. Every AI software requires hardware and a surrounding material infrastructure to run [10]. Second, this means realising that hardware does not exist as an isolated entity. Its existence is always dependent on complex global networks of production, supply, and use [8,9]. In the specific case of AI, these global networks are: socio-environmental, connecting labour force and raw materials to produce and distribute the technical components of the hardware running AI; material, providing the technical products and socio-technical relations that together produce AI; and digital, providing computational analysis.

We do not claim here to have exhaustively explored the materiality of AI technologies or to have given a definition that suits all contexts. Instead, we intend our conceptualisation of AI to act as an invitation to shift perspective and to facilitate discussion on the sustainability of AI.

This discussion is a multi-faceted one. For this reason, this collection of papers explores the issue of Sustainable AI from a variety of different angles. The first set of papers deals with diverse, fundamental ethical questions in relation to the sustainability of AI. In the vein of reconceptualising AI, Bolte, Vandemeulebroucke, and van Wynsberghe [14] argue that problems with current AI ethics guidelines are due to a conceptualisation of AI as isolated artefacts, which can be revised by conceptualising sustainability as a property of complex systems. Halsband [15] can be read as an elaboration on this topic, arguing for intergenerational justice as the normative core of the sustainability concept, while Robbins and van Wynsberghe [16] point out the consequences of infrastructural lock-in if the interconnectedness of AI with ecological, social, and economic systems is disregarded. However, AI can also help us achieve our visions for a more sustainable future, as shown by Bartman [17], who argues for the legitimacy of certain forms of AI-powered climate nudging.

The second set of papers focusses on sustainability frameworks for AI in diverse contexts. Continuing the theme of ethical groundwork, Genovesi and Mönig [18] connect sustainability to an ethics of responsibility and investigate how sustainability can be included in an ethical AI certification. Two papers propose frameworks based on the UN Sustainable Development Goals (SDGs): Gupta and Rhyner [19] strive to connect digitalisation with the SDGs, while Sætra [20] develops a framework for more comprehensive corporate reporting on the sustainability impact of AI. To make corporate impacts more visible to consumers, Kindylidi and Cabral [21] assess whether the current EU consumer protection framework is sufficient to promote sharing and substantiation of sustainability information to consumers.

Two of the papers in this collection present methodologies to relevant stakeholders. Ligozat et al. [22] point out the lack of attention on the negative sustainability impacts of AI for Green and introduce Life Cycle Assessment as a useful methodology for anyone developing AI for Green solutions. Raper et al. [23], with SDG 13 "Climate Action" in mind, present the notion of sustainability budgets addressed at software developers.

Getting clearer on how the relation between sustainability and AI is construed in the field, Samuel, Lucivero, and Somavilla [24] survey how stakeholders researching,

governing, or working on the environmental impacts of digital technologies utilise different conceptions of 'environmental sustainability'.

Finally, we close our collection with Bliek [25], who reviews successful sustainability applications of machine learning that use the technique of surrogate-based optimisation and gives recommendations to researchers who work on or apply that technique.

It is evident from this collection that Sustainable AI can only be tackled in an interdisciplinary manner. The question of 'What is sustainable AI?' must be approached from conceptual, ethical, political, sociological, empirical, technical, and many more perspectives. Ultimately, in asking this question, we ask how we envision our future with AI and how this vision may be blurred by leaving its material impact out of sight. We ask this question at a crucial time where climate and environmental worries accelerate while enthusiasm for AI runs high. The co-editors hence urge researchers to come together and work on the interrelations between these two grand developments of our time to find a sustainable path forward. With this collection of papers, we take a necessary step in this direction.

Author Contributions: Conceptualization, A.v.W., T.V., L.B., and J.N. writing—original draft preparation, T.V., and L.B.; writing—review and editing, A.v.W., T.V., L.B., and J.N.; supervision, A.v.W.; funding acquisition, 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 The Applied Ethics of Artificial Intelligence endowed by the Federal Ministry of Education and Research to Prof. Dr. Aimee van Wynsberghe.

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

References

1. Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 28 July–2 August 2019; Available online: https://arxiv.org/pdf/1906.02243.pdf (accessed on 14 October 2022).
2. Coeckelbergh, M. AI for Climate: Freedom, Justice, and Other Ethical and Political Challenges. *AI Ethics* **2021**, *1*, 61–72. [CrossRef]
3. Nordgren, A. Artificial Intelligence and Climate Change: Ethical Issues. *J. Inf. Commun. Ethics Soc.* **2022**. *ahead-of-print*. [CrossRef]
4. Dodge, J.; Prewitt, T.; des Combes, R.T.; Odmark, E.; Schwartz, R.; Strubell, E.; Luccioni, A.S.; Smith, N.A.; DeCario, N.; Buchanan, W. Measuring the Carbon Intensity of AI in Cloud Instances. In Proceedings of the FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; pp. 1877–1894. [CrossRef]
5. Hilty, L.M.; Aebischer, B. ICT for Sustainability: An Emerging Research Field. In *ICT Innovations for Sustainability*; Hilty, L.M., Aebischer, B., Eds.; Springer: Cham, Switzerland, 2015.
6. Sustainable AI Conference. 2021. Available online: https://sustainable-ai-conference.eu/archive/conference-2021 (accessed on 14 October 2022).
7. 1st International Sustainable AI Workshop (ISAW). 2022. Available online: https://sustainai.github.io/icdm2022_workshop/ (accessed on 14 October 2022).
8. Van Wynsberghe, A. Sustainable AI: AI for Sustainability and The Sustainability of AI. *AI Ethics* **2021**, *1*, 213–218. [CrossRef]
9. Crawford, K. *Atlas of AI: Power, Politics, and The Planetary Costs of Artificial Intelligence*; Yale University Press: New Haven, CT, USA, 2021.
10. Coeckelbergh, M. *AI Ethics*; The MIT Press: Cambridge, MA, USA, 2020.
11. Boden, M. *AI Its Nature and Future*; Oxford University Press: Oxford, UK, 2016.
12. Serres, M. *Le Contrat Naturel*; Francois Bourin: Paris, France, 1994.
13. Feenberg, A. *Technosystem: The Social Life of Reason*; Harvard University Press: Cambridge, MA, USA; London, UK, 2017.
14. Bolte, L.; Vandemeulebroucke, T.; van Wynsberghe, A. From an Ethics of Carefulness to an Ethics of Desirability: Going Beyond Current Ethics Approaches to Sustainable AI. *Sustainability* **2022**, *14*, 4472. [CrossRef]
15. Halsband, A. Sustainable AI and Intergenerational Justice. *Sustainability* **2022**, *14*, 3922. [CrossRef]
16. Robbins, S.; van Wynsberghe, A. Our New Artificial Intelligence Infrastructure: Becoming Locked into an Unsustainable Future. *Sustainability* **2022**, *14*, 4829. [CrossRef]
17. Bartmann, M. The Ethics of AI-Powered Climate Nudging—How Much AI Should We Use to Save the Planet? *Sustainability* **2022**, *14*, 5153. [CrossRef]
18. Genovesi, S.; Mönig, J.M. Acknowledging Sustainability in the Framework of Ethical Certification for AI. *Sustainability* **2022**, *14*, 4157. [CrossRef]
19. Gupta, S.; Rhyner, J. Mindful Application of Digitalization for Sustainable Development: The Digitainability Assessment Framework. *Sustainability* **2022**, *14*, 3114. [CrossRef]

20. Sætra, H.S. A Framework for Evaluating and Disclosing the ESG Related Impacts of AI with the SDGs. *Sustainability* **2021**, *13*, 8503. [CrossRef]
21. Kindylidi, I.; Cabral, T.S. Sustainability of AI: The Case of Provision of Information to Consumers. *Sustainability* **2021**, *13*, 12064. [CrossRef]
22. Ligozat, A.-L.; Lefevre, J.; Bugeau, A.; Combaz, J. Unraveling the Hidden Environmental Impacts of AI Solutions for Environment Life Cycle Assessment of AI Solutions. *Sustainability* **2022**, *14*, 5172. [CrossRef]
23. Raper, R.; Boeddinghaus, J.; Coeckelbergh, M.; Gross, W.; Campigotto, P.; Lincoln, C.N. Sustainability Budgets: A Practical Management and Governance Method for Achieving Goal 13 of the Sustainable Development Goals for AI Development. *Sustainability* **2022**, *14*, 4019. [CrossRef]
24. Samuel, G.; Lucivero, F.; Somavilla, L. The Environmental Sustainability of Digital Technologies: Stakeholder Practices and Perspectives. *Sustainability* **2022**, *14*, 3791. [CrossRef]
25. Bliek, L. A Survey on Sustainable Surrogate-Based Optimisation. *Sustainability* **2022**, *14*, 3867. [CrossRef]

sustainability

MDPI

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

Citation: Bolte, L.; Vandemeulebroucke, T.; van Wynsberghe, A. From an Ethics of Carefulness to an Ethics of Desirability: Going Beyond Current Ethics Approaches to Sustainable AI. *Sustainability* **2022**, *14*, 4472. https://doi.org/10.3390/su14084472

Academic Editor: Peng-Yeng Yin

Received: 28 February 2022
Accepted: 4 April 2022
Published: 8 April 2022

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 count instances of impacts, but 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 fulfil 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	Nicht-Diskriminierung und Fairness
	Technical Reliability and Human Oversight	Technische Verlässlichkeit und Menschliche Aufsicht
	Autonomy and Data Protection	Selbstbestimmung und Datenschutz
	Inclusive and Participatory Design	Inklusives und Partizipatives Design
	Cultural Sensibility	Kulturelle Sensibilität
Economic Sustainability	Market Diversity and Exhaustion of Innovative Potential	Marktvielfalt und Ausschöpfung des Innovationspotenzials
	Distribution Effect in Target Markets	Verteilungswirkung in Zielmärkten
	Working Conditions and Jobs	Arbeitsbedingungen und Arbeitsplätze
Ecological Sustainability	Energy Consumption	Energieverbrauch
	CO_2 and Greenhouse Gas Emissions	CO_2- und Treibhausgasemissionen
	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 carefulness', 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 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 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 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].

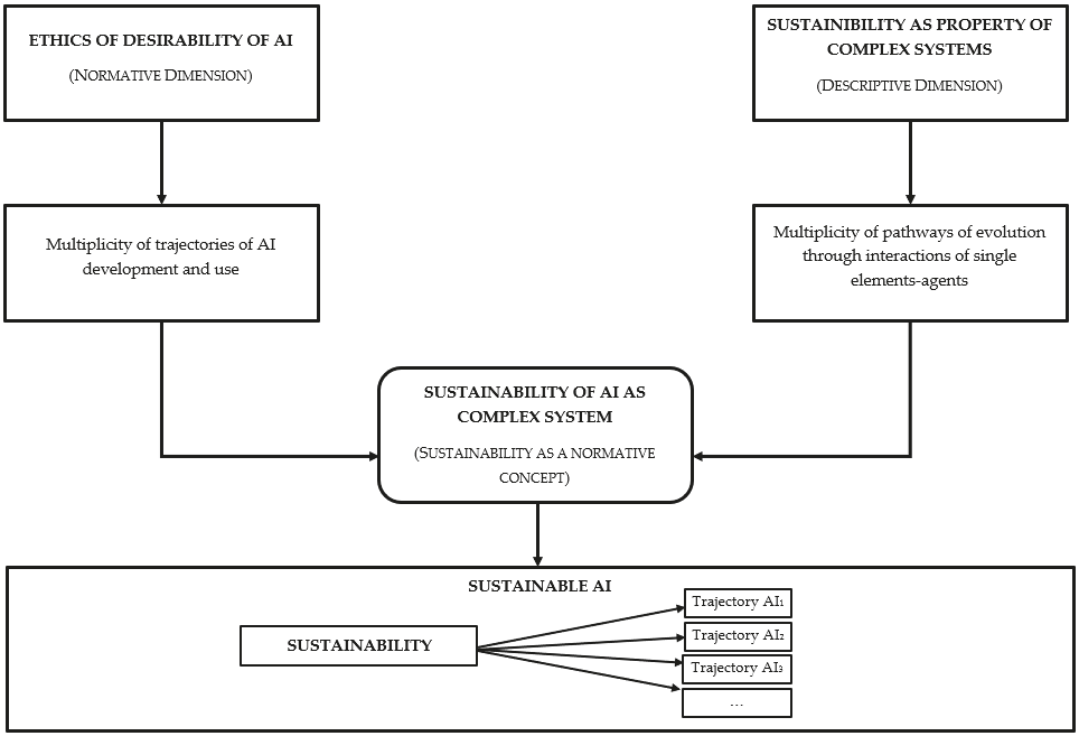


Figure 1. Sustainability of AI as complex system.

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]. In

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."

References

1. Zhou, N.; Zhang, Z.; Nair, V.N.; Singhal, H.; Chen, J.; Sudjianto, A. Bias, Fairness, and Accountability with AI and ML Algorithms. *arXiv* **2021**. Available online: https://arxiv.org/abs/2105.06558 (accessed on 29 March 2022).
2. AlgorithmWatch. AI Ethics Guidelines Global Inventory. Available online: https://inventory.algorithmwatch.org/ (accessed on 29 March 2022).
3. Strubell, E.; Ganesh, A.; McCallum, A. Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, July 2019; Available online: https://arxiv.org/pdf/1906.02243.pdf (accessed on 29 March 2022).
4. Schwartz, R.; Dodge, J.; Smith, N.A.; Etzioni, O. Green AI. *Comm. ACM* **2020**, *63*, 54–63. [CrossRef]
5. Bolger, M.; Marin, D.; Tofighi-Niaki, A.; Seelman, L. *'Green Mining' is A Myth: The Case for Cutting EU Resource Consumption*; European Environmental Bureau & Friends of the Earth Europe: Brussels, 2021; 51p. Available online: https://eeb.org/wp-content/uploads/2021/10/Green-mining-report_EEB-FoEE-2021.pdf (accessed on 26 February 2022).
6. Schomberg, A.C.; Bringezu, S.; Flörke, M. Extended life cycle assessment reveals the spatially-explicit water scarcity footprint of a lithium ion battery storage. *Comm. Earth Environ.* **2021**, *2*, 1–10. [CrossRef]
7. Andrews, D.; Newton, E.; Naeem, A.; Chenadex, J.; Bienge, K. A circular economy for the data centre industry: Using design methods to address the challenge of whole system sustainability in a unique industrial sector. *Sustainability* **2021**, *13*, 6319. [CrossRef]
8. Navas, G.; D'Alisa, G.; Martínez-Alier, J. The role of working-class communities and the slow violence of toxic pollution in environmental health conflicts: A global perspective. *Glob Environ. Change* **2022**, *73*, 102474. [CrossRef]
9. Van Wynsberghe, A. Sustainable AI: AI for sustainability and the sustainability of AI. *AI Ethics* **2021**, *1*, 2013–2218. [CrossRef]
10. Rohde, F.; Wagner, J.; Reinhard, P.; Petschow, U.; Meyer, A.; Voß, M.; Mollen, A. *Entwicklung eines Kriterien- und Indikatorensets für die Nachhaltigkeitsbewertung von KI-Systemen entlang des Lebenszyklus*; Report No.: 220/21; Schriftenreihe des IÖW: Berlin, Germany, 2022; 80p.
11. Cave, S.; ÓhÉigeartaigh, S. An AI race for strategic advantage: Rhetoric and risks. In Proceedings of the AIES '18: Proceedings of the 2018 AAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; ACM: New York NY, USA, 2018; pp. 36–40. [CrossRef]
12. Fjeld, J.; Achten, N.; Hilligoss, H.; Nagy, A.; Srikumar, M. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. Berkman Klein Center for Internet & Society. 2020. Available online: https://dash.harvard.edu/bitstream/handle/1/42160420/HLS%20White%20Paper%20Final_v3.pdf?sequence=1&isAllowed=y (accessed on 27 March 2022).
13. Mittelstadt, B. Principles alone cannot guarantee ethical AI. *Nat. Mach. Intell.* **2019**, *1*, 501–507. [CrossRef]
14. Beauchamp, T.L.; Childress, J.F. *Principles of Niomedical Ethics*, 8th ed.; Oxford University Press: New York, NY, USA, 2019.
15. Jobin, A.; Ienca, M.; Vayena, E. Artificial intelligence: The global landscape of ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
16. European Group on Ethics in Science and New Technologies. *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*; Publications Office of the European Union: Luxembourg, 2018.
17. Green Digital Working Group. Position on Robotics and Artificial Intelligence. 2016. Available online: https://felixreda.eu/wp-content/uploads/2017/02/Green-Digital-Working-Group-Position-on-Robotics-and-Artificial-Intelligence-2016-11-22.pdf (accessed on 29 March 2022).
18. Vinuesa, R.; Azizpour, H.; Balaam, M.; Dignul, V.; Domisch, S.; Felländer, A.; Langhans, S.D.; Tegmark, M.; Nerini, F.F. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nat. Comm.* **2020**, *11*, 233. [CrossRef] [PubMed]
19. Sætra, H.S. AI in context and the sustainable development goals: Factoring in the unsustainability of the sociotechnical system. *Sustainability* **2021**, *13*, 1738. [CrossRef]
20. Hagendorff, T. The ethics of AI ethics: An evaluation of guidelines. *Minds Mach.* **2020**, *30*, 99–120. [CrossRef]
21. Barley, S.R. *Work and Technological Change*; Oxford University Press: Oxford, UK, 2020.
22. Boddington, P. *Towards a Code of Ethics for Artificial Intelligence*; Springer: Cham, Germany, 2017.
23. Sætra, H.S. A typology of AI applications in politics. In *Artificial Intelligence and Its Contexts*; Visvizi, A., Bodziany, M., Eds.; Springer: Cham, Germany, 2021; pp. 27–43. [CrossRef]
24. Bowie, N. Organisational integrity and moral climates. In *Oxford Handbook of Business Ethics*; Brenkert, G.G., Ed.; Oxford Handbooks Online; Oxford University Press: Oxford, UK, 2009.
25. Marvin, L.J.H. The problem of 'thick in status, thin in content' in Beauchamp and Childress' principlism. *J. Med. Ethics* **2010**, *36*, 525–528. [CrossRef]
26. Ruggerio, C.A. Sustainability and sustainable development: A review of principles and definitions. *Sci. Total Environ.* **2021**, *786*, 147481. [CrossRef]
27. High-Level Expert Group set up by European Commission. *Ethics Guidelines for Trustworthy AI*; European Commission: Brussels, Switzerland, 2019.
28. Poel, I.V. Translating values into design requirements. In *Philosophy and Engineering: Reflections on Practice, Principles and Process*; Springer: Dordrecht, The Netherlands, 2013; pp. 253–266.

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. Vandemeulebroucke, 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. Feenberg, A. Concretizing Simondon and constructivism: A recursive contribution to the theory of concretization. *Sci. Technol. Hum. Values* **2017**, *42*, 62–85. [CrossRef]
35. 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.
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: 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]

sustainability

MDPI

Article

Sustainable AI and Intergenerational Justice

Aurélie Halsband

German Reference Centre for Ethics in the Life Sciences (DRZE), University of Bonn, 53113 Bonn, Germany; ahalsban@uni-bonn.de

Abstract: Recently, attention has been drawn to the sustainability of artificial intelligence (AI) in terms of environmental costs. However, sustainability is not tantamount to the reduction of environmental costs. By shifting the focus to intergenerational justice as one of the constitutive normative pillars of sustainability, the paper identifies a reductionist view on the sustainability of AI and constructively contributes a conceptual extension. It further develops a framework that establishes normative issues of intergenerational justice raised by the uses of AI. The framework reveals how using AI for decision support to policies with long-term impacts can negatively affect future persons. In particular, the analysis demonstrates that uses of AI for decision support to policies of environmental protection or climate mitigation include assumptions about social discounting and future persons' preferences. These assumptions are highly controversial and have a significant influence on the weight assigned to the potentially detrimental impacts of a policy on future persons. Furthermore, these underlying assumptions are seldom transparent within AI. Subsequently, the analysis provides a list of assessment questions that constitutes a guideline for the revision of AI techniques in this regard. In so doing, insights about how AI can be made more sustainable become apparent.

Keywords: artificial intelligence; sustainable AI; intergenerational justice; future generations; policy-making; explainability; transparency

Citation: Halsband, A. Sustainable AI and Intergenerational Justice. *Sustainability* **2022**, *14*, 3922. https://doi.org/10.3390/su14073922

Academic Editor: Fabrizio D'Ascenzo

Received: 22 February 2022
Accepted: 23 March 2022
Published: 26 March 2022

1. Introduction

Within its broad field of application, artificial intelligence (AI) is increasingly framed as a promising tool to enhance sustainable development. The European Commission sees AI as one of the digital technologies that are a "critical enabler for attaining the sustainability goals of the Green deal" i.a. by accelerating and maximizing "the impact of policies to deal with climate change and protect the environment" [1] (p. 9).

Recently, attention has been drawn to the environmental impact of AI itself under the umbrella term of "sustainable AI" [2] (see also [3]), stressing the need to critically assess especially the immense energy consumption of AI. However, sustainability is not tantamount to the reduction of environmental costs. By shifting the focus to intergenerational justice as one of the constitutive normative pillars of sustainability, the paper demonstrates and addresses the threat of a reductionist view on sustainable AI. It identifies the question of whether and, if so, to what extent AI can be sustainable as a major research question necessitating a theoretical underpinning. The ethical analysis contributes to the assessment of AI's long-term impacts on sustainability by revealing major implications of intergenerational justice as the underlying normative component (see [4] (p. 2,4)).

Although "sustainability" is a frequently mentioned standard that institutions and persons commit themselves to, the definition and use of this concept are often inconsistent [5]. While the concept's applicability itself is contested [6], as are different interpretations of its content, there is at least a consensus on its core idea: sustainability is the presupposition of intergenerational equity, implying the obligation to conserve "what matters for future generations" [7] (p. 54) (see also [8], p. 60). It is this shared perspective on obligations towards future persons that I will use as the starting point for my analysis.

That is to say, instead of defending a specific interpretation of sustainability, the goal of my analysis is to focus on intergenerational justice as one of its constitutive normative pillars. In so doing, the encompassing demands that are implied with the objective of creating sustainable AI become apparent: if sustainability is fundamentally about conserving "what matters for future generations" [8] (p. 54), this conservative effort will exceed a mere reduction of environmental costs such as those resulting from high energy consumption. This comprehensive approach to sustainable AI is also reflected in the European Commission's description of the conditions that AI must satisfy in regard to 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" [9] (p. 19).

By addressing the question of whether and, if so, to what extent the development and use of AI can be sustainable from the specific normative angle of intergenerational justice, the analysis contributes to closing two research gaps. Firstly, it depicts the reductionist understanding of sustainability in the context of sustainable AI, which has been focused on the welcome call for emission reductions and carbon footprint assessments of AI [10], yet without reference to the further demands of sustainability. This merely implicit reference to intergenerational justice in spite of its fundamental normative function has also been an issue of criticism [11,12] of the United Nations' understanding of sustainability that underlies the formulation of its 17 Sustainable Development Goals (SDGs) [13]. Secondly, the integration of the concept of intergenerational justice provides an addendum to previous analyses of justice issues raised by AI. Although the principle of justice has frequently been applied to evaluate the different uses of AI, these have been focused on issues of discrimination resulting from biased algorithms or on broader issues of distributive justice, e.g., arising from exclusive access to AI technologies because of diverging financial means (cf. e.g., [14], p. 699). Within the emerging application of AI to climate mitigation, additional issues of justice have been discussed such as using AI to nudge people into climate-friendly behaviour or the question of who within the global community should bear the costs of using AI to enhance climate mitigation [3]. Yet, intergenerational justice opens the view on "novel forms of ethical challenges" raised by the use of AI in the context of climate change mitigation and the broader field of environmental policies [15] (p. 13). While issues of *intra*generational justice raised by AI have been addressed before, the intergenerational justice dimension has received little attention up to now [3] (p. 70) and, to my knowledge, there has been no analysis in the context of AI.

To address this gap, the analysis turns to a specific field of application of AI that can significantly impact future persons. Challenges of intergenerational justice are especially raised by the use of AI in those fields of application in which AI provides decision support to issues with long-term impacts, such as environmental protection policies or climate mitigation policies. Other areas where especially policies can have significant impacts on future generations are, e.g., funding strategies of pension schemes or public debt management [16] (p. 62). This paper focuses on the former field of application. Thus, AI, with its specific feature of self-learning (machine learning, ML), is being employed as a tool for climate policy analysis "[. . .] evaluating the outcomes of past policies and assessing future policy alternatives [. . .]. ML can provide data for policy analysis, help improve existing tools for assessing policy options, and provide new tools for evaluating the effects of policies" [17] (p. 52f). In addition, AI has been applied to other environmental issues such as monitoring the extent of deforestation or simulating the effects of climate change [15,17].

As a first step, this analysis provides a normative framework that helps to explore those applications of AI in the context of climate mitigation and environmental protection that raise issues of intergenerational justice, especially those that may have detrimental impacts on future generations. This shall help to contribute to a conceptually informed understanding of sustainability. In a second step, the analysis provides a list of assessment questions that constitutes the first guideline for the revision of AI techniques in this regard.

Overall, the framework offers insights into how sustainable some uses of AI are with the specific normative focus on issues of intergenerational justice.

Although I will mostly refer to ML applications, I use the broader term of AI throughout the paper. The framework and assessment questions will also provide guidance for identifying those types of AI that raise the depicted issues of intergenerational justice.

2. The Normative Framework

2.1. Two Ethical Dimensions of Sustainability

From an ethical perspective, sustainability can be described as a concept with two central dimensions. My analysis starts from the consensus among the debate on defining sustainability that it is majorly based on the obligation to conserve "what matters" for future persons. Obligations to future generations are, in turn, embedded in theories of intergenerational justice which constitute one of the "key components" [16] (p. 62) and the first ethical dimension of sustainability [18] (p. 897).

Note that most concepts of sustainability limit their search for "what matters for future generations" to (parts of) the environment. This specific focus on natural resources as prerequisites for providing future generations with "what matters" constitutes the second ethical dimension of sustainability. An important debate in this context is the dispute between adherents to the concepts of weak sustainability and strong sustainability, where the latter reject, and the former assume that multiple aspects of the "natural capital" currently required to satisfy basic humans needs can prospectively be substituted by technological or other artificial means [18] (p. 904ff). Both ethical dimensions of sustainability heavily rest on normative considerations debated extensively in theories of intergenerational justice. Why we are obligated towards future persons in the first place and if so, to what extent are major subjects of discussion within these theories. For debates within sustainability about the extent or scope of the obligations towards future persons, i.e., if prerequisites for basic needs satisfaction for future persons or a more encompassing perspective on prerequisites for the realisation of different conceptions of the good life should guide the selection of natural resources that ought to be preserved, refer to the general debate about the most convincing distributive principle of intergenerational justice. An example for the former approach is the well-known definition of sustainable development in the World Commission on Environment and Development's 1987 Brundtland report, in which sustainable development is defined as "development that meets the needs of the present without compromising the ability of future generations to meet their own needs" [19].

For the purpose of my analysis, it is the first ethical dimension of sustainability, i.e., the normative concept of intergenerational justice and its focus on impacts on future persons that will serve as a starting point. This will not constitute an encompassing definition of sustainability. Rather, it is conceptualised as a normative module among both further ethically informed modules (i.a. integrating the second ethical dimension of sustainability) and other evaluative modules informed by the additional constitutive perspectives on sustainability such as those of natural, economic, and social sciences.

2.2. Intergenerational Relations as a Framework

How to evaluate present persons' actions when they may have negative impacts on future persons is usually understood as an issue of intergenerational justice within ethical theory, presupposing that the concept of justice can be applied to those not yet alive. Within theories of intergenerational justice, the term "future generations" refers as a shorthand to those persons who will come into existence after the presently alive persons' lifetime, i.e., a group of persons that will have no possibility to directly interact with those presently alive (see [7], p. 43). The need for an ethical assessment of current persons' actions and their impacts on future persons within a specific theory (of justice) can be justified by the features of the relations between members of different generations. One specific feature of the relation between present persons and future persons is *contingency*, referring to the fact that "future people's existence, number, and specific identity depend (are contingent)

upon currently living people's decisions and actions" [20]. Most importantly, which and if future persons come into existence depends on the present person's decisions if, and when, to reproduce. Another genuine feature of the relation between members of different generations is the lack of *reciprocity*, stressing the impossibility of direct interaction between persons currently alive and those who are not yet alive. This relation is closely connected to the intergenerational *power-asymmetry*, describing the fact that only present persons can exercise actions affecting—either positively or negatively—future persons during their lifetime. Finally, intergenerational relations are characterised by *uncertainty*, especially about the identities and preferences of future persons.

As with every innovation, developing and using AI will affect who, how many, and which persons will come into existence (*contingency*). Therefore, I will not treat 'intergenerational contingency' as a genuine normative challenge in the context of AI. Moral implications of the intergenerational relation of contingency have been prominently discussed within the still ongoing debate about the "non-identity problem" [21] (pp. 351–441). In contrast, the focus on the power-asymmetry, as well as the intergenerational relations of non-reciprocity and uncertainty, will help to explore specific uses of AI that raise issues of intergenerational justice. These relations are being used as a framework to break down the encompassing concern of intergenerational justice as to how different entitlements of different persons living at different times—i.e., different generations—should be specified and weighed when they are in conflict.

More generally speaking, the framework supports a continuous ethical assessment of AI as a set of emerging technologies with the specific focus on potentially detrimental impacts that directly result from the use of these technologies in the present but primarily will affect future persons. It rests on past experiences with detrimental side-effects of emerging technologies such as nuclear energy generation and the issue of radioactive waste or high carbon-emitting industries and climatic changes, which both predominantly will affect future generations.

3. Power-Asymmetry and Intertemporal Discounting

With AI's strong potential in the evaluation of large sets of data, it is successively being used to improve policy addresses to the complex phenomenon of climate change and its interdependent causes. Integrated assessment models (IAMs) play an important role in predicting and evaluating the interaction of socioeconomic and climate-related factors [17] (p. 53). The goal of IAMs is "to project alternative future climates with and without various types of climate change policies in place in order to give policymakers at all levels of government and industry an idea of the stakes involved in deciding whether or not to implement various policies" [22] (p. 116). Due to the complexity of the involved models, as well as the amount of data, AI and especially ML are being applied to various sub-models which, together, form the IAMs [17] (p. 53). AI has thus been used to support policy-making in domains with a multitude of factors and stakeholders interacting, such as policies on sustainable development [23] (pp. 22,27) or agricultural public policy [24].

However, this support of policy-making with the help of AI is also confronted with some of the criticism brought forward against features of these policy models in general. One branch of models that are part of IAMs and have important implications regarding intergenerational justice is cost–benefit analyses of climate policies. These models assess the costs and benefits of climate mitigation across a long period of time, surpassing the lifetime of presently alive persons. They assess how costs and benefits are being distributed between different people (i.e., different generations) across different times. How to weigh costs and benefits between persons living at different times within cost–benefit analysis is usually addressed by the inclusion of a social discount rate. Setting the discount rate high involves assigning a significantly smaller value to benefits that accrue in the distant future. This has important normative implications which can be illustrated regarding carbon emission reduction policies:

> "[. . .] intertemporal equity is extremely important in determining the appropriate rate of implementation of policies designed to reduce carbon emissions [. . .]. Low discount rates generally make rapid implementation of such policies much more urgent than high discount rates because damages are projected to grow steadily over time at a much more rapid rate than mitigation costs" [22] (p. 126f).

Against this background, the practice of discounting within cost–benefit analyses with large time horizons—such as those on climate mitigation policies—are faced with considerable objections. On the practical level, it may lead to an underestimation of potentially severe costs for future persons and underplay the urgency of action required in the present to reduce these costs. This is because mitigation policies in the context of climate change imply costs (of climate mitigation) that predominantly accrue to present persons and their losses in consumption. The benefits, however, are reduced risks of climate change which most importantly benefit future persons [25] (p. 401). Present persons thus face potentially higher burdens and are consequently tempted to include an elevated discount rate to reduce these burdens. On a more general level, if and at which rate to discount refers to a disputed field of normative assumptions. Different justifications for discounting the future have been discussed, for example, that it may be justified to give less weight to benefits for future persons as they will overall be better off under the assumption of an overall steadily increasing wealth [26] (p. 48f). Whether there are legitimate reasons to discount benefits for future persons has been subject to an extensive discussion within philosophy and between philosophers and economists (see e.g., [21,27]). With respect to applying AI to this domain of policy evaluation, it suffices to state in a first step that the integration of a social discount rate in those contexts with large time horizons needs to be accessible for a normative evaluation. Among other considerations, strongly discounting benefits for future persons can bear the risk of assigning excessively high costs to them. This may then equal a negative manifestation of the intergenerational power-asymmetry.

The issue of discounting is, however, not a normative issue genuinely raised by the application of AI. Instead, applying AI to this domain can only be justified if the already discussed limitations of these models are adequately considered. Yet a specific challenge genuine to some of the AI techniques is the issue of providing an explanation for generated decisions. As has been shown, the setting of a social discount rate can have important normative implications regarding future persons. To address these limitations, cost–benefit analyses conducted by AI need to be explainable and transparent regarding the setting of the discount rate, thus leaving the possibility for later revisions of the settings. I will come back to the aspect of explainable AI below. With regard to the limitations of the integrated models, constructive insights for potential revision can be gained from general critical assessments of these models [22] (pp. 124,128f) and from objections to the practice of discounting, e.g., in climate mitigation [25] (pp. 401,405).

Regarding the use of AI to support assessments with large time frames such as climate mitigation policies, another aspect under dispute, which has important implications regarding intergenerational justice, is the underlying calculation of costs. A focus on static costs has been shown to neglect the long-term aspect of climate change by neglecting the dynamics between potentially slightly higher costs in the present that may, however, reduce mitigation costs in the distant and near future [28] (p. 54), thus generating an overall improved cost–benefit ratio. Hence, the calculation of costs represents another aspect that must be accessible for potential revision within assessments that are being conducted or supported by AI.

Finally, policies with long-term impacts will only be able to represent potentially detrimental consequences for future persons if the time frames are set in a way that includes those persons. This illustrates a third aspect that needs to be accessible for potential revision not only within cost–benefit analyses conducted or supported by AI but for all types of policy assessments that may include AI. For example, policy-making regarding energy management relies among others on electricity demand forecasting which is increasingly being supported by AI. Within these forecasts, time horizons for long-term projections range

from a couple of years to projections about the next 50 years [29] (p. 15ff). Consequently, insights about the time frames and thus implicitly about the representation of potential impacts affecting persons in the distant future need to be made accessible within AI-based policy support assessments.

Using AI on contexts and decisions affecting different persons and different times—especially future generations—thus adds to the general challenge of creating AI that is transparent and explainable. Explainability is addressing "the need to understand and hold to account the decision-making processes of AI" [14] (p. 700). The principle of explainability has been established as a genuine principle for the normative evaluation of AI along with the established bioethical principles of beneficence, non-maleficence, autonomy, and justice. Impacts on future persons constitute a yet-underestimated societal area that ought to be assessed using this principle. This will also contribute to the critical assessment of using AI within policy-making that has importantly been focused on issues of acceptance and trust [23] (p. 33f).

4. Uncertain Preferences and "Intergenerational Transfer Bias"

Intergenerational relations are characterised by uncertainty in important domains, such as uncertainty about the preferences of future persons. Consequently, there is no data or only fragmentary data that AI can use in this regard. Using AI for assessments with large time frames will accordingly involve assumptions about preferences that future persons will have and how these can be 'translated' into opportunities that present persons should leave open for them. For example, the implications for the use of IAMs in the context of climate mitigation can be described like this:

> "People making decisions today on behalf of those not yet alive need to make collective ethical choices about what kind of opportunities (usually characterized as a particular state of the climate system measured by global mean temperature, GHG concentration, or maximum climate damages allowable by some future date) they want to leave future inhabitants of planet Earth [. . .]" [22] (p. 126f).

It is these choices that have normative implications. Take for example a study [30] forecasting both CO_2 emission and energy demand that will arise from the transportation sector in Turkey until 2050 based on machine learning algorithms. Such a forecast necessarily includes assumptions about preferences that persons living in the time frame from 2022–2050 will pursue that are tied to emissions, energy use, and choice of transportation means. However, the longer the time frame of the forecast, the more difficult it will be to anticipate the preferences. A longer time frame of the forecast will also complicate the task of anticipating what the pursuit of these preferences will require, e.g., regarding the use of energy, the emission of greenhouse gases, or the choice of transportation means. This is because the use of these—broadly understood—resources such as the use of energy are tied to the pursuit of preferences but do not represent preferences in themselves. People usually do not enjoy emitting CO_2 but partake in activities that can stand in a causal relation to emissions, such as living in adequately heated buildings when the outside temperature is low. Over longer periods of time, both these causal relations, as well as the preferences, can change.

A simple approach to these assumptions about future preferences within AI-supported assessments could be to presuppose that the preferences of persons in the distant future, including future persons, are broadly overlapping with those of current persons. However, this way to proceed may raise the challenge of a so-called transfer of data bias [31] (p. 4), a challenge especially important in machine learning and its reliance on historic data for training purposes [32] (p. 6f). Simply 'transferring' present preferences may bear the risk of providing insufficiently for opportunities that should be left open for future persons because either future persons' preferences change significantly or the circumstances in which these preferences can be satisfied change. Most importantly, the satisfaction of preferences such as mobility may rely on very different sets of resources in differing circumstances, thus leaving future persons with different opportunities. The fact that resources

may provide different individuals in different circumstances with highly heterogeneous opportunities has been extensively discussed as the issue of "conversion factors" within the literature on the Capabilities Approach [33]. Besides the potentially differing individual conversion of resources, it is even unclear from a philosophical point of view if future persons should be provided *with the same* opportunities. This has been an issue of debate between the adherents of the four most discussed intergenerational "principles" of justice of either equality, proportionality, priority, or sufficiency [5] (p. 7448). To date, there neither emerged a consensus within this philosophical debate nor is AI technology suited to integrate all (theoretical) facets of the debate. However, this specific type of transfer bias, which I have framed as *intergenerational transfer bias*, as well as encompassing questions regarding the choice and extent of opportunities that should be left open for future persons, requires AI applied in these contexts once again to be open for revision. Similar solutions have been proposed for the difficulty of including AI's potential impacts on non-human animals [31] (p. 6). This way, potentially adapted preferences or changed circumstances may be added to the algorithms. In other cases, considering the uncertainty about future persons' preferences may require present persons to provide for broader "choice options" that leave the realisation of different preferences in the distant future open (see [7] (p. 53) and [34] (p. 206ff). How this can be realised within AI-based assessments will constitute a challenge for those involved in the design and implementation of these systems.

5. Non-Reciprocity and Indirect Involvement

Unlike with other issues of fairness or justice raised by using AI [3] (p. 71f), the involvement of stakeholders cannot contribute solutions to the presented issues of intergenerational justice. As future persons are yet unborn, there is no reciprocity between future and present persons. An involvement of future persons can thus only be accomplished indirectly.

The success of indirectly involving future persons by present persons' concern for the well-being of the former can, however, be rather limited [35] (p. 19). A more promising way to take aspects of intergenerational justice into account when using AI is to develop a set of evaluative criteria. As a result of the normative challenges described before, a list of questions guiding the potential revision of AI used in context with long-term impacts emerges (cf. Table 1). The first category of questions is targeted at shaping AI in a way that makes especially those features accessible for potential revision that can have negative impacts on future persons. This way, the threat of having no data on potential detrimental impacts [36] (p. 9) ought to be avoided. Further aspects and data will have to be added. Thus, in the environmental context, a specific focus on irreversible costs such as the acceleration of biodiversity loss or the generation of hazardous waste may have to be added to the evaluation.

The second category of questions supporting the use and assessment of AI in contexts with long-term impacts is targeted at assessing whether the use of AI *itself* negatively impacts future persons. Whereas most of the questions raised above reveal the necessity to revise tools of assessments that are also being operated without AI, the use of AI may itself raise additional challenges to the realisation of intergenerational justice. Here, it is the threat of overseeing insights into potentially detrimental impacts [36] (p. 9) on future persons from available data, as well as the occurrence of unintended adverse impacts [32] (p. 8), that is being targeted. The environmental costs of running AI are an example of a negative impact that refers to AI *itself*, i.e., a genuine impact on future persons caused by using AI.

Table 1. Artificial Intelligence (AI) and Intergenerational Justice: Assessment Questions.

Time Frame	What is the time horizon of the assessment that is supported or entirely conducted by the AI? Is the scope of evaluation surpassing ≈ 20 years, thus making the anticipation of future preferences of both the yet-unborn and those already alive more difficult? If yes, issues of intergenerational justice may be affected by this specific use of AI.		
	Cost–benefit analysis	Does the analysis involve weighing benefits for different people at different times? If yes, a series of follow-up questions guides the further evaluation:	
		Discount rate	How has the discount rate been set? For what reasons?
		Burden distribution	How are potential costs of a project being distributed between different people at different times? Does the assessment assign excessively high burdens to a particular sub-group? Is there an intergenerational transfer bias?
		Cost definition	How are costs as a backside of the benefits being defined (e.g., static or dynamic) and assessed?
		Benefit definition	On what assumptions about the preferences of potentially affected persons have the benefits been defined?
AI itself as impact	Does the use of AI have negative impacts on future persons that are directly linked to methods and infrastructure of AI itself?		
	Environmental impact	Is the environmental impact of AI in proportion to its potential positive impact?	

Overall, this list of assessment questions will have to be adapted and revised on a regular basis as it serves to ethically accompany nascent technologies [31] (p. 8). The hope is to provide a normatively informed standard for using AI "properly", i.e. in accordance with intergenerational justice:

> "If AI is underutilised or misused, it may undermine existing environmental policies, slow down efforts to foster sustainability, and impose severe environmental costs on current and future generations. However, if used properly, AI can be a powerful tool to develop effective responses to the climate emergency. Policymakers and the research community must act urgently to ensure that this impact is as positive as possible, in the interest of an equitable and sustainable future" [37] (p. 779).

The list of normative questions adds to this endeavour of realising AI that is sustainable, where intergenerational justice as one of the two ethical dimensions of sustainability provides a central normative standard to assess AI's sustainability. Starting with the question of whether and, if so, to what extent AI can be sustainable, the presented research developed a normative framework that attempts to integrate major aspects of intergenerational justice which, in turn, can be applied to assess different uses of AI. The application of this framework to specific uses of AI with potentially significant long-term impacts, namely, decision support for climate mitigation and environmental protection policies, resulted in the list of assessment questions presented above. A major implication that has been deduced is the necessity to make AI transparent and open for revision, especially with regard to the setting of a social discount rate and the assumptions about future persons' preferences whenever it is used in this context.

6. Discussion and Outlook: Towards the Sustainability of AI

Measuring the use of AI against the standard of intergenerational justice may overburden the involved technologies. If current decision-making procedures, especially about policies with important impacts on future persons, do not fulfil this standard, why should AI? For instance, the German Federal Constitutional Court ruled in March 2021 that the provisions of the Federal Climate Change Act and its governing national climate targets are insufficient regarding the emission regulations because it shifts an excessively large part of the mitigation burden to future persons [38]. The standard of intergenerational justice is thus already presenting severe challenges to policy-making in general. In addition, the normative approaches to intergenerational justice are highly debated and "[. . .] fall astonishingly short of expectations in attempting to deal with the normative issues raised by environmental and resource depletion problems" [16] (p. 61). This may impede the attempt to use them as guidelines for AI design.

Two replies are in order. First, even if intergenerational justice is a contested issue, this does not rule out normative guidance. It rather urges to reveal the choice and reasons for the selection of specific normative premises regarding future persons (see for a similar point regarding sustainability [7] p. 50). The presented list of guideline questions constitutes a framework that supports this endeavour. Impacts on future persons and their normative evaluation thus constitute a further application context for the criteria of transparency and explainability within the debate about AI.

Second, AI technology may even facilitate the application of intergenerational justice as a normative standard. AI's potential to reduce institutional inefficiency in the context of environmental degradation, climate mitigation, or sustainability policies has already been noted (see e.g., [3] p. 69 and [32]). Regarding the intergenerational impact of policies, AI that has been designed and developed in accordance with normative criteria such as those described above may even be employed as a corrective tool by disclosing settings that refer to contested issues of intergenerational justice.

For the time being, however, the use of AI is faced with several constraints regarding intergenerational justice: "[. . .] AI system adoption practices are heavily technologically determined and reductionist in nature, and do not envisage and develop long-term, ethical, responsible and sustainable solutions" [39] (p. 3) (see also [32]). One such reduction is the reduction of the standard of sustainability to the attempt of reducing environmental costs. Unsurprisingly, AI will thus not be able to realise sustainability in itself and instead needs to be included in an encompassing vision as "[. . .] many of our current sustainability interventions via IT are measures to reduce unsustainability instead of creating sustainability, which means that we have to significantly shift our thinking towards a transformation mindset for a joint sustainable vision of the future" [4] (p. 11).

The elaborated normative framework provides a list of assessment questions that explore normative issues regarding impacts on future persons and subsequently the potential need for revision of AI techniques within such a technological approach to a sustainable future. In so doing, insights about how AI can be made more sustainable become apparent. This way, AI may contribute to the pervasive political effort of promoting sustainable development.

To this end, topics for future research are distributed between different scientific disciplines. As an addendum to the ethically informed analysis, the future AI-based support for policies on climate mitigation and environmental protection, and its conformity with the concept of sustainability ought to be assessed from the perspective of policy research. The above-developed framework and assessment guide is conceptualised as a normative module that can be complemented by further normative modules. These would have to represent, for example, issues of *intra*generational justice and the use of natural resources as the second ethical dimension of sustainability. Furthermore, they would have to be interlinked with more empirically oriented sustainability assessments of AI to form an encompassing standard assessing the sustainability of AI. Attempts for more encompassing evaluations of AI and its impacts on sustainability have been conducted

against the UN's sustainable development goals (SDGs) [40–42], however, not representing issues of intergenerational justice. Also, future research topics include the question of how the policy decisions support provided by AI can be designed to be open for revision in the relevant way described above.

7. Conclusions

The analysis developed a normative framework to assess whether and, if so, to what extent the development and use of AI can be sustainable from the specific normative angle of intergenerational justice. Starting from the observation that recent calls for more sustainable AI are based on a narrow understanding of sustainability, it instructed a return to intergenerational justice as a central ethical dimension of sustainability. This contributed to a conceptually informed understanding of sustainability, moving beyond an equation of sustainability with the reduction of environmental costs. The normative framework used intergenerational power asymmetries, as well as the intergenerational relations of non-reciprocity and uncertainty to explore specific uses of AI that raise issues of intergenerational justice. Due to its long-term impacts, the policy decisions support provided by AI in the context of climate mitigation and environmental protection was identified as a significant application field in need of a normative assessment. More specifically, the setting of a social discount rate and the assumptions about future persons' preferences within AI-supported policy assessments were presented as potentially having detrimental impacts on future generations. A major implication has thus been the insight that AI must be made transparent and open for revision, especially with regard to social discounting and assumed preferences over large time horizons. To instruct the implementation of these insights, the analysis provided a list of assessment questions that constitute a first guideline for the revision of AI techniques. It operationalises key aspects of intergenerational justice as one of the constitutive concepts of sustainability and thus contributes a normative module for an ethically informed assessment of the sustainability of AI.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. European Commission (EC). *The European Green Deal. COM (2019) 640 Final*; European Commission: Geneva, Switzerland, 2019; Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1576150542719&uri=COM%3A2019%3A640%3AFIN (accessed on 21 February 2022).
2. Van Wynsberghe, A. Sustainable AI: AI for sustainability and the sustainability of AI. *AI Ethics* **2021**, *1*, 213–218. [CrossRef]
3. Coeckelbergh, M. AI for climate: Freedom, justice, and other ethical and political challenges. *AI Ethics* **2021**, *1*, 67–72. [CrossRef]
4. Khakurel, J.; Penzenstadler, B.; Porras, J.; Knutas, A.; Zhang, W. The rise of artificial intelligence under the lens of sustainability. *Technologies* **2018**, *6*, 100. [CrossRef]
5. Stumpf, K.H.; Baumgärtner, S.; Becker, C.U.; Sievers-Glotzbach, S. The Justice Dimension of Sustainability. A Systematic and General Conceptual Framework. *Sustainability* **2015**, *7*, 7438–7472. [CrossRef]
6. Beckerman, W. 'Sustainable Development': Is it a Useful Concept? *Environ. Value* **1994**, *3*, 191–209. [CrossRef]
7. Barry, B. Sustainability and intergenerational justice. *Theoria* **1997**, *44*, 43–64. [CrossRef]
8. Ott, K. The case for strong sustainability. In *Greifswald's Environmental Ethics. From the Work of the Michael Otto Professorship at Ernst Moritz Arndt University. 1997–2002*; Ott, K., Thapa, P.P., Eds.; Steinbecker: Greifswald, Germany, 2003; pp. 59–64.
9. European Commission. European Commission. European Group on Ethics in Science and New Technologies (EGE). In *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*; European Commission: Brussels, Belgium, 2018; Available online: https://data.europa.eu/doi/10.2777/786515 (accessed on 21 February 2022).
10. Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. *arXiv* **2019**, arXiv:1906.02243.
11. Vasconcellos Oliveira, R. Back to the Future: The Potential of Intergenerational Justice for the Achievement of the Sustainable Development Goals. *Sustainability* **2018**, *10*, 427. [CrossRef]
12. Spijkers, O. Intergenerational Equity and the Sustainable Development Goals. *Sustainability* **2018**, *10*, 3836. [CrossRef]

13. United Nations General Assembly. *Transforming our World: The 2030 Agenda for Sustainable Development, Resolution 70/1, Adopted 25 September 2015*; United Nations: New York, NY, USA, 2015.
14. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef]
15. Cowls, J.; Tsamados, A.; Taddeo, M.; Floridi, L. The AI Gambit—Leveraging artificial intelligence to combat climate change: Opportunities, challenges, and recommendations. *AI SoC* **2021**. [CrossRef] [PubMed]
16. Gosseries, A. Theories of intergenerational justice: A synopsis. *SAPIENS* **2008**, *1*, 61–71. [CrossRef]
17. Rolnick, D.; Donti, P.L.; Kaack, L.H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A.S.; Milojevic-Dupont, N.; Jaques, N.; Waldman-Brown, A.; et al. Tackling climate change with machine learning. *arXiv* **2019**, arXiv:1906.05433.
18. Ott, K. Institutionalizing Strong Sustainability: A Rawlsian Perspective. *Sustainability* **2014**, *6*, 894–912. [CrossRef]
19. United Nations (UN). Report of the world commission on environment and development. In *Our Common Future*; Oxford University Press: Oxford, UK, 1987; Available online: http://www.un-documents.net/wced-ocf.htm (accessed on 21 February 2022).
20. Meyer, L. Intergenerational Justice. In *The Stanford Encyclopedia of Philosophy*; Stanford University: Stanford, CA, USA, 2021; Available online: https://plato.stanford.edu/archives/sum2021/entries/justice-intergenerational (accessed on 21 February 2022).
21. Parfit, D. *Reasons and Persons*, 3rd ed.; Oxford University Press: Oxford, UK, 1987.
22. Weyant, J. Some contributions of integrated assessment models of global climate change. *Rev. Environ. Econ. Policy* **2017**, *11*, 115–137. [CrossRef]
23. Milano, M.; O'Sullivan, B.; Gavanelli, M. Sustainable policy making: A strategic challenge for artificial intelligence. *AI Mag.* **2014**, *35*, 22–35. [CrossRef]
24. Sánchez, J.M.; Rodríguez, J.P.; Espitia, H.E. Review of artificial intelligence applied in decision-making processes in agricultural public policy. *Processes* **2020**, *8*, 1374. [CrossRef]
25. Davidson, M.D. Climate change and the ethics of discounting. *WIREs Clim Change* **2015**, *6*, 401–412. [CrossRef]
26. O'Neill, J. *Ecology, Policy and Politics: Human Well-Being and the Natural World*; Routledge: London, UK; New York, NY, USA, 2002.
27. Broome, J. Discounting the Future. *Philos. Public Aff.* **1994**, *23*, 128–156. [CrossRef]
28. Gillingham, K.; Stock, J.H. The cost of reducing greenhouse gas emissions. *J. Econ. Perspect.* **2018**, *32*, 53–72. [CrossRef]
29. Mir, A.A.; Alghassab, M.; Ullah, K.; Khan, Z.A.; Lu, Y.; Imran, M. A review of electricity demand forecasting in low and middle income countries: The demand determinants and horizons. *Sustainability* **2020**, *12*, 5931. [CrossRef]
30. Ağbulut, Ü. Forecasting of transportation-related energy demand and CO_2 emissions in Turkey with different machine learning algorithms. *Sustain. Prod. Consum.* **2022**, *29*, 141–157. [CrossRef]
31. Galaz, V.; Centeno, M.A.; Callahan, P.W.; Causevic, A.; Patterson, T.; Brass, I.; Baum, S.; Farber, D.; Fischer, J.; Garcia, D.; et al. Artificial intelligence, systemic risks, and sustainability. *Technol. Soc.* **2021**, *67*, 101741. [CrossRef]
32. Nishant, R.; Kennedy, M.; Corbett, J. Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. *Int. J. Inf. Manag.* **2020**, *53*, 102104. [CrossRef]
33. Robeyns, I.; Byskov, M.F. The Capability Approach. In *The Stanford Encyclopedia of Philosophy*; Stanford University: Stanford, CA, USA, 2021; Available online: https://plato.stanford.edu/archives/win2021/entries/capability-approach (accessed on 21 February 2022).
34. Halsband, A. *Konkrete Nachhaltigkeit. Welche Natur wir für künftige Generationen erhalten sollten*; Baden-Baden: Nomos, Germany, 2016.
35. Klockmann, V.; Von Schenk, A.; Villeval, M.C. Artificial Intelligence, Ethics, and Intergenerational Responsibility. *GATE Work Pap.* **2021**. [CrossRef]
36. Walsh, T.; Evatt, A.; de Witt, C.S. Artificial Intelligence & Climate Change: Supplementary Impact Report. 2020. Available online: https://www.semanticscholar.org/paper/Artificial-Intelligence-%26-Climate-Change-%3A-Impact-a-Walsh-Evatt/a840e7c4af0f10b3fac136ddc99e31c6c7d58507 (accessed on 21 February 2022).
37. Taddeo, M.; Tsamados, A.; Cowls, J.; Floridi, L. Artificial intelligence and the climate emergency: Opportunities, challenges, and recommendations. *One Earth* **2021**, *4*, 776–779. [CrossRef]
38. German Federal Constitutional Court. Constitutional Complaints against the Federal Climate Change Act Partially Successful. Press Release No. 31/2021 of 29 April 2021. Order of 24 March 2021. 1 BvR 2656/18, 1 BvR 288/20, 1 BvR 96/20, 1 BvR 78/20. Available online: https://www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/EN/2021/bvg21-031.html (accessed on 21 February 2022).
39. Yigitcanlar, T.; Mehmood, R.; Corchado, J.M. Green artificial intelligence: Towards an efficient, sustainable and equitable technology for smart cities and futures. *Sustainability* **2021**, *13*, 8952. [CrossRef]
40. Vinuesa, R.; Azizpour, H.; Leite, I.; Balaam, M.; Dignum, V.; Domisch, S.; Felländer, A.; Langhans, S.D.; Tegmark, M.; Fuso Nerini, F. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nat. Commun.* **2020**, *11*, 233. [CrossRef]
41. Truby, J. Governing Artificial Intelligence to benefit the UN Sustainable Development Goals. *Sustain. Dev.* **2020**, *28*, 946–959. [CrossRef]
42. Sætra, H.S. AI in Context and the Sustainable Development Goals: Factoring in the Unsustainability of the Sociotechnical System. *Sustainability* **2021**, *13*, 1738. [CrossRef]

sustainability

MDPI

Article

Our New Artificial Intelligence Infrastructure: Becoming Locked into an Unsustainable Future

Scott Robbins [1,*] and Aimee van Wynsberghe [2]

[1] Center for Science and Thought, University of Bonn, Poppelsdorfer Allee 28, 53115 Bonn, Germany
[2] Institute for Science and Ethics, University of Bonn, Bonner Talweg 57, 53113 Bonn, Germany; aimee@uni-bonn.de
* Correspondence: srobbins@uni-bonn.de

Abstract: Artificial intelligence (AI) is becoming increasingly important for the infrastructures that support many of society's functions. Transportation, security, energy, education, the workplace, the government have all incorporated AI into their infrastructures for enhancement and/or protection. In this paper, we argue that not only is AI seen as a tool for augmenting existing infrastructures, but AI itself is becoming an infrastructure that many services of today and tomorrow will depend upon. Considering the vast environmental consequences associated with the development and use of AI, of which the world is only starting to learn, the necessity of addressing AI alongside the concept of infrastructure points toward the phenomenon of carbon lock-in. Carbon lock-in refers to society's constrained ability to reduce carbon emissions technologically, economically, politically, and socially. These constraints are due to the inherent inertia created by entrenched technological, institutional, and behavioral norms. That is, the drive for AI adoption in virtually every sector of society will create dependencies and interdependencies from which it will be hard to escape. The crux of this paper boils down to this: in conceptualizing AI as infrastructure we can recognize the risk of lock-in, not just carbon lock-in but lock-in as it relates to all the physical needs to achieve the infrastructure of AI. This does not exclude the possibility of solutions arising with the rise of these technologies; however, given these points, it is of the utmost importance that we ask inconvenient questions regarding these environmental costs before becoming locked into this new AI infrastructure.

Keywords: sustainable AI; artificial intelligence; AI ethics; climate justice; infrastructure

Citation: Robbins, S.; van Wynsberghe, A. Our New Artificial Intelligence Infrastructure: Becoming Locked into an Unsustainable Future. *Sustainability* **2022**, *14*, 4829. https://doi.org/10.3390/su14084829

Academic Editor: Amir Mosavi

Received: 28 February 2022
Accepted: 15 April 2022
Published: 18 April 2022

1. Introduction

Artificial intelligence (AI) is becoming increasingly important for the infrastructures that support many of society's functions. Transportation, security, energy, education, the workplace, government, have all incorporated AI into their infrastructures for enhancement and/or protection. Not only is AI seen as a tool for augmenting existing infrastructures, but AI itself is becoming an infrastructure that many services of today and tomorrow will depend upon. There is a growing body of research addressing the impact of AI on the environment. This body of literature shows that AI development and use requires an amazing amount of computational power which creates increased carbon emissions. The effects of AI on environmental justice will be vast considering too the mining of precious minerals and the vulnerable demographics exploited for these processes. This is deeply concerning given the grave situation the world finds itself in regarding the climate. The Intergovernmental Panel on Climate Change (IPCC) goes so far as to say that it is "code red for humanity" [1]. Given that there is high confidence that climate change is to a large extent human-induced [2], we should be asking more questions before introducing a new human-made carbon-emitting infrastructure powered by AI.

The field of sustainable AI has been put forward as a way of addressing the environmental justice issues associated with AI throughout its lifecycle [3]. Sustainable AI is

about more than applying AI to achieve climate goals (Though much work in the field is devoted to this idea. See e.g., [4–10]), it is about understanding and measuring the environmental impact of developing and using AI. The little information we have on the environmental impact of AI is, to say the least, not encouraging [11]. Thus, many of the questions surrounding the sustainability of AI remain unanswered. These answers are needed for society to make an informed choice regarding the use of AI in a particular context. This makes AI a huge environmental risk as AI continues to be implemented in a broad range of contexts despite this opacity regarding its environmental consequences.

It may not be immediately clear why AI researchers and developers, in particular, must pay attention to issues of environmental sustainability. Does not everything need to consider issues of sustainability? In this paper, we argue that the environmental consequences associated with AI are essential issues of AI ethics. The way we choose to build and implement AI today will have profound consequences for our future sustainability that warrants a specific focus on its sustainability. This special attention is due to the connection between AI and the concept of infrastructure.

In what follows, we illustrate how AI has traditionally been understood as conceptually distinct from infrastructure. From this vantage point, AI can be used to enhance or protect existing infrastructures. We also point out that AI is dependent on vast infrastructures which are climate intensive, e.g., AI needs electricity, precious minerals, data to be transferred, etc. AI is increasingly being used to power the next generation of digital services. That is, AI is now the infrastructure relied upon by digital services. Look to the Facebook outage of 2021 that showed how many businesses in Ghana were unable to function without the Facebook infrastructure. Facebook's services are AI-powered services. Everything from how content is displayed, moderated, and sorted is powered by AI [12]. Furthermore, the advertising ecosystem which Facebook makes money from is AI-powered [13]. It is safe to say that without AI there is no Facebook. Consider also, the business model of social networking companies that rely on targeted advertising to generate revenue. The necessity of addressing AI alongside the concept of infrastructure points toward the phenomenon of carbon lock-in—whereby society's ability to technologically, economically, politically, and socially reduce carbon emissions are constrained due to the inherent inertia created by entrenched technological, institutional, and behavioral norms [14]. The negative outcomes that AI adoption creates may also give rise to innovative, environmentally sound, solutions. However, without knowing the extent of the problem and giving that problem the attention it deserves, those solutions will never come about. Given these points, we must ask inconvenient questions regarding these environmental costs before becoming locked into this new AI infrastructure. No amount of convenience provided by AI can justify further decimating our planet.

2. AI Ethics and the Sustainability of AI

It is important here to note what we mean by AI. The concept is overused and can refer to many different things. For this article, AI refers to the methodology of creating algorithms driven by the rise of machine learning (ML). ML algorithms "use statistics and probability to "learn" from large datasets" [15]. This learning is not restricted to picking out features that humans could understand—which gives the resulting algorithm greater power than we have seen before. This is a pragmatic definition as it excludes other methodologies which should fall under the definition of AI. For example, expert systems and decisions trees were for decades the only AI algorithms out there. However, they are not what is driving the rise of AI in our everyday lives and their impact is similar to traditional software applications. ML is what has driven the need for more data, sensors, computing power, etc. We would prefer that everyone simply use ML rather than AI (because that is usually what is being referred to). However, as it stands, AI is the standard concept that people hear about in academic literature, popular culture, and the media. In what follows we use AI to refer to ML algorithms. This means that while autonomous cars and medical technologies are not AI—more and more of these technologies are powered by ML algorithms.

As AI applications increase across society AI ethicists have begun to uncover risks associated with the technology. Risks, for example, concerning the use of historical data to train algorithms when such historical data embeds stereotypes and discriminatory assumptions about individuals and groups in society. The consequence of this practice is oftentimes further discrimination of said individuals and groups. AI ethics, in short, is dedicated to uncovering and understanding the ethical issues associated with the development and use (i.e., the entire life cycle) of AI-powered machines—how does AI threaten the ability of individuals and groups to live a "good life". Once the risks have been identified it is then the goal to prevent and/or mitigate said risks.

The field of AI ethics has grown in importance in the last decade seen through an increase in academic publications on the topic (For example the Berkman Klein Center identified 36 "prominent AI principles" documents [16]), the involvement of AI ethics in the policy forum (e.g., European Commission High-Level Expert Group on AI) [17], and the adoption of AI ethics into the business and consulting space (see e.g., [18,19]). In each of these sectors, there are certain canonical ethical issues pertaining to AI that are being discussed, most often concerning particular AI methodologies. Machine learning, for example, has been described as a method that creates a kind of opacity given that it is often impossible to know and/or to understand the rules generated by the model used to make a prediction. Stemming from this technical feature come ethical issues related to transparency (e.g., should a particular technology be used if it is impossible to understand how it arrives at an output); responsibility (e.g., should a particular technology be used if this lack of transparency leads to confusion in terms of who is responsible for the consequences of a decision that are not known or understood by the programmer); and, security (e.g., how can we ensure that security of a system when we do not entirely understand its functioning). To be sure, none of these concerns have been rectified.

Without diminishing the significance of the above issues, it is also important to note that little attention has been paid, to date, to the environmental consequences of making and using AI. A small group of researchers has begun to study carbon emissions [11] and computing power [20]; however, there is little incentive for academics and/or industry to incorporate this systematically into research and production methods. There is no regulation to demand an environmental assessment of the impacts of making and/or using AI/ML. The systematic accounting of these environmental impacts is necessary to have a better idea of the large-scale impact of making and using AI systems. Moreover, "accurate accounting of carbon and energy impacts aligns with energy efficiency [21], raises awareness, and drives mitigation efforts, among other benefits" [20]. It is this connection—between AI and environmental consequences—that drives the points made in this paper. Namely, that we must know the specifics of this connection before we become (more) dependent on AI.

To be sure, the environmental costs of making and using AI do not end with direct carbon emissions or computing power. The systems used to create and run AI models require precious minerals that are mined in often horrible conditions for the individuals involved [22]. There is water needed for the cooling of the computing centers. There will be electronic waste (e-waste) resulting from the updating of materials, computers, and data centers. Historically, e-waste has been dumped in underdeveloped countries exposing inhabitants to the toxic chemicals in their water supplies and agricultural [23]. These concerns are essentially issues of environmental justice and while they focus on environmental consequences, they point to societal concerns that have been, to date, invisible from public discourse. As Hasselbalch describes, data ethics is not only about power but also is power [24]. AI ethics is not only about power asymmetries but is power in so far as the loudest voices are the ones who determine the ethical issues of importance and priority. The movement to focus on sustainability is about revealing the hidden demographics who suffer and will continue to suffer as AI becomes more and more pervasive in our daily lives.

Sustainable AI was first defined by van Wynsberghe in 2021 as a "movement to foster change in the entire lifecycle of AI products (i.e., idea generation, training, re-tuning, implementation, governance) towards greater ecological integrity and social justice" [3]. Given the high costs already identified, we suggest that AI researchers (both ethicists and computer scientists) along with AI practitioners (AI developers) and policymakers (those involved with drafting legislation concerning the governance of AI) ought to shift focus to explicitly, and quickly, address the hidden environmental costs associated with AI development and usage.

Reframing AI ethics discussions in terms of sustainability opens up novel insights. First, to use the phrase "sustainable AI" demands that one consider sustainability as a value within the AI ethics domain, one that is deserving of greater attention. Second, the label of sustainability invokes the recognition of the environment as a starting point for addressing AI ethics issues. The environment becomes a lens through which societal and economic issues are uncovered, conceptualized, and understood. Third, sustainability as a concept emphasizes issues of intra- and inter-generational justice. Attention to environmental consequences demands consideration of the impacts, and our responsibilities to mitigate said impacts, on younger generations as well as those yet to come.

Fourth, sustainability demands the recognition of AI on a larger scale rather than on one or two specific applications. To date, the focus of AI ethics has been on mitigating concerns of privacy, safety, and fairness, to name a few. With this narrow view of the impacts of AI, researchers run the risk of overlooking the larger structural issues of AI as infrastructure, researchers cannot see "the forest from the trees". By this, we mean that in focusing on issues of design, or how to implement the technology, researchers to date have been unable to take a step back and understand the magnitude of AI development and use. AI is not one or two models that will be restricted to a particular sector or for a particular application. Instead, AI is being promoted as an innovation suitable for any sector, for any application. From our perspective, it is thus paramount to address AI alongside the notion of infrastructure.

3. AI and Infrastructure

We begin by asking: "What is the relationship between AI and infrastructure?" As Kate Crawford describes in her book "Atlas of AI" there is a fascinating phenomenon concerning the materiality of AI/ML; the language used to describe the materials refers to algorithms and "the cloud", making AI seem not of the physical. However, in reality, there is a vast physical infrastructure behind the production of AI. Water is needed to cool computing centers and the water obtained for this comes from public infrastructures. Electricity is needed to fuel computing centers and the pipelines through which the electricity travels are often publicly funded networks. Minerals are required for batteries and microchips. These minerals are part of a long chain of procurement in which humans often work in slave-like conditions and degradation of the environment results from the way minerals are sourced (see e.g., [25–27]). These realities are kept hidden to ensure enthusiasm toward AI. Consequently, the hidden materiality of AI fosters a lack of understanding of the breadth of the physical infrastructures powering AI. This does not entail that AI and its materiality are worse than other industries regarding carbon emissions. Rather, the materiality of AI points to a non-negligible impact on the environment. This must be included in the cost-benefit analysis of specific AI-powered services and products.

Not only does the development and use of AI rely on existing infrastructures but AI is seen as a powerful tool to support, enhance, or protect infrastructure. AI was used by Google in 2016 to understan42d how to conserve electricity in their data centers—allowing the company to enhance its energy conservation efforts [28]. AI is used in the banking sector to predict when/if a fraudulent transaction has occurred allowing banks to react faster for their customers' protection [29,30]. AI is used in both the public and private sectors to protect against spam and phishing schemes [31,32]. AI is used across the transportation

sector to enhance in a variety of ways from managing traffic lights [33] to the idea of autonomous vehicles for the reduction of fatalities [34].

As we see, AI can be understood as dependent on existing infrastructure and/or as enhancing existing infrastructure. Our aim now is to argue that AI should itself be understood as infrastructure. And it is this understanding that adds urgency to the environmental concerns. Infrastructure is not easily defined, and we do not attempt here to settle any debates on that subject. What we can do is take some properties of infrastructure and show how they relate to AI.

3.1. Infrastructure Properties

Susan Leigh Star [35] lists 9 such properties (which she calls dimensions). We highlight a few here concerning AI. First, infrastructure has the feature of embeddedness. That is, it exists within other structures, social arrangements, and technologies [35]. AI can clearly be said to have this property as it is embedded into the technologies and structures that we interact with daily, e.g., simple tools such as Google Maps or the advertising shown to us whenever we are online. When AI is implemented, it often does not stand alone—but interacts with the technologies we use and takes data from our social arrangements (and/or actions) to generate its outputs, e.g., advertisements require data from our search history and previous purchases fed to an AI to predict what might be appealing.

Second, is the property of transparency. Infrastructure is transparent to use. When we turn on a light switch, we do not see the infrastructure of wiring and power grids that enable the light to come on. We simply enjoy the convenience of light. Likewise, when we turn on Netflix, we do not see the infrastructure of cables, servers, and algorithms (often AI) that enable those recommendations to populate the home screen. Our attention is drawn to the result that infrastructure enables—not the process that leads to the result. In our many daily interactions with AI, we could be excused for not even knowing that AI was driving what was happening.

Third, infrastructure becomes visible upon breakdown. When infrastructure ceases to function properly our attention directs itself toward that infrastructure. When the light does not turn on upon flipping the light switch, we direct our attention to the fuse box and if that does not work, we may have to call our attention to the company that runs the infrastructure that provides our electricity. Much attention has been given to AI when it functions improperly. When Google's AI-powered image labeling system incorrectly labeled people of color as gorillas it quickly drew people's attention to the algorithm and the data that serves as the infrastructure to that system.

Fourth, infrastructure is modular. Infrastructure does not simply grow from nothing. It is put on top of other infrastructure and must take on the benefits and negatives that come with it. The original wiring of the internet was done through the existing phone lines. Only incrementally was this replaced with fiber optic cables that power the internet that we have today. This is because the infrastructure we have come to rely on has its own inertia—it has to work with the existing infrastructure because we depend on it. AI must also be placed on top of existing infrastructure. It interacts with platforms, algorithms, and the infrastructure that powers the internet. We see new phones with processors that enable AI features [36]—thereby starting the modular process that slowly replaces old infrastructure.

3.2. AI as Infrastructure

This listing of infrastructure properties provides a base of understanding into how AI can already be considered infrastructure and how this will continue in the years to come. Currently, AI is evaluated in terms of its impact on infrastructure (i.e., as being conceptually distinct from infrastructure); however, in (the near) future AI must be evaluated as the infrastructure itself. Following this, any new infrastructure—because of its importance and resistance to change—should be an environmentally sustainable one. Consequently, evaluating AI requires insight into the environmental consequences of understanding AI as infrastructure.

Part of the reason for writing this paper is that the environmental sustainability of AI is unknown—and for the reasons outlined above, this is an unacceptable situation. Furthermore, one cannot state the environmental sustainability of AI in broad strokes. Particular systems in particular contexts may be environmentally sustainable (e.g., green servers) while others not. The point here is that for governments and consumers to make informed decisions regarding AI-powered solutions, the environmental sustainability of those solutions themselves must be known and factored in.

4. Locked in with AI

The crux of this paper boils down to this: in conceptualizing AI as infrastructure, we can recognize the risk of lock-in, not just carbon lock-in but lock-in as it relates to all the physical needs to achieve the infrastructure of AI.

The phenomenon of lock-in is most referenced in terms of carbon lock-in and the concern for greenhouse gas (GHG) emissions. Carbon Lock-In refers to "the dynamic whereby prior decisions relating to GHG-emitting technologies, infrastructure, practices, and their supporting networks constrain future paths, making it more challenging, even impossible, to subsequently pursue more optimal paths toward low-carbon objectives" [37]. Coal power plants are an oft-cited example of a carbon lock-in [37,38]. While expensive to build carbon plants, they are cheap to operate. This creates political, economic, and social conditions that make it difficult to replace this high carbon-emitting infrastructure.

This points to the fact that the choices we make now regarding our new AI-augmented infrastructure not only relate to the carbon emissions that it will have; but also relate to the creation of constraints that will prevent us from changing course if that infrastructure is found to be unsustainable.

Self-driving vehicles require a large amount of energy to capture, store, and process the large amount of data required to navigate their environments. One estimate shows that it takes roughly 2500 Watts, which is enough to light 40 incandescent light bulbs. That is for just one car [39]. Multiple studies have attempted to estimate the energy savings and costs of self-driving cars (see e.g., [40,41]). They factor in the energy that the sensors capturing data consume, the onboard computers and processors, data transfer energy costs, as well as the efficiency gained by automating driving. However, there is a range of variables that are not accounted for in such analyses, such as hardware production. Thus, we argue here that it is not enough to address a limited number of variables; rather the entire system (from procurement to development to recycling) must be considered.

In what follows we highlight some of the major processes that come with the rise of AI. This points to what must be measured and accounted for when we evaluate the cost of a particular AI application. The costs of producing the hardware running the algorithms, the costs of collecting and transmitting data used and processed by AI, the computational cost of training and using the model, the disposal of the network of hardware needed by AI, and the costs of ensuring that the algorithms are aligned with ethical principles all must factor in. This is not supposed to be exhaustive; rather, it should point to the fact that a lot of work must be done before we even have the information necessary to make informed decisions regarding the use of a particular AI system.

4.1. Hardware Production

The hardware used in the AI lifecycle is, to say the least, non-negligible in terms of energy consumption. There are the obvious components such as the servers and their components (e.g., hard drives, GPUs, etc.) that are required to run the algorithms and store large amounts of data. However, there are also many devices used to collect data such as video cameras, lidar sensors, motion detectors, and so on. It has been shown that the manufacturing of these devices "as opposed to hardware use and energy consumption, accounts for most of the carbon output attributable to hardware systems" [42]. The rise of "edge computing" is fueling the rise in these devices.

Edge computing has been defined as "the enabling technologies allowing computation to be performed at the edge of the network, on downstream data on behalf of cloud services and upstream data on behalf of IoT services" [43]. This simply means that the processing of data is happening at the edge of the network closer to the source of the data rather than in some central cloud server. This could, for example, mean that facial recognition processing happens on the smart CCTV camera rather than the video footage being sent to the cloud. This can save on the cost of transferring data and reduce the need for energy-intensive cloud servers; however, it increases the need for complex devices. It is estimated that the number of these devices will almost five-fold by 2030 to 7.8 billion devices [44].

Many of these modern technological devices have rare earth elements (REE). For example, REEs are found in "hybrid vehicles, rechargeable batteries, wind turbines, mobile phones, flat-screen display panels, compact fluorescent light bulbs, laptop computers, disk drives, catalytic converters, etc." [45]. The production of these devices has a huge impact—not only on the environment but also on vulnerable populations that suffer human rights violations [25]. The environmental impacts are not fully understood; however, it is understood that they are significant: "REE mining and refining generate significant amounts of liquid and solid wastes, with potentially deleterious effects on the environment, and it is expected to continue increasing in the future because they are irreplaceable in many technological sectors" [45].

As we increasingly depend upon AI-powered technologies, we increase the need for REEs and the processes that produce them. Currently, these processes are terrible for the environment and the people who work in the mines. This cost cannot be ignored when we tally the benefits and consequences of delegating more and more to AI.

4.2. Data Collection and Transmission

AI applications require input data to be processed. This data can come from virtually any kind of data. Security services use video feeds as input for facial recognition algorithms. Biometric sensors attached to people (e.g., smartwatches) collect and send data to healthcare AI algorithms used to detect, for example, heart problems. Smart cities use a vast network of sensors and devices (see Section 4.1 above) to collect data to use as inputs for many AI algorithms promising to better our cities. It has been calculated that "Internet use has a carbon footprint ranging from 28 to 63 g CO_2 equivalent per gigabyte (GB)" [46]. The power necessary to keep these sensors active as well as the energy required to transmit this data is non-negligible.

The Shift Project estimates that the digital era is responsible for 4% of greenhouse gas emissions in 2020 [47]. This is similar to the emissions caused by pre-Covid level commercial aviation [48]. The Shift Project further estimates an 8% rise year over year due to several factors including the rise of the Internet of Things (IoT) and an explosion in data traffic [47]. Increasingly relying upon AI will exacerbate these factors.

The massive amounts of video, image, pollution, temperature, biometric, radar, lidar, etc. data that must be transmitted to cloud servers for processing by AI algorithms takes energy. By increasingly relying upon AI to run our society, we become locked into needing this vast network of data transmission. We should know more about its energy cost to responsibly evaluate whether or not certain AI applications are worth it.

4.3. AI Model Creation and Data Processing

The most often cited statistic regarding the creation of AI models is that common large AI models emit more than 626,000 pounds of carbon dioxide—equivalent to five times the lifetime emissions of an automobile [11]. While this number may be far lower depending on the specific context—for example, when designers are simply fine-tuning a model that has already been trained—there is no question that AI requires an exponentially increasing amount of computing power [49]. Once the model is trained and the algorithm is live, inputs must be given to that model for processing. Videos, images, text, sound, etc. all need to be classified using the model in question. This has its own associated cost—and

with, for example, video input, this can be a large cost. Efforts are being made to reduce this cost by, for example, only feeding the model-specific frames of the video rather than the whole thing. Other methods are also being explored [50].

Once the hardware is set up, the coding is done, the model is trained on the collected training data and everything is running smoothly, there is the problem that all of this will need updating. We learned that many AI systems failed during the COVID-19 pandemic simply because our behavior changed drastically—making many ML models useless [51]. New behavior requires new models—which can then cause some of the processes listed above to need to be re-done—furthering the environmental impact, in terms of carbon emissions, of these systems.

4.4. Hardware Disposal

Finally, the process of recycling and disposing of hardware must be accounted for. In 2019 the world generated "53.6 Mt [million metric tons] of e-waste . . . and is projected to grow to 74.7 Mt by 2030" [52]. This, of course, factors in all types of e-waste including appliances and personal devices, and not just AI devices alone. The point is that an increased reliance upon AI will require the disposal of more e-waste. While it may seem reasonable for anyone with AI application design not to spend time thinking about this; ignoring this fact while setting up a society that depends more and more on AI would be a critical failure.

Not all computer hardware is used to power AI; however, AI requires an extreme amount of computational power—which requires not only more hardware—but new hardware. Anything which relies upon computer hardware should factor in the cost of the disposal and recycling of that hardware. Here we only want to point out that this is also a cost of using AI. Furthermore, "there is a growing demand for specialized hardware accelerators with optimized memory hierarchies that can meet the enormous compute and memory requirements of" machine learning [53]. McKinsey, in a report, found that "AI-related semiconductors will see growth of about 18 percent annually over the next few years—five times greater than the rate for semiconductors used in non-AI applications" [54]. This shows that there is a rise in hardware specifically designed for AI.

There must be a plan for the recycling of all of this hardware—and the environmental cost associated with such recycling must be factored in when setting up a society dependent on AI and the hardware it requires.

4.5. Ethics Alignment

The rise of AI has precipitated a rise in those pointing out that there are many ethical issues associated with AI. Methods for overcoming these risks have been proposed and implemented. Some of these methods themselves come with a cost. For example, many contemporary AI methodologies (e.g., deep neural networks) are not explainable. That is, the considerations which contribute to the output are unknown to even the designers of the algorithm [15,55–57]. When we are delegating the task of certain decisions to AI this lack of explanation will not be acceptable. Delegating judicial decisions [58] or moral decisions [59] to AI requires an explanation for the outputs generated.

Various methodologies have come out to overcome this lack of explainability. For example, one proposal has suggested that we can use counterfactual explanations. That is, an explanation can be provided by knowing the smallest change in the input that would yield a positive outcome [60]. Visual methods that apply to specific models have also been proposed such as Gradient and Guided Back Propagation. These yield visual explanations which may show us which features of an image most contributed to an output. Other methods are more general, for instance LIME and SHAP, which aim to highlight feature importance for a particular output (for a review of such methods see e.g., [61]).

These methods require their own trained model which then exacerbates the environmental costs pointed out in the above sections. When the use of AI will require the use of

more AI to overcome ethical issues, then the environmental cost of this further model must also be calculated.

5. Conclusions

It is no secret that AI requires a vast amount of energy to accomplish its tasks. Any industry uses energy to accomplish its tasks. What we have shown to be special about AI is that AI is increasingly becoming the infrastructure that is required for society to function. Governments, schools, cars, hospitals, banking, etc. are all becoming dependent upon this AI-powered infrastructure. This is a choice that society is making. Choices as important as these cannot be done without thinking about the environmental consequences. And there is little known about the breadth of environmental consequences associated with AI as infrastructure.

Choosing a path that leads to greater harm to the environment is unacceptable. Choosing a path out of ignorance to its impact on the environment is also unacceptable. So far, we are blindly going forward with the creation of a dependence relationship on a technology whose environmental impact, based on the little we do know, is extremely high. While there is much work being done to mitigate this impact, that work, and its results should be known before creating this dependence. We run the risk of locking ourselves into a technological infrastructure that is energy-intensive in both its development and use, as well as energy-intensive to mitigate certain ethical concerns. This is precisely the aim of the Sustainable AI domain—to investigate and make clear that there are a plethora of environmental risks associated with AI and to argue that these risks ought to be the starting point in any ethical analysis of AI/ML.

The argument from large tech companies that most of the energy they use is renewable—and therefore has little impact on the environment is frivolous. The use of energy, renewable or not, during a time that has been called "code red for humanity" is of great importance. The question before any AI model is created should be: is this worth the environmental cost that we will be locked into for decades? The answer will often be no.

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

Funding: This research was funded by The Alexander von Humboldt Foundation (The Alexander von Humboldt Stiftung) in Germany in the form of a Professorship for 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.

References

1. McGrath, M. Climate Change: IPCC Report Is "Code Red for Humanity". *BBC News*, 2021. Available online: https://www.bbc.com/news/science-environment-58130705(accessed on 22 March 2022).
2. IPCC. *Climate Change 2022 Impacts, Adaptation and Vulnerability: Summary for Policymakers*; Intergovernmental Panel on Climate Change: Geneva, Switzerland, 2022.
3. van Wynsberghe, A. Sustainable AI: AI for Sustainability and the Sustainability of AI. *AI Ethics* **2021**, *1*, 213–218. [CrossRef]
4. Vinuesa, R.; Azizpour, H.; Leite, I.; Balaam, M.; Dignum, V.; Domisch, S.; Felländer, A.; Langhans, S.D.; Tegmark, M.; Fuso Nerini, F. The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. *Nat. Commun.* **2020**, *11*, 233. [CrossRef] [PubMed]
5. Tomašev, N.; Cornebise, J.; Hutter, F.; Mohamed, S.; Picciariello, A.; Connelly, B.; Belgrave, D.C.M.; Ezer, D.; van der Haert, F.C.; Mugisha, F.; et al. AI for Social Good: Unlocking the Opportunity for Positive Impact. *Nat. Commun.* **2020**, *11*, 2468. [CrossRef] [PubMed]
6. Sætra, H.S. AI in Context and the Sustainable Development Goals: Factoring in the Unsustainability of the Sociotechnical System. *Sustainability* **2021**, *13*, 1738. [CrossRef]

7. Nishant, R.; Kennedy, M.; Corbett, J. Artificial Intelligence for Sustainability: Challenges, Opportunities, and a Research Agenda. *Int. J. Inf. Manag.* **2020**, *53*, 102104. [CrossRef]
8. Lahsen, M. Should AI Be Designed to Save Us From Ourselves?: Artificial Intelligence for Sustainability. *IEEE Technol. Soc. Mag.* **2020**, *39*, 60–67. [CrossRef]
9. Dauvergne, P. *AI in the Wild: Sustainability in the Age of Artificial Intelligence*; MIT Press: Cambridge, MA, USA, 2020; ISBN 978-0-262-53933-3.
10. Tsolakis, N.; Zissis, D.; Papaefthimiou, S.; Korfiatis, N. Towards AI Driven Environmental Sustainability: An Application of Automated Logistics in Container Port Terminals. *Int. J. Prod. Res.* **2021**, 1–21. [CrossRef]
11. Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. *arXiv* **2019**, arXiv:190602243.
12. Macaulay, T. Here's How AI Determines What You See on the Facebook News Feed. Available online: https://thenextweb.com/news/heres-how-ai-determines-what-you-see-on-facebook-news (accessed on 22 March 2022).
13. Facebook How Does Facebook Use Machine Learning to Deliver Ads? Available online: https://www.facebook.com/business/news/good-questions-real-answers-how-does-facebook-use-machine-learning-to-deliver-ads (accessed on 22 March 2022).
14. Seto, K.C.; Davis, S.J.; Mitchell, R.B.; Stokes, E.C.; Unruh, G.; Ürge-Vorsatz, D. Carbon Lock-In: Types, Causes, and Policy Implications. *Annu. Rev. Environ. Resour.* **2016**, *41*, 425–452. [CrossRef]
15. Robbins, S. AI and the Path to Envelopment: Knowledge as a First Step towards the Responsible Regulation and Use of AI-Powered Machines. *AI Soc.* **2020**, *35*, 391–400. [CrossRef]
16. Fjeld, J.; Achten, N.; Hilligoss, H.; Nagy, A.; Srikumar, M. *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*; Social Science Research Network: Rochester, NY, USA, 2020.
17. High Level Expert Group on AI Ethics Guidelines for Trustworthy AI. Available online: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed on 15 January 2020).
18. AI at Google: Our Principles. Available online: https://www.blog.google/technology/ai/ai-principles/ (accessed on 14 January 2019).
19. Nadella, S. Microsoft's CEO Explores How Humans and A.I. Can Solve Society's Challenges—Together. Available online: https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html (accessed on 14 January 2019).
20. Henderson, P.; Hu, J.; Romoff, J.; Brunskill, E.; Jurafsky, D.; Pineau, J. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *J. Mach. Learn. Res.* **2020**, *21*, 1–43.
21. Schwartz, R.; Dodge, J.; Smith, N.A.; Etzioni, O. Green AI. *Commun. ACM* **2020**, *63*, 54–63. [CrossRef]
22. Crawford, K. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*; Yale University Press: New Haven, CT, USA, 2021; ISBN 978-0-300-20957-0.
23. Vidal, J. Toxic "e-Waste" Dumped in Poor Nations, Says United Nations. *The Guardian*, 2013. Available online: https://www.theguardian.com/global-development/2013/dec/14/toxic-ewaste-illegal-dumping-developing-countries(accessed on 22 March 2022).
24. Hasselbalch, G. *Data Ethics of Power: A Human Approach in the Big Data and AI Era*; Edward Elgar Publishing: Cheltenham, UK, 2021; ISBN 978-1-80220-311-0.
25. Amnesty International. *"This Is What We Die for" Human Rights Abuses in the Democratic Republic of the Congo Power the Global Trade in Cobalt*; Amnesty International: London, UK, 2016.
26. Searcey, D.; Lipton, E.; Gilbertson, A. Hunt for the 'Blood Diamond of Batteries' Impedes Green Energy Push. *New York Times*, 29 November 2021.
27. *Precious Metal, Cheap Labor: Child Labor and Corporate Responsibility in Ghana's Artisanal Gold Mines*; Human Rights Watch: New York, NY, USA, 2015.
28. DeepMind DeepMind AI Reduces Google Data Centre Cooling Bill by 40%. Available online: https://www.deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40 (accessed on 13 April 2022).
29. Pierre, R. Detecting Financial Fraud Using Machine Learning: Winning the War Against Imbalanced Data. Available online: https://towardsdatascience.com/detecting-financial-fraud-using-machine-learning-three-ways-of-winning-the-war-against-imbalanced-a03f8815cce9 (accessed on 30 June 2019).
30. West, J.; Bhattacharya, M. Intelligent Financial Fraud Detection: A Comprehensive Review. *Comput. Secur.* **2016**, *57*, 47–66. [CrossRef]
31. Karim, A.; Azam, S.; Shanmugam, B.; Kannoorpatti, K.; Alazab, M. A Comprehensive Survey for Intelligent Spam Email Detection. *IEEE Access* **2019**, *7*, 168261–168295. [CrossRef]
32. Basit, A.; Zafar, M.; Liu, X.; Javed, A.R.; Jalil, Z.; Kifayat, K. A Comprehensive Survey of AI-Enabled Phishing Attacks Detection Techniques. *Telecommun. Syst.* **2021**, *76*, 139–154. [CrossRef]
33. Srivastava, M.D.; Sachin, S.; Sharma, S.; Tyagi, U. Smart traffic control system using. *Int. J. Innov. Res. Sci. Eng. Technol.* **2012**, *1*, 169–172.
34. Fleetwood, J. Public Health, Ethics, and Autonomous Vehicles. *Am. J. Public Health* **2017**, *107*, 532–537. [CrossRef]
35. Star, S.L. The Ethnography of Infrastructure. *Am. Behav. Sci.* **1999**, *43*, 377–391. [CrossRef]
36. Molloy, D. Google's Pixel 6 Processor Brings AI Photo Features. *BBC News*, 2021. Available online: https://www.bbc.com/news/technology-58955304(accessed on 22 March 2022).
37. Erickson, P.; Kartha, S.; Lazarus, M.; Tempest, K. Assessing Carbon Lock-In. *Environ. Res. Lett.* **2015**, *10*, 084023. [CrossRef]
38. OECD. *Energy, Climate Change and Environment: 2014 Insights*; International Energy Agency: Paris, France, 2014.

39. Stewart, J. Self-Driving Cars Use Crazy Amounts of Power, and It's Becoming a Problem. *Wired*, 6 February 2018. Available online: https://www.wired.com/story/self-driving-cars-power-consumption-nvidia-chip/(accessed on 21 October 2021).
40. Lee, J.; Kockelman, K.M. Energy implications of self-driving vehicles. In Proceedings of the 98th Annual Meeting of the Transportation Research Board, Washington, DC, USA, 13–17 January 2019; Available online: https://www.caee.utexas.edu/prof/Kockelman/public_html/TRB19EnergyAndEmissions.pdf (accessed on 22 March 2022).
41. Liu, Z.; Tan, H.; Kuang, X.; Hao, H.; Zhao, F. The Negative Impact of Vehicular Intelligence on Energy Consumption. *J. Adv. Trans.* **2019**, *2019*, e1521928. [CrossRef]
42. Gupta, U.; Kim, Y.G.; Lee, S.; Tse, J.; Lee, H.-H.S.; Wei, G.-Y.; Brooks, D.; Wu, C.-J. Chasing Carbon: The Elusive Environmental Footprint of Computing. In Proceedings of the 2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA), Seoul, Korea, 27 February–3 March 2021; pp. 854–867.
43. Shi, W.; Cao, J.; Zhang, Q.; Li, Y.; Xu, L. Edge Computing: Vision and Challenges. *IEEE Internet Things J.* **2016**, *3*, 637–646. [CrossRef]
44. Transforma Insights Edge Computing Set for Rapid Growth, across Both IoT Devices and 'Campus Edge'. Available online: https://transformainsights.com/edge-computing-rapid-growth-iot (accessed on 25 February 2022).
45. Edahbi, M.; Plante, B.; Benzaazoua, M. Environmental Challenges and Identification of the Knowledge Gaps Associated with REE Mine Wastes Management. *J. Clean. Prod.* **2019**, *212*, 1232–1241. [CrossRef]
46. Obringer, R.; Rachunok, B.; Maia-Silva, D.; Arbabzadeh, M.; Nateghi, R.; Madani, K. The Overlooked Environmental Footprint of Increasing Internet Use. *Resour. Conserv. Recycl.* **2021**, *167*, 105389. [CrossRef]
47. The Shift Project. *The Shift Project Lean ICT: Towards Digital Sobriety*; The Shift Project: Paris, France, 2019.
48. Griffiths, S. Why Your Internet Habits Are Not as Clean as You Think. Available online: https://www.bbc.com/future/article/20200305-why-your-internet-habits-are-not-as-clean-as-you-think (accessed on 25 February 2022).
49. Thompson, N.C.; Greenewald, K.; Lee, K.; Manso, G.F. The Computational Limits of Deep Learning. *arXiv* **2020**, arXiv:200705558.
50. Martineau, K. Shrinking Deep Learning's Carbon Footprint. Available online: https://news.mit.edu/2020/shrinking-deep-learning-carbon-footprint-0807 (accessed on 28 February 2022).
51. Heavon, W. Our Weird Behavior during the Pandemic is Messing with AI Models. Available online: https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/ (accessed on 28 February 2022).
52. Forti, V.; Cornelis, P.B.; Kuehr, R.; Bel, G. *The Global E-Waste Monitor 2020: Quantities, Flows and the Circular Economy Potential*; United Nations University (UNU): Geneva, Switzerland; United Nations Institute for Training and Research (UNITAR)—Co-hosted SCYCLE Programme, International Telecommunication Union (ITU): Bonn, Switzerland; International Solid Waste Association (ISWA): Rotterdam, Switzerland, 2020.
53. Capra, M.; Bussolino, B.; Marchisio, A.; Shafique, M.; Masera, G.; Martina, M. An Updated Survey of Efficient Hardware Architectures for Accelerating Deep Convolutional Neural Networks. *Future Internet* **2020**, *12*, 113. [CrossRef]
54. Batra, G.; Jacobson, Z.; Madhav, S.; Queirolo, A.; Santhanam, N. *Artificial-Intelligence Hardware: New Opportunities for Semiconductor Companies*; McKinsey & Company: Hong Kong, China, 2018.
55. Robbins, S. A Misdirected Principle with a Catch: Explicability for AI. *Minds Mach.* **2019**, *29*, 495–514. [CrossRef]
56. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef]
57. Robbins, S.; Henschke, A. The Value of Transparency: Bulk Data and Authoritarianism. *Surveill. Soc.* **2017**, *15*, 582–589. [CrossRef]
58. McKay, C. Predicting Risk in Criminal Procedure: Actuarial Tools, Algorithms, AI and Judicial Decision-Making. *Curr. Issues Crim. Justice* **2020**, *32*, 22–39. [CrossRef]
59. van Wynsberghe, A.; Robbins, S. Critiquing the Reasons for Making Artificial Moral Agents. *Sci. Eng. Ethics* **2019**, *25*, 719–735. [CrossRef]
60. Wachter, S.; Mittelstadt, B.; Russell, C. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harv. J. Law Technol.* **2017**, *31*, 841. [CrossRef]
61. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. [CrossRef] [PubMed]

sustainability

MDPI

Article

The Ethics of AI-Powered Climate Nudging—How Much AI Should We Use to Save the Planet?

Marius Bartmann

German Reference Centre for Ethics in the Life Sciences (DRZE), University of Bonn, Bonner Talweg 57, 53113 Bonn, Germany; bartmann@uni-bonn.de

Abstract: The number of areas in which artificial intelligence (AI) technology is being employed increases continually, and climate change is no exception. There are already growing efforts to encourage people to engage more actively in sustainable environmental behavior, so-called "green nudging". Nudging in general is a widespread policymaking tool designed to influence people's behavior while preserving their freedom of choice. Given the enormous challenges humanity is facing in fighting climate change, the question naturally arises: Why not combine the power of AI and the effectiveness of nudging to get people to behave in more climate-friendly ways? However, nudging has been highly controversial from the very beginning because critics fear it undermines autonomy and democracy. In this article I investigate the ethics of AI-powered climate nudging and address the question whether implementing corresponding policies may represent hidden and unacceptable costs of AI in the form of a substantive damage to autonomy and democracy. I will argue that, although there are perfectly legitimate concerns and objections against certain forms of nudging, AI-powered climate nudging can be ethically permissible under certain conditions, namely if the nudging practice takes the form of what I will call "self-governance".

Keywords: sustainability; climate change; artificial intelligence; nudging; digital nudging; libertarian paternalism; autonomy; intergenerational justice

Citation: Bartmann, M. The Ethics of AI-Powered Climate Nudging—How Much AI Should We Use to Save the Planet? *Sustainability* **2022**, *14*, 5153. https://doi.org/10.3390/su14095153

Academic Editors: Aimee van Wynsberghe, Larissa Bolte, Jamila Nachid and Tijs Vandemeulebroucke

Received: 28 February 2022
Accepted: 22 April 2022
Published: 24 April 2022

1. Introduction

The number of areas in which artificial intelligence (AI) technology is being employed increases continually, and climate change is no exception [1]. Several challenges posed by AI have been discussed extensively over the past years—privacy, bias, opacity, to name but a few examples [2]. More recently, the environmental impact of AI itself has also come into focus, for example, the high energy consumption needed to train and run algorithms [3,4]. This research contributes to the insight that we must be wary not to indulge in a questionable form of "technological solutionism" [4] (p. 71) that views AI as a panacea for all kinds of problems. Thus, we must carefully balance the costs and benefits of AI's employment, with particular scrutiny of the ethical challenges involved [5,6].

There are already growing efforts to encourage people to engage more actively in sustainable environmental behavior, so-called "green nudging", such as certifying consumer products with eco-labels and providing households with peer comparisons to improve energy conservation [7]. Nudging in general is a widespread policymaking tool designed to influence people's behavior while preserving their freedom of choice [8–10]. With people spending more time in the digital sphere and making more decisions online, nudging has also become widespread in digital environments [11–13]. Big data and AI are being used in the service of "Big Nudging" to steer people's choices [14,15]. Given the enormous challenges humanity is facing in fighting the severe, harmful and possibly irreversible effects of climate change [16], the question naturally arises: Why not combine the power of AI and the effectiveness of nudging to get people to behave in more climate-friendly ways? Climate nudging powered by AI may thus suggest itself as a suitable strategy to

change people's behavior so that their decisions contribute to a cleaner, safer, and more sustainable planet.

However, nudging has been highly controversial from the very beginning [17]. Advocates praise it as an effective means for making people's lives better. Critics object that nudging compromises people's autonomy by interfering with their capacity to make their own choices. Employing AI to nudge people is no less controversial, even if it is being done with the well-intentioned effort to prevent further harmful climate change. Critics reject AI-powered nudging as large-scale paternalism, which not only disrespects people's autonomy but may also lead to authoritarian societies in which a digital "Green Leviathan" [15] or "wise king" [14] manipulates our lives behind our backs.

In this article I investigate the ethics of AI-powered climate nudging and address the question whether implementing corresponding policies may represent hidden and unacceptable costs of AI in the form of a substantive damage to autonomy and democracy. I will argue that, although there are perfectly legitimate concerns and objections against certain forms of nudging, AI-powered climate nudging can be ethically permissible under certain conditions. To this end, I first elaborate on nudging in general, its background, and rationale (Section 2). In a second step, I review the main argument for nudging as well as the issue of autonomy as the main ethical concern critics have raised (Section 3). Third, I briefly present relevant facts about climate change and the specific ethical challenges they raise (Section 4). Finally, I consider AI-powered climate nudging and discuss whether the main ethical concerns revolving around autonomy also apply in this context. I argue that AI-powered climate nudging can be ethically permissible if the nudging practice takes the form of what I will call "self-governance" (Section 5).

2. Libertarian Paternalism and Nudging

In their highly influential 2008 book *Nudge*, Richard Thaler and Cass Sunstein develop a policymaking approach they call *libertarian paternalism*:

> We strive to design policies that maintain or increase freedom of choice. When we use the term *libertarian* to modify the word *paternalism*, we simply mean liberty-preserving. And when we say liberty-preserving, we really mean it. Libertarian paternalists want to make it easy for people to go their own way; they do not want to burden those who want to exercise their freedom. The paternalistic aspect lies in the claim that it is legitimate for choice architects to try to influence people's behavior in order to make their lives longer, healthier, and better. [8] (p. 5)

Early on, commentators have noted that the definition of paternalism given by Thaler and Sunstein deviates significantly from standard accounts [18] (pp. 126–130). Compare their definition with a standard definition of paternalism:

> Paternalism is the interference of a state or an individual with another person, *against their will*, and defended or motivated by a claim that the person interfered with will be better off or protected from harm. [19] (emphasis added)

In both definitions, increasing an agent's individual welfare represents the primary aim of the interference and serves at the same time as a justification for it. The difference consists in the means employed to achieve this aim. According to the standard account of paternalism, interferences to increase individual welfare infringes on the liberty or autonomy of the targeted person in some way or other. More often than not, the infringement consists in altering or restricting the space of options among which people can choose, for example, banning smoking in public buildings or prescribing motorcyclists to wear helmets [19]. Libertarian paternalism, on the other hand, purports to increase individual welfare while respecting people's liberty and autonomy by preserving freedom of choice. Libertarian paternalism leaves the space of options intact, that is, the interference consists not in *what* options are made available but rather in *how* the options are presented. A standard example is a cafeteria where salads, fruits, vegetables, and other healthy food items are deliberately displayed at eye level so that visitors are more likely to choose what is

better for them [8] (pp. 1–4). This is precisely what Thaler and Sunstein call a "nudge", the notion at the heart of libertarian paternalism. It is defined in the following way:

> A nudge, as we will use the term, is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid. Nudges are not mandates. Putting fruit at eye level counts as a nudge. Banning junk food does not. [8] (p. 6)

The term "choice architecture" refers to the space of options within which people make choices, and nudges are intentional modifications of this space to influence people's choices for their own benefit in a way that does not interfere with their freedom of choice. The number of options is supposed to remain constant, only the presentation of options is changed. A key feature of standard paternalism frequently raising ethical concerns—interfering with people's liberty or autonomy in some form—is thus declared absent from libertarian paternalism.

Before elaborating on nudging and choice architecture in more detail, I want to emphasize right from the outset a rather general point often highlighted by the authors themselves. Sunstein calls this general point the "trap of abstraction" [10] (p. 424). The point is plain and simple but important: nudges can assume a vast and diverse variety of forms; they involve different ends, different means, and different justifications. Hence, a proper ethical assessment of nudges should always consider their characteristic features and the specific circumstances of their implementation.

To get an impression of the heterogeneity of nudges, consider the following examples given by Sunstein himself [10] (p. 424). A nudge can consist in merely providing consumers with information or warnings on products (a GPS device, nutritional information on food items, or graphic images on cigarette packages), in reminders for unpaid bills, in rearranging items in the supermarket to increase their salience, or in changing the default from opt-in to opt-out in the context of enrollment in savings or retirement plans. Paradigmatically, nudges like this aim at people's individual welfare, they are intended to promote people's own ends and to make their lives better. Unsurprisingly, that is why the issue of paternalism has taken center stage in the nudging debate. Yet many nudges are not paternalistic in that they rather aim at social welfare or at protecting the environment, for example, campaigns and programs to reduce energy consumption and greenhouse gas (GHG) emissions [8] (chapter 12). Since the purpose of nudges benefitting the environment is to correct market failures, Sunstein later on expressly distinguishes between "paternalistic nudges" and "market failure nudges" [10] (pp. 426–427).

The upshot of the general point is this: the ends, means, and justifications constituting the respective nudges can differ substantially, therefore the ethical questions involved and the ethical assessment required to decide them may be just as different. The lesson to be drawn is that rather than trying to arrive at a general verdict as to the ethical permissibility of nudging per se we are better off proceeding in a somewhat piecemeal fashion and discuss relevant ethical principles always with an eye towards specific nudging practices.

Now, what is the rationale for libertarian paternalism and nudging in the first place? Libertarian paternalism is based on a certain picture of human cognition. More specifically, it is rooted in a particular understanding of the mechanisms underlying decision-making processes. Thaler and Sunstein largely draw on behavioral science research, especially on the works of Daniel Kahneman [8] (chapter 1). According to Kahneman, the human mind contains two cognitive systems that process information in significantly different ways. System 1 "operates automatically and quickly", whereas System 2 "allocates attention to the effortful mental activities that demand it, including computations" [20] (pp. 20–21). System 1 is in charge of activities such as sensing a particular mood in a voice, driving on an empty highway, or processing simple sentences; System 2 activities require concentration and focus, such as following the clowns in a circus show, trying to remember a particular sound, or doing the taxes [20] (pp. 21–22). Although both systems are fallible, System 1 is particularly susceptible to what is nowadays known under the umbrella term "heuristics

and biases" [20] (part II). Among the many lapses and blunders of System 1 Thaler and Sunstein point out are, for example, overconfidence, loss aversion, and framing [8] (chapter 1). Take one of the most famous experiments illustrating the framing fallacy. Physicians were given information about a certain medical procedure. Those who read the description "The one-month survival rate is 90%" were much more likely to decide in favor of the procedure than those physicians who read the description "There is 10% mortality in the first month", even though both descriptions are logically equivalent [20] (p. 367). Examples like these and the fallacies associated with System 1 abound.

Thaler and Sunstein conclude from the ubiquitous shortcomings of System 1 that the *homo economicus* promulgated by many economists is an illusion. Choices and decisions are simply not always the outcome of fully rational, fully informed, and strong-willed individuals who always act in accordance with their best interests:

> The false assumption is that almost all people, almost all of the time, make choices that are in their best interest or at the very least are better than the choices that would be made by someone else. We claim that this assumption is false–indeed obviously false. [8] (p. 8)

For example, they point to the scientifically corroborated link between obesity and the increased risk of several medical conditions on the one hand, and the high obesity rate in the U.S. on the other to cast doubt on the idea that all U.S. citizens keep to an ideal diet [8] (p. 7). So why not take advantage of the weaknesses of System 1, so the reasoning goes, and harness the power of nudges to get people to choose what is better for them? The food items in the cafeteria must be arranged somehow—why not make a virtue of necessity and design the choice architecture in a way that helps people achieve their ends (a healthy lifestyle) and still preserves their freedom (they can choose a less healthy alternative if they want)?

3. Nudgers and Their Critics

Roughly, there are three main arguments for nudging. The first argument simply builds on the effectiveness of nudging (this is largely uncontroversial, and I will not dispute it here); the second argument maintains that nudging preserves freedom of choice and autonomy (this is very controversial, and I will discuss it in the context of climate nudging in Section 5) [8] (p. 252). Here, I want to review briefly the third argument because it is presented by Sunstein and Thaler as one of their central arguments and also sheds light on a core notion of libertarian paternalism: choice architecture.

The argument is that choice architecture is inevitable, therefore nudging is permissible ([8] (p. 237)), ([9] (p. 14)), ([10] (pp. 415, 420–422)). Just as homo economicus was an illusion, so was the neutrality of the space of options within which people make choices. Indeed, this claim seems hard to challenge. The food items in the cafeteria must be arranged somehow, and the information about mortality and survival rates of medical procedures must be worded in some way. Choice architecture cannot but influence people in one way or other regardless of whether it is the result of deliberate design or coincidence. This also applied to government regulation. In many cases, public officials cannot avoid acting as choice architects in policymaking. This was particularly true when it came to defaults, for example in organ transplantation. *Some system* has to be put in place (for example, opt-in or opt-out), and whichever system is chosen will have consequences for people. Since choice architecture is thus inevitable, nudging is permissible.

It's unclear to me whether the conclusion follows from the premise. Even if the premise is true, and people are influenced by choice architecture regardless of whether it is the result of deliberate design or coincidence, does it really follow that nudging is ethically permissible per se? All that seems to follow is that the specific design of a choice architecture matters because every difference in the space of options potentially makes a difference for people's choices. However, this does not give nudgers carte blanche to interfere with choice architecture. On the contrary, the inevitability of choice architecture rather *increases* the responsibility of policy makers *precisely because* choice architecture always influences

people in some way, regardless of whether it is the result of deliberate design or coincidence. According to the nudgers' own premise, choice architecture influences people either way, and thus policy makers are responsible even in case they decline to interfere with it in some area. If anything, then, the inevitability of choice architecture entails that there is a burden of justification for both interference and non-interference with choice architecture. In the context of government regulation, for example, the inevitability of choice architecture means that every governmental action—as well as every inaction—has consequences for people's lives and must therefore be justified. Whether a particular nudging practice is defensible, on the other hand, is a further question requiring careful ethical assessment of the characteristic features and the specific circumstances of its implementation.

I will now turn briefly to the main argument against nudging. The main ethical concerns critics have with nudging revolves around autonomy [21]. Broadly speaking, threats to autonomy in the context of paternalism can arise in two ways. Either paternalism interferes with the ends people set for themselves, or it interferes with the means people employ to achieve their ends, which is reflected in the common distinction between ends paternalism and means paternalism [9] (p. 19). Ends paternalism is seen as problematic because it imposes ends on people that are not necessarily their own. This leads to a mismatch between people's choices and what they actually want, which undermines their autonomy. When people are nudged in a certain direction, strictly speaking the choices they make are not their own, their "actions reflect the tactics of the choice architect rather than exclusively their own evaluation of alternatives" [18] (p. 128).

In reply, many libertarian paternalists claim that, contrary to standard forms of paternalism, nudging represents a form of the weaker means paternalism because it neither imposes ends on people nor questions the ends people have. Rather, nudging is intended to help people realize the ends they already set for themselves [10] (p. 433). Libertarian paternalists thus turn the tables on their critics and argue that nudging even promotes people's autonomy. For example, the autonomy of someone who wants to eat healthy but is tempted by less healthy options is actually enhanced, and not impaired, because the nudge only helped achieve an end the agent had anyways. Since nudges must be easy to avoid by definition, it is unlikely they would lead to changes of mind in staunch meat-eaters, which libertarian paternalists agree would be problematic because it would interfere with people's ends. In any case, they argue, autonomy is either preserved (no one is coerced, and freedom of choice is secured) or even promoted (the weak-willed are supported in realizing their ends).

However, means paternalism has also been met with criticism. Even if people's ends are respected, interfering with the means people employ to achieve them nevertheless represents a form of manipulation by "*bypassing their capacity for reason*" [22] (p. 5), which again is seen as undermining autonomy. As elaborated on in Section 2, nudges primarily target System 1 and thus exploit psychological vulnerabilities and faulty reasoning, such as inertia and framing. In doing so, critics argue, choice architects would not take people seriously as rational agents. Rather, they would take advantage of their cognitive weaknesses, even if they did so for their own benefit. For choices to be genuinely autonomous, critics insist, people not only have to be in control of setting their ends but also in control of the means and processes to realize those ends ([18] (p. 128)), ([23] (p. 209)).

In reply, some proponents of nudging complain that being fully in control of both ends *and* means represented too high a bar for choices to be autonomous and relied on an implausible conception of rational agency ([7] (p. 337)), ([21] (pp. 145–146)). Since choice architecture is inevitable, people are influenced one way or the other anyways, regardless of whether they are nudged or not. It would be an illusion to think that there could be purely rational processes free of any external influences. Thus, proponents argue, as long as people's ends are promoted or at least left intact, interfering with decision-making processes does not undermine autonomy.

The arguments for and against nudging are still subject to ongoing debate. As noted, in order to decide whether a given nudging practice is ethically problematic, its specific

features and the circumstances of its implementation must be taken into account. In the following section, I will therefore present relevant facts about climate change and the specific ethical challenges they raise, after which I will consider whether the objections from autonomy also apply to AI-powered nudges intended to induce more climate-friendly behavior.

4. Climate Change as an Ethical Challenge

Over the past decades, overwhelming scientific evidence has been gathered to substantiate the thesis that the current climate change—in particular global warming, rising sea levels, and increased frequency of extreme weather events—is caused primarily by anthropogenic GHG emissions [16]. Emitting GHGs of such magnitude has grave and long-lasting effects on the global climate system and will increase the probability of "irreversible impacts for people and ecosystems" [24] (p. 8). Negative effects of climate change outweigh the positive effects by far. Adaptation measures—adjustments to the adverse effects of climate change—are necessary in any case, but without substantial mitigation efforts—GHG emission reduction—the risk of harmful effects of more frequent and more intense climate and weather events will rise significantly [24] (pp. 18–19).

What is distinctive about the atmosphere is that it "comes closest to being a pure public good in that GHGs released anywhere have similar effects, making it a common as well as an essential resource" [25] (p. 79). While an essential resource, it is also finite in that its use without harmful environmental consequences is limited. Many ethicists thus consider climate change primarily a problem of justice, in particular a problem of intergenerational justice and distributive justice [25–27]. In essence, climate change as an ethical problem of intergenerational and distributive justice revolves around the question what present generations owe future generations [28] (chapter 1). The intergenerational aspect is due to climate change being a time-delayed phenomenon in that its effects extend far into the future because most GHGs have a very long lifetime in the atmosphere [24] (p. 87). Therefore, climate policy today will inevitably affect future generations. The distributive aspect is due to the fact that climate change brings with it an unequal distribution of burdens and benefits, which immediately raises questions such as how the cost of mitigation policies and the rights to emit GHGs are to be distributed fairly. For example, for a two-thirds chance of limiting global warming to 1.5 °C by 2050 the remaining carbon budget is roughly 420 $GtCO_2$ [29] (p. 12). If this goal is to be reached, then the atmosphere becomes a finite resource raising the question of just allocation of rights to use it.

The complexity of climate change forms a unique ethical challenge. Two aspects, in particular, pose extraordinary difficulties for an adequate response. First, the response necessary to fight climate change can be understood as a collective action problem [30]. Collective action problems often involve a multitude of agents who have an interest in using collective resources but are disincentivized to pay their fair share because of the possibility to benefit from the resources without carrying any burdens (what is also called "free riding"). Applied to the problem of climate change, this means all agents have an interest in using the atmosphere—through emission of GHGs—but individually they are disincentivized to contribute to the costs because from an individual perspective it is in their interest to free ride on the emission reduction efforts of others. Therefore, responses to climate change are often considered a variant of the tragedy of the commons, more specifically a prisoner's dilemma with a collective resource, in which it is collectively rational to reduce GHG emissions but not individually so [31] (p. 89). Second, GHG emissions are "externalities and are the biggest market failure the world has seen" [32] (p. 39). The costs of emissions in the form of harmful climate change are not fully paid by those who are causing them, but rather transferred to future generations. Thaler and Sunstein concede that in the face of market failures even libertarians think some form of government intervention may prove necessary [8] (p. 184). Examples are taxes on GHG emissions or cap-and-trade systems [8] (pp. 185–188). However, although Thaler and Sunstein do not reject such incentive-based approaches, they believe this approach should be supported with a nudging practice to reduce emissions because incentives like

taxes are often unpopular and therefore difficult for policy makers to implement. Instead, they argue, the hidden costs of emissions should be made visible to nudge people into action. To this end, they proposed that the government should devise a "Greenhouse Gas Inventory" in which the emissions by the biggest emitters are documented. This is supposed to raise public awareness, increase the pressure to act, and lead to more emissions reduction efforts [8] (p. 191).

The need for legal regulation becomes particularly apparent in view of the fact that large portions of global GHG emissions can be attributed to a comparatively small number of industrial companies, on which citizens have only limited influence [33]. Critics of nudging sometimes suggest that nudging practices are a bad substitute for structural reform, but proponents point out that there is no reason why we cannot do both [17]. I would thus agree with Sunstein and Thaler that market failures such as GHG emissions have to be primarily addressed by appropriate legislation but can also be accompanied by suitable nudging practices provided they are implemented in an ethically responsible way.

5. AI-Powered Climate Nudging

Against this backdrop, and given the opportunities of AI technology, one may ask: why not use green nudging as proposed by Thaler and Sunstein and combine it with AI in the effort to reduce GHG emissions? AI technology is already being used to fight climate change. For example, AI for Good, a non-profit organization, promotes using AI to advance the UN's Sustainable Development Goals (SDGs), among which is also "Climate Action" (SDG 13) [34]. In addition, Capgemini, a research institute, found that AI can contribute to combatting climate change when employed, for example, by companies to reduce emissions, improve energy efficiency, and optimize waste management [35].

In a recent book, Mark Coeckelbergh devises a thought experiment and envisions a "green brave new world" in which climate-related policy decisions are delegated to an AI-powered "Green Leviathan" because humanity did not manage to fight climate change by itself [15] (pp. 1–2). This, of course, represents a highly undesirable Huxleyan dystopia. Yet I think Coeckelbergh is perfectly right in highlighting the underlying problem he deals with in his book, namely "the problem of freedom in the light of climate change and AI" [15] (p. 5). In this context, he also explores AI-powered climate nudging, a seemingly freedom-preserving alternative to the authoritarian Green Leviathan:

> One could imagine that nudging is used for changing individual behavior in a more environmentally and climate-friendly direction. [. . .] And maybe AI, having analyzed the data of entire populations or even the entire world, could give us statistical information about our collective carbon footprints and communicate this information in a way that has similar effects on us. Not just by providing information as such, and not by persuasion by means of rational arguments, but by working with human biases and emotions. In this way, nobody is forced to do the right thing, as an authoritarian regime would do; instead, people are 'gently' pushed in a direction. But they can always opt out, they can always make other choices. It seems that freedom is preserved. [15] (pp. 36–37)

Eventually, however, Coeckelbergh rejects AI-powered climate nudging as a form of paternalism infantilizing people because the government would treat its citizens as irrational children incapable of doing the right thing. Although libertarian paternalists insisted on promoting people's own ends "in practice someone else judges for them: the nudger" [15] (p. 38). Coeckelbergh further argues that exploiting cognitive weaknesses–System 1 vulnerabilities—represents a form of unacceptable manipulation undermining people's autonomy by disrespecting their rational capacities [15] (p. 39). He concludes that the tactics employed by choice architects revealed a profound distrust in people since they operated on the assumption "that humans are weak-willed or irrational, and do not always know what is good for them" [15] (p. 41).

Coeckelbergh's critique echoes some aspects of the arguments against nudging touched on in Section 3. Additionally, the employment of AI technology to nudge people into more

climate-friendly behavior exacerbates the ethical concern with autonomy because of the large-scale effects it may produce. Arranging food items in a cafeteria is a nudging practice with quite local effects, but in a digital sphere with millions of users, nudges could have a rather pervasive impact. For this scenario to be plausible we do not need to imagine a Green Leviathan. There is already evidence how character traits can be derived from digital footprints for effective mass persuasion [36]. For example, just think of the infamous Facebook experiment in which the news feeds of almost 690,000 people were manipulated [37].

The potentially large-scale effects of AI-powered nudging add the societal dimension to threats against autonomy. Not only the autonomy of specific (groups of) individuals is endangered, but also the autonomy of society as a whole, its collective autonomy to determine societal ends and the means to pursue them. This means AI-powered nudging may pose a threat to democracy itself. Objections against nudging based on a concern for individual autonomy can thus also be raised out of worries about collective autonomy since in both cases the self-determination of ends and means is interfered with. In Section 3 I elaborated on two ways in which individual autonomy can be undermined, either by interfering with people's ends (ends paternalism) or with the means people employ to realize their already existing ends (means paternalism). These two ways of violating people's autonomy can also occur on the societal level. With regard to ends paternalism, critics fear that a "data-empowered 'wise king'" would be in a position "to produce desired economic and social outcomes almost as if with a digital magic wand", which could ultimately lead to a "top-down controlled society" [14]. In this scenario, people would no longer govern themselves through democratic processes but rather be controlled by a quasi-totalitarian regime imposing its ends on the citizenry. With regard to means paternalism, critics have argued that just as individual autonomy is undermined by exploiting cognitive weaknesses because it bypasses people's capacity for reason, in the same way the collective autonomy of a democratic society is undermined if choice architects "bypass public debate and opt for psychological manipulation instead" [38]. In this scenario, even if people's ends are respected they are still manipulated because hidden influence is exerted to interfere with the means they use for achieving their ends.

These are all legitimate concerns and objections, but do they also apply to the case of climate nudging? As emphasized in Section 2, the specific ends, means and justifications involved in a particular nudging practice must be taken into account to determine its ethical permissibility. In addition, there seem to be certain differences between climate nudging and more standard or conventional forms of nudging such as the cafeteria example.

First, the distinctive characteristic of paternalistic nudges consists in their aim to increase an agent's individual welfare. Yet this is not the case where nudges aim at getting people to behave in more climate-friendly ways. Rather, climate nudging aims at protecting the environment and future generations from harm caused by excessive GHG emissions generated in the present. As pointed out, climate change is a time-delayed phenomenon, therefore cutting emissions now would particularly benefit future generations and not (only) the people at which the nudges contributing to the reduction are aimed. As I also pointed out, emissions are an externality and thus represent a type of market failure. Drawing on the distinction in Section 2, climate nudges—just as green nudges in general [7] (p. 331)—can be categorized as market failure nudges rather than paternalistic nudges.

Second, the difference regarding the aim of climate nudging (protecting the environment and future generations) compared to paternalistic nudging (improving individual welfare) also opens up the possibility of a different justification. What critics of paternalism generally take issue with is that the interference is purported to be justified with reference to the presumption of a third party to know better what is good for individuals than individuals themselves. This presumption is indeed problematic because in order to know what is best for an individual a third party would have to know the individual's personal preferences. However, who is in a better position to know one's personal preferences than oneself? That is why many critics of paternalism draw on Mill's famous no-harm principle, according to which the only condition under which interference with people's

liberty is legitimate "is to prevent harm to others. His own good, either physical or moral, is not sufficient warrant" [39] (p. 80). However, as pointed out in the previous section, the atmosphere is an inherently public good. Determining whether certain activities harm the environment in the form of adverse effects of climate change does not necessarily require taking into account the personal preferences of particular individuals, or at least they are less relevant. Rather, one needs to look primarily at scientific research, and here the jury is in: overuse of the atmosphere's capacity to absorb emissions will have harmful effects for the environment and future generations [16]. It is therefore not uncommon in the debate over moral obligations regarding climate change to appeal to the no-harm principle [40] (p. 218). Accordingly, it could also be used to justify climate nudging because it would be done to protect others from harm—future generations—and not because the government presumes to know what is better for particular individuals.

The third ethical issue concerns the means employed by climate nudging. Assuming that protecting the environment and future generations from harm is a legitimate end, would climate nudging then be justified? Here I agree with critics that for choices to be genuinely autonomous a mere focus on ends is insufficient. In my view, a central discomfort underlying ethical qualms about paternalism in general and nudging in particular is the structural asymmetry between nudgers and nudgees. *Someone else* occupies an allegedly superior vantage point and interferes with one's choices and decision-making processes. Even if this third party had our best interests at heart, exploiting our cognitive weaknesses to further our interests would still disrespect autonomy. While it may be true that we should have a realistic understanding of the inner workings of our cognitive functions and acknowledge that our choices can also be influenced on a subconscious level in some way, exploiting these influences with non-transparent manipulation techniques seems not the right way to deal with those weaknesses.

Given all these considerations, is it possible to implement climate nudging practices that avoid undermining autonomy and democracy? I think climate nudging can be ethically permissible if it takes the form of *self-governance*. Nudging as self-governance comprises at least the following three conditions, which have to be met in order for a climate nudging practice implemented by policy makers to be ethically permissible. These conditions are best thought of as interrelated aspects of a single overall ethical assessment rather than isolated boxes that could be checked completely independently from one another:

- *Symmetry Condition*: nudgers and nudgees should be at least structurally identical, that is, those groups of individuals or their representatives initiating or approving a particular nudging practice should be the same groups of individuals potentially affected by this practice.
- *Democracy Condition*: a policy implementing a particular nudging practice should possess a democratic mandate in some form, that is, the implementation of a nudging practice should require a procedure of public debate and approval.
- *Transparency Condition*: a particular nudging practice should be implemented in a way so that, in principle, everyone can identify the practice and learn about its mechanisms.

The point of a nudging practice taking the form of self-governance that satisfies these three conditions is that a society implementing such a nudging practice would effectively nudge itself in a self-determined, democratically legitimate, and transparent way. Ethical problems associated with asymmetry and manipulation can be avoided because people's ends are respected and their means are not exploited. Take, for example, the cafeteria scenario again. Even if people's ends are respected and promoted, some critics still object to putting healthy food items at eye level because it interfered with people's decision-making process in a non-transparent and therefore manipulative way. But imagine the cafeteria was a school cafeteria and all parties involved—for example, students, parents, teachers, etc.—decided together to make the menu healthier and implemented in a transparent way a nudging practice and accompanied it with an information campaign. I think in this case there would be much less occasion for ethical concerns. In the following, I present and discuss an example of a green nudging practice and of an AI-powered climate nudging

practice, from the public sector and from the private sector, respectively, that may serve to illustrate what I have called self-governance.

In the late 1980s, an environmental initiative was founded in the German town Schönau [41]. It proposed to take over the local power grid and energy provider, and after campaigns and public debate they put it to a vote. The proposal was accepted, the energy provider became the standard utility and adopted a "green default", meaning most of the energy comes from renewable sources. As a consequence, customers are provided with green energy unless they opt out. By 2006, almost all Schönau households used green energy.

In this example, a default—a standard nudging tool—was implemented after public debate and a subsequent vote, thus satisfying all three conditions I characterized above. Many critics of nudging take issue with defaults because they make use of a System 1 psychological vulnerability (inertia) and thus represent a form of manipulation. However, I think this is not the case in this example. For one thing, the nudge is transparent and does not operate behind people's back. Assuming that people do in fact stick to a non-green energy default out of inertia, in the Schönau case this System 1 psychological vulnerability was not dealt with as a weakness to be exploited surreptitiously, but rather as an issue to be addressed openly. For another, the nudge was democratically implemented by the people potentially affected so that there is no asymmetry between nudgers and nudgees, as would have been the case if a government had implemented the policy in a top-down manner. Finally, if you oppose the green default, for whatever reason, you still can opt out, so freedom of choice is also preserved. As a result, the citizens of Schönau basically nudge themselves in a self-determined, democratically legitimate, and transparent way.

A concrete example for AI-powered climate nudging from the private sector is a service offered by Google. Google now provides users of Google Flights with carbon emissions information so users can include the carbon footprint of different flights into their decision-making process and thus may consciously choose alternatives with lower emissions [42]. Search results for flights display in a prominent way estimates of how much kg of CO_2 is specific for a particular flight. Additionally, flights are marked with a "green badge" if they are associated with much less emissions compared to the amount of emissions typical for this route, and also display the amount of emissions saved by choosing this alternative. A further feature is that if for a particular route a train connection is available, then the train connection will also be listed among the results together with the carbon emissions information.

In my view, including carbon emissions information in the search results for flights constitutes a nudging practice, in particular the "green badge" marking flights with significantly less emissions. This type of providing information relevant for climate-friendly (or at least climate-friendlier) behavior can be considered an example of eco-labelling, which in turn is a staple of green nudging [7] (p. 332). One of the mechanisms underlying eco-labelling is the so-called "salience bias", which can be operative in real-world as well as digital environments [43] (p. 7). According to the salience bias, people tend to focus on aspects of their environment that stand out in some way or other. The classic cafeteria scenario in which healthy food items are made salient by putting them at eye level is a case in point. Likewise, making search results salient by marking them with a "green badge" thus seems to qualify as a green nudge, in particular a climate nudge, because this is intended to steer people's choices towards more climate-friendly options.

Now, does this nudging practice satisfy the three conditions I characterized above, and can it therefore be considered an ethically permissible nudging practice assuming the form of self-governance? First of all, since the nudging practice is not part of a policymaking tool but a service from the private sector, the democracy condition does not really apply. However, consumers are free to use the service or not so there need not be a public procedure of approval. Yet this also means that the satisfaction of the other two conditions is even more important. The symmetry condition requires that consumers not be manipulated by a third party, but rather knowingly and voluntarily nudge themselves into climate-friendly behavior by using the service. Here one can see how the conditions are interrelated, because

for the symmetry condition to be fulfilled the transparency condition must simultaneously be fulfilled. I can only nudge myself if I know that I am participating in a corresponding practice. This seems to be the case here because the service is transparent about carbon emissions information being a relevant factor in displaying the search results.

Nevertheless, some critics of nudging argue that exploiting the salience bias is manipulative on the grounds that the salience of an item and its actual importance may come apart. Taking this critique into account, Robert Noggle argues that "a salience nudge is not manipulative if it influences choice by bringing the salience of some fact into closer alignment with its actual importance" [44] (p. 168). Of course, it is often difficult to determine the importance of some fact, in particular when people's personal preferences are involved. Again, as I argued above, what the nudging practice in this case aims at is not the personal welfare of individual agents but an inherently public good—the atmosphere. Additionally, there is no doubt that the carbon emissions produced by air travel contribute to climate change. Making carbon emissions salient thus brings them in alignment with the important role they play in contributing to climate change. Ethicists often point out that one of the problems in dealing with climate change consists in insufficient awareness of the harmful consequences of activities involving emissions because they are quite literally invisible, and their adverse effects are distributed spatially and temporally [31] (p. 88). Drawing attention to these effects and making them visible to raise awareness about the consequences of our actions seems not to be manipulative. Since there is no restriction on the set of options, freedom of choice also seems to be preserved, and it is always possible to opt out of the service entirely.

Of course, what is primarily necessary is structural reform aiming at decarbonization. However, the fight against the harmful consequences of climate change also faces the problem of "institutional inadequacy" [31] (p. 89) because enforceable sanctions required for limiting GHG emissions are difficult to implement on a global level. Thus, to reduce carbon emissions there seems to be no reason why we should not also engage in effective and ethically permissible climate nudging while pursuing structural reform. In sum, if I nudge myself into taking the train instead of a domestic flight because I am made aware in a non-manipulative way of the significant amount of emissions a flight for the same route would cause, then this seems to be a genuinely autonomous choice and not a case of problematic manipulation.

6. Conclusions

Whereas many nudging practices are in fact ethically problematic, I maintain AI-powered climate nudging can be ethically permissible if it takes the form of self-governance satisfying the symmetry condition, the democracy condition, and the transparency condition. A society implementing corresponding policies would nudge itself and therefore avoid the asymmetry between nudgers and nudgees as well as the danger of manipulation asymmetry involves. Of course, the Green Leviathan is definitively not a role model for solving the climate crisis with AI technology. The harm to autonomy and democracy would represent an unacceptable damage, even if motivated by good intentions. The gold standard of policymaking should always be rational persuasion, and information campaigns should not be replaced by nudging practices. However, if the ethical assessment of a particular climate nudging practice powered by AI takes the form of self-governance, then the power of AI can be harnessed in an ethically justifiable way as a supporting measure to fight the adverse effects of climate change.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Rolnick, D.; Donti, P.L.; Kaack, L.H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A.S.; Milojevic-Dupont, N.; Jacques, N.; Waldman-Brown, A.; et al. Tackling climate change with machine learning. *arXiv* **2019**, arXiv:1906.05433. [CrossRef]
2. Mittelstadt, B.D.; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data Soc.* **2016**, *3*, 2053951716679679. [CrossRef]
3. van Wynsberghe, A. Sustainable AI: AI for sustainability and the sustainability of AI. *AI Ethics* **2021**, *1*, 213–218. [CrossRef]
4. Coeckelbergh, M. AI for climate: Freedom, justice, and other ethical and political challenges. *AI Ethics* **2021**, *1*, 67–72. [CrossRef]
5. Floridi, F.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.* **2018**, *28*, 698–707. [CrossRef] [PubMed]
6. van Wynsberghe, A. *Artificial Intelligence. From Ethics to Policy*; Study, Panel for the Future of Science and Technology, European Parliamentary Research Service (EPRS), Scientific Foresight Unit (STOA); European Union: Brussels, Belgium, 2020. Available online: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641507/EPRS_STU(2020)641507_EN.pdf (accessed on 7 April 2022).
7. Schubert, C. Green Nudges: Do they work? Are they ethical? *Ecol. Econ.* **2017**, *132*, 329–342. [CrossRef]
8. Thaler, R.H.; Sunstein, C.R. *Nudge. Improving Decisions about Health Wealth, and Happiness*; Yale University Press: New Haven, CT, USA; London, UK, 2008.
9. Sunstein, C.R. *Why Nudge? The Politics of Libertarian Paternalism*; Yale University Press: New Haven, CT, USA; London, UK, 2014.
10. Sunstein, C.R. The Ethics of Nudging. *Yale J. Regul.* **2015**, *32*, 413–450. [CrossRef]
11. Weinmann, M.; Schneider, C.; vom Brocke, J. Digital Nudging. *Bus. Inf. Syst. Eng.* **2016**, *58*, 433–436. [CrossRef]
12. Mirsch, T.; Lehrer, C.; Jung, R. Digital Nudging: Altering User Behavior in Digital Environments. In Proceedings of the 13th International Conference on Wirtschaftsinformatik, St. Gallen, Switzerland, 12–15 February 2017; pp. 634–648.
13. Yeung, K. 'Hypernudge': Big Data as a Mode of Regulation by design. *Inf. Commun. Soc.* **2017**, *20*, 118–136. [CrossRef]
14. Helbing, D.; Frey, B.S.; Gigerenzer, G.; Hafen, E.; Hagner, M.; Hofstetter, Y.; van den Hoven, J.; Zicari, R.V.; Zwitter, A. *Will Democracy Survive Big Data and Artificial Intelligence?* Scientific American: New York, NY, USA, 2017. Available online: https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/?redirect=1 (accessed on 7 April 2022).
15. Coeckelbergh, M. *Green Leviathan or the Poetics of Political Liberty*; Routledge: New York, NY, USA; London, UK, 2021.
16. IPCC. *Climate Change 2021: The Physical Science Basis*; IPCC: Geneva, Switzerland, 2021. Available online: https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Full_Report.pdf (accessed on 7 April 2022).
17. Schmidt, A.T.; Engelen, B. The ethics of nudging: An overview. *Philos. Compass* **2020**, *15*, e12658. [CrossRef]
18. Hausman, D.M.; Welch, B. Debate: To Nudge or Not to Nudge. *J. Political Philos.* **2010**, *18*, 123–136. [CrossRef]
19. Dworkin, G. Paternalism. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Stanford University: Stanford, CA, USA, 2020. Available online: https://plato.stanford.edu/archives/fall2020/entries/paternalism/ (accessed on 7 April 2022).
20. Kahneman, D. *Thinking, Fast and Slow*; Penguin: London, UK, 2012.
21. Engelen, B.; Nys, T. Nudging and Autonomy: Analyzing and Alleviating the Worries. *Rev. Philos. Psychol.* **2020**, *11*, 137–156. [CrossRef]
22. Blumenthal-Barby, J.S.; Burroughs, S. Seeking Better Health Care Outcomes: The Ethics of Using the "Nudge". *Am. J. Bioeth.* **2012**, *12*, 1–10. [CrossRef]
23. Bovens, L. The Ethics of Nudge. In *Preference Change. Approaches from Philosophy, Economics and Psychology*; Grüne-Yanoff, T., Hansson, S.O., Eds.; Springer: Dordrecht, The Netherlands, 2009; pp. 207–219. [CrossRef]
24. IPCC. *Climate Change 2014: Synthesis Report*; IPCC: Geneva, Switzerland, 2014. Available online: https://www.ipcc.ch/site/assets/uploads/2018/02/SYR_AR5_FINAL_full.pdf (accessed on 7 April 2022).
25. Vanderheiden, S. *Atmospheric Justice*; Oxford University Press: Oxford, UK, 2008.
26. Page, E.A. *Climate Change, Justice and Future Generations*; Edward Elgar: Cheltenham, UK; Northampton, UK, 2006.
27. Caney, S. Climate Change. In *The Oxford Handbook of Distributive Justice*; Olsaretti, S., Ed.; Oxford University Press: Oxford, UK, 2018; pp. 664–688. [CrossRef]
28. Hiskes, R.P. *The Human Right to a Green Future: Environmental Rights and Intergenerational Justice*; Cambridge University Press: Cambridge, UK, 2009.
29. IPCC. *Global Warming of 1.5 °C*; IPCC: Geneva, Switzerland, 2018. Available online: https://www.ipcc.ch/site/assets/uploads/sites/2/2019/06/SR15_Full_Report_Low_Res.pdf (accessed on 7 April 2022).
30. Hayward, T. Climate change and ethics. *Nat. Clim. Chang.* **2012**, *2*, 843–848. [CrossRef]
31. Gardiner, S.M. A Perfect Moral Storm: Climate Change, Intergenerational Ethics, and the Problem of Moral Corruption. In *Climate Ethics. Essential Readings*; Gardiner, S.M., Caney, S., Jamieson, D., Shue, H., Eds.; Oxford University Press: Oxford, UK, 2010; pp. 87–98.
32. Stern, N. The Economics of Climate Change. In *Climate Ethics. Essential Readings*; Gardiner, S.M., Caney, S., Jamieson, D., Shue, H., Eds.; Oxford University Press: Oxford, UK, 2010; pp. 39–76.
33. CDP Carbon Majors Report. 2017. Available online: https://cdn.cdp.net/cdp-production/cms/reports/documents/000/002/327/original/Carbon-Majors-Report-2017.pdf?1501833772 (accessed on 7 April 2022).

34. AI for Good. Available online: https://ai4good.org/ai-for-sdgs/goal-13-climate-action/ (accessed on 7 April 2022).
35. Capgemini Research Institute. Climate AI: How Artificial Intelligence Can Power Your Climate Action Strategy. 2020. Available online: https://www.capgemini.com/de-de/wp-content/uploads/sites/5/2020/11/Report_Climate_AI_Capgemini_Research_Institute.pdf (accessed on 7 April 2022).
36. Matz, S.C.; Netzer, O. Using Big Data as a window into consumers' psychology. *Curr. Opin. Behav. Sci.* **2017**, *18*, 7–12. [CrossRef]
37. Kramer, A.D.I.; Guillory, J.E.; Hancock, J.T. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 8788–8790. [CrossRef] [PubMed]
38. Furedi, F. Defending Moral Autonomy against an Army of Nudgers. Spiked. 2018. Available online: https://www.spiked-online.com/2011/01/20/defending-moral-autonomy-against-an-army-of-nudgers (accessed on 7 April 2022).
39. Mill, J.S. *On Liberty*; Yale University Press: New Haven, CT, USA; London, UK, 2003.
40. Singer, P. Climate Change. In *Practical Ethics*, 3rd ed.; Cambridge University Press: Cambridge, UK, 2011; pp. 216–237.
41. Pichert, D.; Katsikopoulos, K.V. Green defaults: Information presentation and pro-environmental behaviour. *J. Environ. Psychol.* **2008**, *28*, 63–73. [CrossRef]
42. Find Flights with Lower Carbon Emissions. Available online: https://blog.google/products/travel/find-flights-with-lower-carbon-emissions/ (accessed on 7 April 2022).
43. Caraban, A.; Karapanos, E.; Gonçalves, D.; Campos, P. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–15. [CrossRef]
44. Noggle, R. Manipulation, salience, and nudges. *Bioethics* **2018**, *32*, 164–170. [CrossRef] [PubMed]

sustainability

MDPI

Article

Acknowledging Sustainability in the Framework of Ethical Certification for AI

Sergio Genovesi * and Julia Maria Mönig *

Center for Science and Thought, University of Bonn, Poppelsdorfer Allee 28, 53115 Bonn, Germany
* Correspondence: genovesi@uni-bonn.de (S.G.); moenig@uni-bonn.de (J.M.M.)

Abstract: In the past few years, many stakeholders have begun to develop ethical and trustworthiness certification for AI applications. This study furnishes the reader with a discussion of the philosophical arguments that impel the need to include sustainability, in its different forms, among the audit areas of ethical AI certification. We demonstrate how sustainability might be included in two different types of ethical impact assessment: assessment certifying the fulfillment of minimum ethical requirements and what we describe as nuanced assessment. The paper focuses on the European, and especially the German, context, and the development of certification for AI.

Keywords: AI certification; sustainability; ethics of AI

Citation: Genovesi, S.; Mönig, J.M. Acknowledging Sustainability in the Framework of Ethical Certification for AI. *Sustainability* **2022**, *14*, 4157. https://doi.org/10.3390/su14074157

Academic Editors: Aimee van Wynsberghe, Tijs Vandemeulebroucke and Marc A. Rosen

Received: 21 February 2022
Accepted: 30 March 2022
Published: 31 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Due to growing concerns about ethical, legal, and social issues around AI systems, over the past few years, both private corporations and public institutions have started developing quality and trustworthiness certification for AI [1] (p. 26 f.). In the EU, in April 2021, a proposal for an "Artificial Intelligence Act" was published, which foresees "standards, conformity assessment, certificates [and] registration" as a means to deal with "high-risk AI systems" [2] (Chapter 5, Art. 6). Although the route to a standardized and generally accepted certification is still a long one, several actors have laid the groundwork for the development of an assessment of what constitutes trustworthy AI. A High-Level Expert Group (HLEG) on Artificial Intelligence appointed by the European Commission indicated four ethical principles for AI based on fundamental rights and seven key AI requirements [3]. The ethical principles are respect for human autonomy, the prevention of harm, fairness, and explicability. The key requirements are supporting human agency and oversight; technical robustness and safety; respecting privacy and allowing good data governance; transparency; guaranteeing diversity; non-discrimination and fairness; improving societal and environmental well-being; and accountability for the outcomes of AI systems [3,4]. Part of the "social and environmental well-being" requirement is the need for a "sustainable and environmentally friendly AI" [3] (p. 30). In 2018, the European Group on Ethics in Science and New Technologies (EGE) had already identified sustainability as one of nine "ethical principles and democratic prerequisites" for a "shared Ethical Framework for Artificial Intelligence, Robotics and 'Autonomous' Systems," alongside human dignity, autonomy, responsibility, justice, equity and solidarity, democracy, the rule of law and accountability, security, safety, bodily and mental integrity, data protection and privacy. In parallel with the work of the HLEG, many stakeholders published guidelines for the development of trustworthy AI systems identifying, among other things, what ethical and technical minimum requirements should be considered in their development and audited using an AI certification [5–8]. To name but a few, the German Data Ethics Commission listed in a report the indispensable ethical and legal principles that should guide the development of AI systems and their regulation: these being human dignity, self-determination, privacy, safety, democracy, justice and solidarity, and sustainability [9]. The platform "Lernende Systeme" indicated the following minimum requirements for

AI certification: transparency, traceability, verifiability, and accountability; product safety and reliability; avoidance of unintended consequences on other systems, people, and the environment; justice in the sense of equality and non-discrimination; privacy and personal rights protection; allowing human self-determination, guaranteeing transparency about the use of the AI system and the role of the human being in the decision-making process [10]. Similarly, in a white paper by Fraunhofer IAIS, in cooperation with the Universities of Bonn and Cologne, Cremer et al. defined the following minimum requirements as a basis for an audit catalog: respect for social values and laws, human autonomy and control, fairness, transparency, reliability, security, and data protection [11]. An initial suggestion on how to put these requirements into actual practice has been detailed in an inspection catalog [12]. In its "Standardization Roadmap for AI", the German Institute for Standardization (DIN) refers to these requirements as quality criteria for AI products, also noting the contributions by the Data Ethics Commission and the platform "Lernende Systeme" [13]. Remarkably, only the Data Ethics Commission, whose report does not directly focus on the development of certification, indicates that sustainability is a *basic* principle. The white paper by Fraunhofer IAIS et al. does not mention sustainability as an audit area for AI certification and the platform of "Lernende Systeme" states that sustainability might be considered an additional, optional requirement for a "Certification plus"—but not as a minimum requirement [10] (p. 25).

In contrast to the position put forward by "Lernende Systeme", we will argue that assessing sustainability should be a key part of any ethical certification for AI. Since this is a philosophical paper, we deploy the method of conceptual analysis. Addressing the three dimensions of sustainability, in Section 2, we briefly review some of the major issues that AI systems present when considering the environmental, economic, and social impact of system development and use. Starting from the idea of ethical behavior as embodying just and responsible behavior toward other human and non-human beings, in Section 3, we show that sustainability is at root an ethical issue, since it involves responsibility toward other human beings and the environment, and is required to guarantee international, intergenerational, and interspecies justice. Based on this, in Section 4, we highlight the relevance of a sustainability audit in the context of ethical certification for AI and suggest two audit methods that could be used in the process of a certification: a "minimum requirements" checklist demanding the fulfillment of specific prerequisites, and a "nuanced assessment" attributing a score to evaluate the performance of a system in a given audit area. In conclusion, we call to action the stakeholders responsible for the development of ethical certification of AI, to implement AI sustainability.

2. AI and the Three Dimensions of Sustainability

Sustainability is defined differently by different actors depending on their aims and fields of interest. One famous definition of sustainability, or, more precisely, of "sustainable development", is often quoted from the Brundtland Report, also known as "Our Common Future": "Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs" [14] (p. 41). Regarding AI, it became clear that this ability addresses all three so-called "pillars of sustainability" [5] (p. 395), [15], namely, the environmental, economic, and social dimensions of sustainability. "Environmental sustainability" generally refers to the impact of our actions on planet Earth. To be environmentally sustainable, human development should identify planetary boundaries that must not be transgressed and work to prevent unacceptable environmental change [16]. "Economic sustainability" refers to "practices that support long-term economic growth without negatively impacting [the] social, environmental, and cultural aspects of the community" [17]. Finally, "social sustainability" includes, among other things, "achieving a fair degree of social homogeneity, equitable income distribution, employment that allows the creation of decent livelihoods, and equitable access to resources and social services" [18], as well as "[encouraging] communities to promote social interaction and [fostering] community investment while respecting social

diversity" [19]. In the following, we will list some of the major issues concerning AI in these three domains.

(1) Environmental Sustainability is not only important regarding those CO_2 emissions caused by the electricity needed for computing operations, but also concerning the whole life cycle of products. This includes—but is not limited to—the production of the very hardware needed to run AI and software in general, using, for instance, plastics, metal, and raw materials. Considering the whole life-cycle does also include recycling and re-use processes (e.g., addressing the premature obsolescence of hardware and software by designing them in a "technically sustainable" way [20] or unifying electric chargers for smartphones to reduce attendant electronic waste [21,22]). The complex issues related to environmental sustainability can be illustrated by the example of electric vehicles and the common narrative that asserts that they will solve many problems related to CO_2 emissions. This narrative might be somewhat misleading since it does not take into account those other environmental costs involved in the production of electric vehicles [23]. Indeed, to assess the sustainability of a product, its whole "material footprint" should be considered [24]. In addition, the rhetoric about "cloud computing" contributes to cloaking the fact that physical computers are performing the computing operations. This is part of the overall claim that the use of AI technologies should not only prevent negative outcomes for the environment but also, in a broader sense, be "favorable to the environment" [25].

(2) Economic Sustainability. As with the case of environmental sustainability, many economic sustainability issues result from the very hardware production process for AI systems. Indeed, a type of new colonialist exploitation of those populations who live near raw material extractions sites can be observed and reported. Returning to the already noted case of electric vehicles, the production of car batteries can raise serious ethical issues, e.g., when the cobalt mining for the batteries is being done by children (and adults) working in conditions of slavery and without adequate safety measures [26–28]. Likewise, extracting and processing the necessary materials used to build hardware raises important sustainability issues concerning health, working conditions, and the environmental and resource exploitation of many populations in developing countries, directly contributing to inequality between the global north and global south. To address this problem, it has been claimed that we need a decolonizing engagement to fight institutionalized oppression [29].

(3) Social Sustainability. Regarding the real-world applications of AI systems, biased data and a lack of a diversity-oriented perspective in the designing, developing, testing, launching, and post-marketing phases of a given product might lead companies to release software that discriminates against minorities and vulnerable social groups, replicating racist and sexist biases in application fields such as risk scoring in justice [30] or credit scoring for mortgage lending [31]. These algorithmic-fairness issues affect the social sustainability of a product directly since they facilitate the spread of inequality and social conflicts, resulting in compromising societal well-being.

To address social, economic, and environmental fairness at large, several approaches show how "AI for social good" [32] and "AI for sustainability" [15] should be used to foster the achievement of the UN's Sustainable Development Goals through a value-sensitive design [33] aiming, among other things, at "reducing inequality within and among countries" (Goal 10) and achieving gender equality (Goal 5) [34,35]. These three levels of sustainability lay the groundwork for the investigation of sustainability as an ethical topic and, therefore, shall be considered in the ultimate assessment of the ethical implications of trustworthy AI. It should be remarked that these dimensions are tightly interwoven in actual real-world scenarios. The sustainability assessment in the framework of a given certification should, therefore, focus on those concrete sustainability issues pertaining to cases of specific use, and these issues might encompass different dimensions all at the same time.

3. Sustainability and the Ethics of Responsibility

Today, it is possible for us to understand that our economic behavior has material consequences on a global scale. In 2020, this evidence was amplified, for instance, by the shortage of many products during the first lockdown of the coronavirus pandemic, highlighting how heavily many societies depend on outsourced work in the globalized world economy [36,37]. How goods are designed and produced, what goods we purchase, how long we use them for, and how we dispose of them can positively or negatively affect the environment and other humans—and their rights—around the world, even though the causal connections might be neither straightforward nor directly visible. At the same time, the increasing number of catastrophic climatic events over the past few years shows the impact that human behavior, and especially mass consumption, has on our environment in a way that cannot be ignored anymore.

Even though the negative consequences felt on a global scale produced by individual behaviors might not be caused intentionally and might "just" be the result of shortsighted and profit-oriented conduct, sustainability issues raise questions of (in-) justice. Justice, in a philosophical sense at least, can be understood as respect for others, as the struggle to ensure equal rights and preserve human dignity, and as the will not to harm others through violence or subjugation. In moral philosophy, it is generally considered unjust and wrong to conceive of and treat others as a mere means to one's own ends, and to not see them at the same time as an end in themselves [38] (p. 428), to reduce otherness to the totality of one's own limited and limiting representation of it, ignoring the fact that the other infinitely exceeds this representation because of their own complexity and freedom [39] (Chapters I.C. and III.B.), or to deny recognition of their identity, values, and rights [40]. Behaviors that are unsustainable on an environmental and/or social level evidence this lack of consideration toward others. For example, there is an unfair distribution in bearing the cost of pollution since only relatively few enjoy the benefits of polluting production processes and activities, but everyone in the world is, in some way and to some degree, affected by them [41,42]. In this sense, those polluting the most overlook other people's needs, suffering, and discomfort and focus solely on their own advantages. Similarly, exploited workers in countries to which production is outsourced are looked upon merely as means if no thought is given to their working conditions. Accordingly, in a global ethical framework, the undeniable evidence of the impact of consumers' and producers' actions makes them not only causally co-responsible for climatic and humanitarian disasters but also morally accountable for the injustice caused by their economic behavior.

Ignoring the consequences of one's actions necessarily implies ignoring the central capacity of modern humanity: to plan one's actions and to assess the possible consequences and future risks for other human beings and their environment. This is also stressed in the above mentioned Brundtland Report: "Humanity has the ability to make development sustainable to ensure that it meets the needs of the present without compromising the ability of future generations to meet their own needs" [14] (p. 16). In 1979, Hans Jonas highlighted that while "new" challenges do come with the development of "modern" technology, previous generations had neither the knowledge nor the power to take the potential future outcomes of their immediate actions into account. Acting ethically was, therefore, synchronous, considered to only affect humans directly surrounding the actor, and responsibility was backward-looking [43,44]. However, even if from today's point of view, current developments look more complex, human action itself has been thought to produce unforeseeable outcomes, as, for instance, Hannah Arendt argues. According to her, this had at least three consequences: (a) in politics people tried to "substitute making for acting" and behavior for action to control other humans; (b) as a remedy for the "irreversibility" of one's actions the "power to forgive" was suggested; and (c) the "power of promise" was supposed to deal with the unpredictability of actions [45] (p. 220 ff.). This forward-looking aspect of responsibility, thus, adds up to the retrospective liability for one's actions. For Hannah Arendt, the limits of human responsibility are related to human

"plurality", the fact that all humans are born into a world that has already been inhabited and shaped by other human beings [45] (p. 234).

Despite its limitations, the assessment of the long-term consequences of unsustainable behavior should, according to Jonas and the Brundtland Report, include some consideration of the issue intergenerational justice: if the enjoyment of goods today will cause harm to and limit the freedom of the next generations, and equally those who have already been born and those who will inhabit Earth in the future [42,46,47] and this is also unfair behavior that distributes the environmental and social costs of the actions of fewer people unequally [48]. Hans Jonas sees the reason for this also in the anthropological argument that he adds to the temporal dimensions mentioned above. According to him, we should not only think about other human beings of future generations but also about humankind as a whole and whether we consider it desirable that humans keep on living (on Earth). Destroying our planet logically might mean destroying humanity's habitat and, hence, stripping future generations of the chance to exist at all:

"[. . .] we are, strictly speaking, not responsible to the future human individuals but to the idea of Man, which is such that it demands the presence of its embodiment in the world. [. . .] It is this ontological imperative, emanating from the idea of Man, that stands behind the prohibition of a va-banque gamble with mankind. Only the idea of Man, by telling us why there should be men, tells us also how they should be" [43] (p. 43).

This becomes even more relevant when we consider that, since the twentieth century, humans have been able to destroy not only what they have made, as they have always been able to do, but also, with the invention of the atomic bomb, they even have the capability to destroy what they have not made: nature, all species, and the whole planet [45] (p. 3), [49] (p. 6). Likewise, climate change can be seen as a slow process of destroying life forms and things that humans did not create. In addition to taking future generations into account when reflecting on the possible impact of our behavior, we, therefore, also need to consider its consequences for other species, especially given the complex relationships between biodiversity, nutrition, habitat conservation, etc. [50]. What is even more striking is that, for decades now, the global north has more than it needs while the global south is—still—being exploited. Human beings are starving while others are wasting food, water, and other resources. By virtue of our duty to our fellow human beings, the environment and future generations, sustainability as an aware and responsible practice should nowadays be a top ethical priority for a globalized society. The evidence that unsustainable behavior results in harm and injustice, and is, therefore, unethical, cannot be ignored anymore. Assessing that we can do, invent, or develop something is not enough of an argument to say we should do it—as, for instance, the so-called technological imperative [49] (p. 7) or Silicon Valley's mantra "Move fast and break things" claim [51] (p. 60). Instead, we need regulations to guide businesses in the sustainable development of new technologies, and the instruments to empower consumers to make responsible choices.

4. Sustainability as an Audit Area for an Ethical Certification of AI

We argue that sustainability, as an ethical issue, should be considered when certifying ethical and trustworthy AI. More specifically, auditing the environmental, economic, and social sustainability of an AI system should be one of the core requirements of an ethical assessment, and not just an option [10] (p. 28). Moreover, sustainability as a core requirement can be seen as matching at least two of the abovementioned requirements for the development of trustworthy AI, namely "Diversity, Non-Discrimination, Fairness," and "Societal and Environmental Well-Being" [4] (pp. 15–20).

The first step in assessing and rating the fulfillment of an ethical requirement is to identify concrete ethical risks that are specific to a particular field of application through expert and stakeholder consultation. In the European context, this is considered an essential procedure in proposals for the development of an ethical impact assessment, among others by the CEN Workshop Agreement for an Ethical Impact Assessment Framework [52]. This allows the translation into practice of ethical goals that otherwise would remain

simply abstract and, therefore, non-auditable. For instance, the ethical implications of a creditworthiness-scoring algorithm and for an AI-powered customer assistance chatbot are different when it comes to ensuring social fairness. In the first case, the algorithm should not (directly or indirectly) discriminate against people based on ethnicity, gender, nationality, or any other category by attributing a lower score to an individual belonging to specific groups than to other individuals with a similar profile. In the second case, the algorithm should not output offensive language and should not discriminate against any social group by perpetrating racist, sexist, homophobic, or other stereotypes. Moreover, if the chatbot uses voice recognition, people with non-native or regional accents and people with speech impairments should be able to communicate with the machine in the same way as people whose pronunciation is considered "standard".

Once the concrete risks concerning sustainability and other ethical issues of AI products in a specific field of application have been identified, an effective way to operationalize the results of the risk assessment in the framework of an audit process and to state the ethical acceptability of an AI system in a particular use case scenario is to set specific minimum requirements, which should be met to avoid unethical consequences, such as, in the case of sustainability, the unnecessary waste of resources or social discrimination. In the German framework, the "minimum requirements approach" is suggested as a basis for the certification of AI systems by different developers [10,11]. As well as the fulfillment of minimum requirements, a nuanced assessment might be a useful resource to audit the abovementioned sustainability issues. The reason is that, once a threshold for ethical acceptability has been determined, different software may perform differently within the acceptability domain. Therefore, the specific function of this more fine-grained level of audit is to provide different stakeholders, such as producers, consumers, governments with a common tool to compare similar products. Nevertheless, if a minimum requirement is not satisfied, the product should be classified as being unsustainable, irrespective of how well the product performs in the nuanced assessment of other features.

Sustainability, therefore, could and should have a direct impact on an ethical assessment in at least two ways. Primarily, adding environmental, economic, and social sustainability to the minimum ethical requirements of an AI application in the form of concrete, domain-specific goals to be fulfilled will prevent unsustainable products from being certified as ethical in the first place. As the CEN Workshop Agreement suggests, in light of the complexity of the process, the defined threshold criteria to be met should be carried out by a multidisciplinary board of experts and stakeholders [52] (pp. 19–21). Indeed, concerning sustainability, the definition of specific threshold values might be particularly difficult in those cases in which an integrated consideration of different dimensions of sustainability is required. Moreover, in the case of a nuanced assessment, the attribution of an audit-area-specific score showing the (expected) performance of a product in the domains of environmental, economic, and social sustainability, would affect the choice of those consumers valuing sustainability and will increase developers' attention toward these audit areas. However, it should be remarked that similar metrics, as accurate as they may be, are just proxies and should not be mistaken for sustainability as a moral and societal goal. Indeed, this goal might be missed if businesses excessively focus on quantitative proxy measures [53]—e.g., by neglecting other important issues, by misallocating funds or, in a worst case scenario, by cheating.

Nuanced assessments already exist in the field of environmental sustainability, for example in assessing the energy performance of household appliances, and there is already at least one attempt to produce a similar "Care Label" certification suite for Machine Learning, labeling not only energy consumption but also other features such as runtime, memory usage, expressivity, usability, and the reliability of the AI software [54,55]. This kind of assessment would be an excellent tool to audit the mentioned features separately and optimize AI systems to achieve a better balance of the performances in the different audit areas. To accomplish this, the assessment could be embedded into a more general ethical framework. An attempt to unify aspects of the three pillars of sustainability (environment,

society, and economy) in a unique, comprehensive sustainability index, is being carried out by the project "SustAIn", funded by the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection. The SustAIn team proposed sets of criteria, indicators, and operationalizable sub-indicators for the evaluation of AI systems' sustainability [56] (pp. 57–64). Among the sub-indicators for social sustainability, they list the level of discrimination potential based on an impact assessment, the proportion of a company's AI systems that use methods to measure fairness and bias, and the diversity of the developer team, measured by the represented percentage of gender, age group, and ethnicity group. Among the sub-indicators for economic sustainability, the evaluation of working conditions throughout the entire production chain is mentioned. Finally, environmental sustainability sub-indicators include, among others, the measurement of energy consumption and direct CO_2 emissions, the percentage of recycled material in the hardware, and the recommendation of environmentally sustainable alternatives by automated decision making (ADM) and recommender systems. Because of their operationalizability, such sustainability sub-indicators could be easily integrated into a certification process either in the definition of the minimum requirements or as relevant indicators for a nuanced assessment. In the first case, a use-case-specific threshold value should be agreed on for the chosen sub-indicator. Not exceeding (or falling below, according to the case) the threshold value should then be taken as a minimum requirement. In the second case, the level of performance above the acceptance threshold of a given AI system could be showcased and rated. We suggest that this kind of integration is essential for the full development of a comprehensive ethical assessment.

It should be stressed that, due to the constant evolution of society and technology, no single certification can guarantee conformity with ethical standards indefinitely. On the one hand, concrete ethical requirements will need to be continually adapted to absorb new research findings from different disciplines and societal dynamics [57]. On the other hand, the development of more advanced technology will create new application scenarios and new ethical challenges. This problem directly concerns the so-called "Collingridge dilemma," according to which it is impossible to predict the impact of a new technology until the given technology is fully developed and deployed [58]. Therefore, ethical assessments should have a de facto expiry date and the continued conformity of a product with the ethical standards of society should be revisited periodically.

Certifying that AI systems are compliant with periodically updated ethical standards would allow us to acknowledge the achievement of increasingly more challenging sustainable and other ethical goals. Indeed, the advancement of technology should be valued not only from a technical point of view but also from a moral perspective. Technology should help translate into reality those values that ethical reflection recognizes as indispensable for future life and well-being in society, such as respecting human rights, protecting the environment, and distributing resources and opportunities fairly. These values should not remain abstract, and it is possible to measure their gradual achievement. Among other factors, the larger our CO_2 emissions, the larger our raw material exploitation and waste production will be in turn, and the further society will be from climate justice. The further exploitative practices to produce goods are spread and minorities are discriminated against, the further society will be from global justice. These trends are reversible. Striving to achieve ethical goals through the improvement of technology and its regulation can be defined as moral progress [59]. An ethical certification aims to foster moral progress by providing consumers and producers with a clear assessment of a product's compliance with these ethical goals.

5. Conclusions

Global ethics of responsibility need a broad picture of the moral community [60] (p. 119). Human beings should be considered moral actors, while the group addressed by moral actions should be even broader than humanity, taking into account the environment at large. It is possible to outline different, coexisting dimensions of ethical responsibility.

First, there is an international, global dimension: people all around the world might bear the consequences of our actions. Moral actors should consider this. Moreover, intergenerational justice should also be considered since future generations will be affected by our current actions and decisions. Finally, especially when considering environmental sustainability, we should address an interspecies dimension: in a globalized society, consumerism is affecting lives and ecosystems all around the world. A rising number of environmentalists are claiming that respect for life and dignity should not be granted only for humans, and destroying ecosystems causes the suffering and impoverishment of life quality for those living beings who survive, directly impacting their freedom and dignity [60] (p. 111), as well as all species' livelihoods.

Fortunately, moral awareness about the ecological and social impacts of globalization and consumerism is rising fast and the urgency of achieving the UN's sustainability goals need new, institutionalized tools to motivate people to act ethically and treat fellow humans, other species, and our planet with respect. Together with sustainability-oriented regulations, a certification for AI software could be effective in motivating consumers to use sustainable products and seek further information about the impact of the products they are using. Furthermore, if the fulfillment of the certified minimum ethical and technical requirements to commercialize a product is made mandatory by law, governments could use certifications to ensure that the AI systems in circulation are sustainable. None of this can be done through lip service. While the European Commission is beating an important path, notably by fostering Responsible Research and Innovation (RRI) [61], pure ethical and conduct codes for enterprises such as the Corporate Digital Initiative Action [62] or strategies for Corporate Social responsibility (CRS) and Responsible Business Conduct (RBC) [63] are not enough in themselves. We need action.

Author Contributions: The authors contributed equally to this paper. All authors have read and agreed to the published version of the manuscript.

Funding: Our research is funded by MWIDE NRW in the framework of the project "Zertifizierte KI" ("Certified AI"); funding number 005-2011-0050.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. DKE/DIN. *Ethik und Künstliche Intelligenz. Was Können Technische Normen und Standards Leisten?* DIN: Berlin, Germany, 2020.
2. European Commission. *Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*; European Commission: Brussels, Belgium, 2021.
3. HLEG on AI. Ethics Guidelines for Trustworthy AI. 8 April 2019. Available online: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed on 29 March 2022).
4. HLEG on AI. The Assessment List for Trustworthy Artificial Intelligence. 17 July 2020. Available online: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (accessed on 29 March 2022).
5. Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
6. Hagendorff, T. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds Mach.* **2020**, *30*, 99–120. [CrossRef]
7. Algorithm Watch. AI Ethics Guidelines Global Inventory. Available online: https://inventory.algorithmwatch.org/ (accessed on 29 March 2022).
8. Zicari, R.V.; Brodersen, J.; Brusseau, J.; Düdder, B.; Eichhorn, T.; Ivanov, T.; Kararigas, G.; Kringen, P.; McCullough, M.; Möslein, F.; et al. Z-Inspection®: A Process to Assess Trustworthy AI. *IEEE Trans. Technol. Soc.* **2021**, *2*, 83–97. [CrossRef]
9. Datenethikkommission. Gutachten der Datenethikkommission. 2019. Available online: https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=6 (accessed on 29 March 2022).
10. Heesen, J.; Müller-Quade, J.; Wrobel, S. *Zertifizierung von KI-Systemen—Kompass für die Entwicklung und Anwendung Vertrauenswürdiger KI-Systeme*; Lernende Systeme: München, Germany, 2020.

11. Cremers, A.; Englander, A.; Gabriel, M.; Hecker, D.; Mock, M.; Poretschkin, M.; Rosenzweig, J.; Rostalski, F.; Sicking, J.; Volmer, J.; et al. Trustworthy Use of Artificial Intelligence. Priorities from a Philosophical, Ethical, Legal, and Technological Viewpoint as a Basis for Certification of Artificial Intelligence. 2019. Available online: https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_Thrustworthy_AI.pdf (accessed on 29 March 2022).
12. Poretschkin, M.; Schmitz, A.; Akila, M.; Adilova, L.; Becker, D.; Cremers, A.B.; Hecker, D.; Houben, S.; Mock, M.; Rosenzweig, J.; et al. Leitfaden zur Gestaltung Vertrauenswürdiger Künstlicher Intelligenz. 2021. Available online: https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf (accessed on 29 March 2022).
13. Wahlster, W.; Winterhalter, C. *Deutsche Normungsroadmap. Künstliche Intelligenz*; DIN: Berlin, Germany, 2020.
14. World Commission on Environment and Development or Brundtland Commission: Our Common Future. Brundtland Report. 1987. Available online: https://sustainabledevelopment.un.org/content/documents/5987our-common-future.pdf (accessed on 29 March 2022).
15. van Wynsberghe, A. Sustainable AI: AI for sustainability and the sustainability of AI. *AI Ethics* **2021**, *1*, 213–218. [CrossRef]
16. Rockström, J.; Steffen, W.; Noone, K.; Persson, A.; Chapin, F.S.; Lambin, E.F.; Lenton, T.M.; Scheffer, M.; Folke, C.; Schellnhuber, H.J.; et al. A safe operating space for humanity. *Nature* **2009**, *461*, 472–475. [CrossRef] [PubMed]
17. University of Mary Washington, Office of Sustainability. Economic Sustainability. Available online: https://sustainability.umw.edu/areas-of-sustainability/economic-sustainability/ (accessed on 31 December 2021).
18. Sachs, I. Social sustainability and whole development: Exploring the dimensions of sustainable development. In *Sustainability and the Social Sciences: A Cross-Disciplinary Approach to Integrating Environmental Considerations into Theoretical Reorientation*; Becker, E., Ed.; Zed Books: London, UK, 1999; ISBN 1856497089.
19. University of Mary Washington, Office of Sustainability. Social Sustainbility. Available online: https://sustainability.umw.edu/areas-of-sustainability/social-sustainability/ (accessed on 31 December 2021).
20. Penzenstadler, B.; Femmer, H. A Generic Model for Sustainability with Process- and Product-Specific Instances. In *Proceedings of the 2013 Workshop on Green in/by Software Engineering*; Association for Computing Machinery: New York, NY, USA, 2013; pp. 3–8, ISBN 9781450318662.
21. European Commission. One Common Charging Solution for All. Available online: https://ec.europa.eu/growth/sectors/electrical-engineering/red-directive/common-charger_en (accessed on 31 December 2021).
22. Fanta, A. *How Apple Lobbied EU to Delay Common Smartphone Charger*; EUobserver: Brussels, Belgium, 2019.
23. Dhara, C.; Singh, V. The Delusion of Infinite Economic Growth. Available online: https://www.scientificamerican.com/article/the-delusion-of-infinite-economic-growth/ (accessed on 29 March 2022).
24. Wiedmann, T.O.; Schandl, H.; Lenzen, M.; Moran, D.; Suh, S.; West, J.; Kanemoto, K. The material footprint of nations. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 6271–6276. [CrossRef] [PubMed]
25. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef] [PubMed]
26. McKie, R. Child Labour, Toxic Leaks: The Price We Could Pay for a Greener Future. Available online: https://www.theguardian.com/environment/2021/jan/03/child-labour-toxic-leaks-the-price-we-could-pay-for-a-greener-future (accessed on 31 December 2021).
27. European Parliament. Answer Given by Ms Urpilainen on Behalf of the European Commission, Question Reference: E-001002/2020. Available online: https://www.europarl.europa.eu/doceo/document/E-9-2020-001002-ASW_EN.html#def3 (accessed on 31 December 2021).
28. Bergmann, R.; Solomun, S. A New AI Lexicon: Sustainability from Tech to Justice: A Call for Environmental Justice in AI. Available online: https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-sustainability-d40fd714d396 (accessed on 31 December 2021).
29. Mohamed, S.; Png, M.-T.; Isaac, W. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philos. Technol.* **2020**, *33*, 659–684. [CrossRef]
30. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine Bias. Available online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed on 29 March 2022).
31. Lee, M.S.A.; Floridi, L. Algorithmic Fairness in Mortgage Lending: From Absolute Conditions to Relational Trade-offs. *Minds Mach.* **2021**, *31*, 165–191. [CrossRef]
32. Floridi, L.; Cowls, J.; King, T.C.; Taddeo, M. How to Design AI for Social Good: Seven Essential Factors. *Sci. Eng. Ethics* **2020**, *26*, 1771–1796. [CrossRef]
33. Umbrello, S.; van de Poel, I. Mapping Value Sensitive Design onto AI for Social Good Principles. In *AI and Ethics*; Springer: Berlin, Germany, 2021; Volume 1.
34. Vinuesa, R.; Azizpour, H.; Leite, I.; Balaam, M.; Dignum, V.; Domisch, S.; Felländer, A.; Langhans, S.D.; Tegmark, M.; Fuso Nerini, F. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nat. Commun.* **2020**, *11*, 233. [CrossRef] [PubMed]
35. Ryan, M.; Antoniou, J.; Brooks, L.; Jiya, T.; Macnish, K.; Stahl, B. The Ethical Balance of Using Smart Information Systems for Promoting the United Nations' Sustainable Development Goals. *Sustainability* **2020**, *12*, 4826. [CrossRef]
36. Gabriel, M. We Need a Metaphysical Pandemic. In *In the Realm of Corona-Normativities: A Momentary Snapshot of a Dynamic Discourse*, 2020th ed.; Gephart, W., Ed.; Vittorio Klostermann: Frankfurt am Main, Germany, 2020; ISBN 9783465145318.

37. Genovesi, S. Support your Local. In *In the Realm of Corona-Normativities: A Momentary Snapshot of a Dynamic Discourse*, 2020th ed.; Gephart, W., Ed.; Vittorio Klostermann: Frankfurt am Main, Germany, 2020; ISBN 9783465145318.
38. Kant, I. *Kritik der Reinen Vernunft (1. Aufl. 1781). Prolegomena. Grundlegung zur Metaphysik der Sitten. Metaphysische Anfangsgründe der Naturwissenschaften, Studienausg, Nachdr. der Ausg. 1968*; de Gruyter: Berlin, Germany, 1978; ISBN 3110014378.
39. Levinas, E. *Totalité et Infini*; Martinus Nijhoff: Den Haag, The Netherlands, 1961.
40. Fraser, N. *Justice Interruptus: Critical Reflections on the "Postsocialist" Condition*; Routledge: New York, NY, USA; London, UK, 2014; ISBN 9781315822174.
41. IPCC. Annex I: Glossary [Matthews, J.B.R. (ed.)]. In *Global Warming of 1.5 °C. An IPCC Special Report on the Impacts of Global Warming of 1.5 °C above Pre-Industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty*; Masson-Delmotte, V., Zhai, P., Pörtner, H.-O., Roberts, D., Skea, J., Shukla, P.R., Pirani, A., Moufouma-Okia, W., Péan, C., Pidcock, R., Eds.; IPCC: Geneva, Switzerland, 2018. Available online: https://www.ipcc.ch/sr15/chapter/glossary/ (accessed on 31 December 2021).
42. Caney, S. *Climate Justice*; Springer: Berlin, Germany, 2020.
43. Jonas, H. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*; Univ. of Chicago Press: Chicago, IL, USA, 1984; ISBN 0226405966.
44. Heidbrink, L.; Langbehn, C.; Loh, J. (Eds.) *Handbuch Verantwortung*; Springer: Berlin, Germany, 2017.
45. Arendt, H. *The Human Condition*; The University of Chicago Press: Chicago, IL, USA, 1998.
46. Caney, S. Justice and Future Generations. *Annu. Rev. Polit. Sci.* **2018**, *21*, 475–493. [CrossRef]
47. *Order of the First Senate of 24 March 2021—1 BvR 2656/18, Paras. 1-270*; Federal Constitutional Court Germany: Karlsruhe, Germany, 2021.
48. Page, E. *Climate Change, Justice and Future Generations*; Reprinted; Edward Elgar: Cheltenham, UK; Northampton, MA, USA, 2007; ISBN 9781847204967.
49. Lenk, H.; Rophol, G. (Eds.) *Technik und Ethik*; Reclam: Stuttgart, Germany, 1993.
50. Robert Garner. *A Theory of Justice for Animals: Animal Rights in a Nonideal World*; Oxford University Press: New York, NY, USA, 2013.
51. Véliz, C. *Privacy is Power*; Bantam Press: London, UK, 2020.
52. CEN/CENELEC. *Ethics Assessment for Research and Innovation—Part 2: Ethical Impact Assessment Framework (SATORI)*; CENELEC: Brussels, Belgium, 2017.
53. Braganza, O. Proxyeconomics, a theory and model of proxy-based competition and cultural evolution. *R. Soc. Open Sci.* **2022**, *9*, 211030. [CrossRef] [PubMed]
54. Morik, K.; Kotthaus, H.; Heppe, L.; Heinrich, D.; Fischer, R.; Mücke, S.; Pauly, A.; Jakobs, M.; Piatkowski, N. Yes We Care!—Certification for Machine Learning Methods through the Care Label Framework. 2021. Available online: http://arxiv.org/pdf/2105.10197v1 (accessed on 29 March 2022).
55. Morik, K.; Kotthaus, H.; Heppe, L.; Heinrich, D.; Fischer, R.; Pauly, A.; Piatkowski, N. The Care Label Concept: A Certification Suite for Trustworthy and Resource-Aware Machine Learning. 2021. Available online: https://arxiv.org/pdf/2106.00512 (accessed on 29 March 2022).
56. Rohde, F.; Wagner, J.; Reinhard, P.; Petschow, U.; Mayer, A.; Voss, M.; Mollen, A. *Nachhaltigkeitskriterien für Künstliche Intelligenz. Entwicklung Eines Kriterien- und Indikatorensets für die Nachhaltigkeitsbewertung von KI-Systemen Entlang des Lebenszyklus*; Schriftenreihe des IÖW 220/21; IÖW: Berlin, Germany, 2021.
57. Rességuier, A.; Rodrigues, R. AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data Soc.* **2020**, *7*, 2053951720942541. [CrossRef]
58. David Collingridge. *The Social Control of Technology*; St. Martin's Press: New York, NY, USA, 1980.
59. Gabriel, M. *Moralischer Fortschritt in dunklen Zeiten*; Ullstein: Berlin, Germany, 2020.
60. Coeckelbergh, M. *Green Leviathan or the Poetics of Political Liberty: Navigating Freedom in the Age of Climate Change and Artificial Intelligence*; Routledge/Taylor & Francis Group: Abingdon, UK, 2021.
61. European Commission. Responsible Research Innovation. Available online: https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation (accessed on 14 January 2022).
62. Corporate Digital Responsibility Initiative. Digitalisation Calls for Responsibility. Available online: https://cdr-initiative.de/ (accessed on 14 January 2022).
63. European Commission. Corporate Social Responsibility & Responsible Business Conduct. Available online: https://ec.europa.eu/growth/industry/sustainability/corporate-social-responsibility_en (accessed on 14 January 2022).

sustainability

MDPI

Article

Mindful Application of Digitalization for Sustainable Development: The Digitainability Assessment Framework

Shivam Gupta * and Jakob Rhyner

Bonn Alliance for Sustainability Research, University of Bonn, D-53113 Bonn, Germany; rhyner@uni-bonn.de
* Correspondence: shivam.gupta@uni-bonn.de

Abstract: Digitalization is widely recognized as a transformative power for sustainable development. Careful alignment of progress made by digitalization with the globally acknowledged Sustainable Development Goals (SDGs) is crucial for inclusive and holistic sustainable development in the digital era. However, limited reference has been made in SDGs about harnessing the opportunities offered by digitalization capabilities. Moreover, research on inhibiting or enabling effects of digitalization considering its multi-faceted interlinkages with the SDGs and their targets is fragmented. There are only limited instances in the literature examining and categorizing the impact of digitalization on sustainable development. To overcome this gap, this paper introduces a new Digitainability Assessment Framework (DAF) for context-aware practical assessment of the impact of the digitalization intervention on the SDGs. The DAF facilitates in-depth assessment of the many diverse technical, social, ethical, and environmental aspects of a digital intervention by systematically examining its impact on the SDG indicators. Our approach draws on and adapts concepts of the Theory of Change (ToC). The DAF should support developers, users as well policymakers by providing a 360-degree perspective on the impact of digital services or products, as well as providing hints for its possible improvement. We demonstrate the application of the DAF with the three test case studies illustrating how it supports in providing a holistic view of the relation between digitalization and SDGs.

Keywords: digitalization; sustainable digitalization; artificial intelligence; sustainable development; SDGs; Assessment Framework; mindful; digital age; digitainability

Citation: Gupta, S.; Rhyner, J. Mindful Application of Digitalization for Sustainable Development: The Digitainability Assessment Framework. *Sustainability* **2022**, *14*, 3114. https://doi.org/10.3390/su14053114

Academic Editors: Tijs Vandemeulebroucke, Aimee van Wynsberghe, Larissa Bolte and Jamila Nachid

Received: 3 February 2022
Accepted: 2 March 2022
Published: 7 March 2022

1. Introduction

The UN Agenda 2030 [1] calls for inclusive action towards sustainable development, covering a broad spectrum of areas ranging from poverty eradication and universal health protection to environmental conservation and peacebuilding. The 17 Sustainable Development Goals (SDGs) and their respective targets, along with particular indicators, require timely monitoring and reporting of the progress in member states of the UN. The frame of 232 indicators developed by the Inter-Agency Expert Group (IAEG) helps measure progress towards the achievement of the SDGs. Accelerating progress towards SDGs is urgently required, augmenting the need for data-driven systemic and context-inclusive approaches to nudge on-ground development. An influential role is played by digitalization, and particularly artificial intelligence (AI) [2]. As Tegmark [3] puts it, "the more intelligent and powerful machines get, the more important it becomes that their goals are aligned with ours", which in our context would rather be the SDGs. Digitalization encapsulates the individual, organizational, and societal transformation triggered by mass adoption of algorithm and data-driven interventions that generate, process, and transfer information [4–6], drawing from a very diverse set of methodologies and technologies [7,8]. It is essential to explore in detail whether and to what extent the opportunities offered by the digitalization interventions (DIs) can be aligned with mindfulness for sustainable development upon particular attention to the achievement of the goals mentioned in the 2030 Agenda [9].

With *mindful* use of digitalization, we are referring to the capability to be aware of where we are in the digitalization development process and how we plan to utilize it

further, considering it in a comprehensive sustainability context. This protects us from being overwhelmed by uncritical optimism on the one hand or by systematic pessimism on the other hand, but rather be more rooted in the explicit contexts for sustainable development. To find the context-driven on-ground function of technology for SDG at the target or indicator level, it is essential that we explicitly comprehend why it is used, what problem it is addressing, who is responsible for action, when is it that the DI offers the desired results and how is it working in reality. The debate about the benefits and risks of utilizing digitalization to support progress towards sustainability has started but remains fragmented [10,11]. Numerous entanglements and contexts remain poorly understood, thus demanding further research towards the mindful use of digitalization.

Several studies have indicated that the DIs collect and utilize diverse amounts of resources, which might not necessarily benefit SDGs as a whole [10–12]. Trade-offs concerning social and environmental sustainability are discussed more broadly [13]. Various frameworks and analytical methods were developed to assess the acceptance and applications of technologies in specific domains and contexts. These tools, qualitative as well as quantitative, can be utilized for predicting the usefulness of technologies across industries and development areas [14]. There also exist some modified versions of the frameworks as mentioned earlier to assess the usability of the technologies under particular conditions incorporating broader sustainability contexts [15]. However, these frameworks are fragmented in their approach to analyzing in isolation considering certain SDG(s) and are domain-specific, thus limiting the understanding of the impacts digitalization offers for SDGs and their indicators [10,12]. This fragmented understanding may lead to gaps in knowledge about the DI's application and an ill-defined prioritization, eroding the holistic aspiration and enactment of the Agenda 2030. In order to transition from the present fragmented knowledge to more holistic evidence about the impact of digitalization for sustainable development , the question of how to systematically assess the impact of DI for SDGs as a whole is the main focus in this article. For this purpose, the paper introduces the Digitainability Assessment Framework (DAF), which maps the impact of a DI on SDG indicators. The inclusive framework garners information about a DI, contexts, process, outcome to reflect the overall impact on the SDGs. The SDG indicator framework [16] offers an internationally agreed structure for the assessment of the sustainability of the DI. Realizing what follows when the DI is brought to practice, e.g., intervention or service is "rolled out" with stakeholders, is vital to address implementation challenges and align it for the advancement of the SDGs. The alignment of the DI with SDGs also renders a more sustainable DI [17]. The DAF should help to guide this alignment.

This paper is organized as follows. In Section 2, three essential dimensions are discussed that need to be taken into account in a comprehensive assessment of the sustainability of the DIs (digitainability assessment), namely, first the synergies and trade-offs between the SDGs, second the context dependency, and third the stakeholder structures. Taking into account these considerations, Section 3 introduces the DAF. The paper then presents the results obtained by the operationalization of the DAF for three test case studies on the diverse DIs in Section 4, namely, spatial optimization for the systematic deployment of citizen-driven air quality monitoring networks, blockchain for healthcare service delivery, and remote sensing-based machine learning approaches for disaster risk management and planning. Finally, the discussion concerning considerations and limitations is presented in Section 5, followed by the conclusions and outlook in Section 6.

2. Critical Dimensions for Digitainability Assessment

Digitalization is generally regarded as an essential element for driving sustainable transformations [18]. There has been rapid adoption of digital technologies as a versatile, complex, and powerful resource capable of performing specific tasks requiring a vast amount of human capacity [19]. Despite several constraints in different circumstances, there is a growing moment in leveraging the DI for addressing matters related to the SDGs and their targets [20]. A couple of recent studies focused on charting the contribution of the

DI for monitoring SDGs indicators within stand-alone "for good" projects in a particular domain [21,22], rather than to also identify the counterproductive effects of digitalization capacities for addressing the complicated challenges of Agenda 2030 [20]. Only a limited effort was devoted to exploring the relationship between the DI and SDGs as a whole [23]. The significance of our work lies in the framework whose purpose is to help in evidencing the impact of DIs in an integrated and consistent manner for SDGs. Further, it helps identify knowledge gaps hindering the mindful use of the power of digitalization for sustainable development.

When assessing the impact of the DI for sustainability, we cannot treat it as a tool or process acting independently of its environment. In this section, we will discuss three specific dimensions, which a comprehensive assessment always needs to consider. First, the intrinsic complexity of the SDG framework with its synergies, but also trade-offs, needs to be taken into account. The DI may support certain indicators while simultaneously being at odds with others. The second one is the context dependency. A given DI may have a different impact in industrialized and in non-industrialized context. The third one is the stakeholder structure. Depending on the intentions and preferences of the stakeholders (providers, users, etc.), the DI may work out differently. In the following, these three essential dimensions are considered in more detail.

2.1. Synergies and Trade-Offs between SDGs

In order to make progress in SDGs, it is crucial to acknowledge the dynamic relations between the indicators of the goals in terms of their potential interactions, both across and within each SDG [24,25]. Progress will significantly depend on utilizing the synergies while addressing the potential trade-offs. Adshead et al. [26] pointed out that the infrastructure intended for delivering SDGs, as well as the potential trade-offs between indicators, are important dimensions to be considered when making investment and policy intervention decisions, asserting the findings of Schroeder et al. [27], especially in the context of the environment, society, and human health.

There is extensive research exploring synergies and trade-offs between SDGs [28–33]. In particular, Pradhan et al. [34] explored the synergies and trade-offs between SDGs indicators in 227 countries, classifying the goals considering their interactions; however, the interactions are identified based on the correlation among indicator level data, not implying causality. Scherer et al. [35] investigated the interactions between SDGs 1, 6, 10, 13, and 15, arguing that social goals usually lead to increase environmental impacts. Mainali et al. [36] explored interaction amid the SDGs 1, 2, 6, and 7 in South Asia and Sub-Saharan Africa context, suggesting that potential synergies and trade-offs among the SDGs vary depending on multiple aspects such as geographical conditions, infrastructure, and the policy measures. Nerini et al. [30] measured the water–energy–food nexus (SDG 2, 6, 12), identifying tensions concerning SDG 7 (for example, energy access), and stressed the need for careful planning of complex energy systems underpinning long term development processes. von Stechow et al. [37] investigated the trade-offs between Climate Action (SDG13) and the rest of SDGs, suggesting that curbing energy demand is crucial across the goals. However, the characterization of how digitalization influences SDG interactions is not well documented yet. Consideration about the interface between social, ethical, environmental, and technical effects of the DI also needs to be understood while approaching sustainability, particularly considering the rebound effects, as already pointed out by Jevons in 1865 when addressing paradoxically increased consumption and offset savings [38]. Nishant et al. [39] also discussed the increase in total consumption of limited resources, even when considering environmentally effective systems.

Thus, it is crucial to understand the cascading impacts caused by digitalization over already discussed trade-offs like negative digital and ecological footprint, environmental impacts, climate crisis, pollution, social tensions, and inequalities together with non-resilient and fragmented practices towards sustainability [40–42]. It is essential to uncover how spillover effects and indirect impacts hinder the realization of several SDGs in the

long- and short term [43], which might not be obvious when taking action. To address these concerns, contextual awareness is a crucial dimension [44–46]. The research mentioned above builds an important basis for the DAF to be introduced in the paper.

2.2. Context Dependency

To understand the impact of the DIs, the awareness of the context is essential for striking the right balance between protecting the essential dimension of sustainability and implementing practices fostering holistic sustainable development. As an example, the impact of innovative technologies and data sources such as satellite images in the context of climate change may have lesser ethical risks than compared to other data sources and domains such as in healthcare and public safety domain, where personal data plays a crucial role [47]. Moreover, the impact of technologies in developed countries vs. developing or low-income countries varies in their level of adoption, local governance priorities, level of acceptance, energy demands, culture, infrastructure, and other particular local aspects [36]. Digitalization might be helpful to amplify scientific discovery and governance towards shedding light on the multifaceted SDGs interlinkages, and their cascading impacts [48–50].

Extant studies do not yet provide or add up to a systematic mapping of the impact of digitalization on SDGs. The relevance of context to demonstrate how or why the specific DI outcomes can be realized and how it might lead to certain synergy and tradeoffs between other targets/indicators of SDGs or limit the generalizability of intervention to different settings or circumstances is fundamental. A DI facilitates us in what we aim to achieve, whereas the context accounts for the specific outcomes [51,52]. The relevance of the context in using the DIs for particular objectives is usually reflected with the help of theories, frameworks, and taxonomies that are useful in exploring the inhibiting and enabling aspects of various potential outcomes [53]. In the literature, methods such as Theory of Change (ToC) [54] and Theory of Acceptance (ToA) [55] seem to cover context as one of determinants. Terms such as "context," "setup," and "environment" were often used synonymously in the reference of technology implementation and related domains [56]. The context-aware assessment about the role of a DI for SDG indicators with the capability to acknowledge the complexities of the SDGs has yet to be developed [57,58], highlighting the need for systemic and context-aware approaches to utilize digitalization mindfully for SDG progress.

2.3. Multi-Stakeholder Structure

Understanding the complex role digitalization can have on the interaction of SDGs, and the context that frames the role it can have for the SDGs, will require multi-stakeholder involvement to systematically exploit the DIs for the progress of the SDGs and beyond. Measuring timely progress on how well the intervention satisfies the key stakeholders of their requirements is vital [59]. A lack of integrated and collective approaches leads to SDG implementation being isolated projects with little or no impact [60]. It is crucial that sustainable development being prioritized and seen as a core value, where trust-building was considered necessary [61]. Krellenberg et al. [62] stress that the SDGs suffer from insufficient ambition and that they compete or overlap with other local context-driven actions. Köhler et al. [63] highlighted the role of the DIs, and policy alone could not facilitate easy transformations to sustainability. Further elements such as coherent strategy, enablers, regulation, and competencies are required [64]. Socio-technical aspects linked to hindrance in a DI adoption, widening digital divide, and lack of trust in machine-driven approaches need to be revisited in the specific contexts for sustainable development with the DIs [40]. To identify further the potential enablers, the role of grassroots and civil society is critical but explored sparsely [65].

Insights concerning the DIs acceptance and creating public awareness for initiating collective actions towards SDGs are also little [66]. Consideration of social factors such as local context, reluctance in technology application, and awareness is crucial for inclusive action [41]. Recent literature recognizes the relevance of encouraging multi-stakeholders

participation and collaboration among key actors to support pathways of sustainable development with digitalization [67–69]. However, the unclear roles and responsibilities and silo-based inertia lead to a lack of integrity in the implementation process for SDGs [68,70]. Existing implementation practices are limited in the transformative power and strategic abilities [60], which hinders the overall implication of digitalization for SDGs. Literature also stresses the lack of tools and guides for reporting and imbalanced alignment among corporate strategies, policies, and scientific outlook, leading to counterproductive outcomes for digitalization and sustainable development [71,72]. The available literature, although still fragmented, shows the importance of the consideration of the stakeholder structure in the assessment.

The synergy and trade-off interrelations between the SDGs, the context dependency, and the stakeholder structure form three critical dimensions for a thorough assessment of the impact of digitalization on sustainable development. Fortunately, as shown in this section, a considerable body of research is already available. It will put us in a position to formulate the Digitainability Assessment Framework (DAF) in the next section.

3. Digitainability Assessment Framework (DAF)

The central idea of the DAF is to systematically examine the impact of a DI on the indicators underpinning the SDGs. In doing so, the DAF draws from and adapts concepts of Theory of Change (ToC), which represent a mapping of causal pathways between changes that have taken place and the activities the transformation processes undertake, track changes, and demonstrate impacts [73,74]. It is important to mention at this point that the ToC is referring to the impact on ("change" for) the SDG indicators and not on possible improvements ("change") of the DI. In this sense, the scope of the DAF is primarily an assessment of the impact of DI on sustainable development and not on the improvement of the DI. While the assessment often may give valuable hints for improvement, its systematic investigation and realization are beyond the scope of the DAF.

The ToC as an approach in itself is not a unique defined theory but rather a set of rules allowing for goal-oriented planning, suitable for identifying and simplifying the complexity, reflecting on the enabling and inhibitory potential change process brought by the DIs. ToC concepts are widely recognized as a practice-oriented approach and is not a fixed methodology. It allows flexibility to work according to the needs. However, there are consensus about the basic elements that constitute ToC (discussed in the following section). Application of ToC as a methodology has been successfully used for testing and validation of two long-standing areas: program theories and development practices [75]. Considering the flexibility ToC offers to different circumstances and adaptable to the different contexts, we believe ToC could guide the assessment of DI considering the planning and implementation activities. Therefore, in the DAF, we are embedding ToC elements to combine practitioner's perspectives not for the evaluation/improvement of DI in itself but rather for the assessment of the impact of the DI on SDGs with the scientifically backed evidence [76,77]. Recently, Li and Thomas [78] also proposed the application of ToC as a methodology and a process to measure the impact of ICT technologies; however, their focus was not on sustainable development.

Typically, the ToC approach is composed of five key parts: Inputs, Activities, Outputs, Outcomes, and Impact [79]. They allow for a combination of quantitative and qualitative criteria for the explanation of the change process [80], and could capture multiple aspects of change, including social, governance, and security perspectives [81]. In applying ToC, appropriate boundaries, scope, and level of complexity are essential [82]. We adapt the ToC concept with a specific focus on the SDGs and a DI for the DAF development. Notably, the adaptations are made in the DAF to capture the ways how the causal linkages are considered between the role of DI and SDG indicators anticipating the particular context and impact types. These adaptations also help capture the multi-level changes brought by a DI, which could be mapped—as the "impact pathway"—presenting each impact in

logical connection to all the relevant SDG indicators and the plausible convergence of crucial information.

We map the ToC parts into the DAF segments, as follows:

- Digitalization Intervention (containing ToC parts Input and Activities).
- Purpose (Outputs and Outcomes).
- Impact (including desired as well as undesired impact).

The three DAF segments are described in detail in the following subsections. When executed together in sequential order, they allow for determining the relevance and impact of a DI on SDGs at the indicator level. Figure 1 illustrates the structure of the DAF. In the following subsections, each segment and associated elements are described.

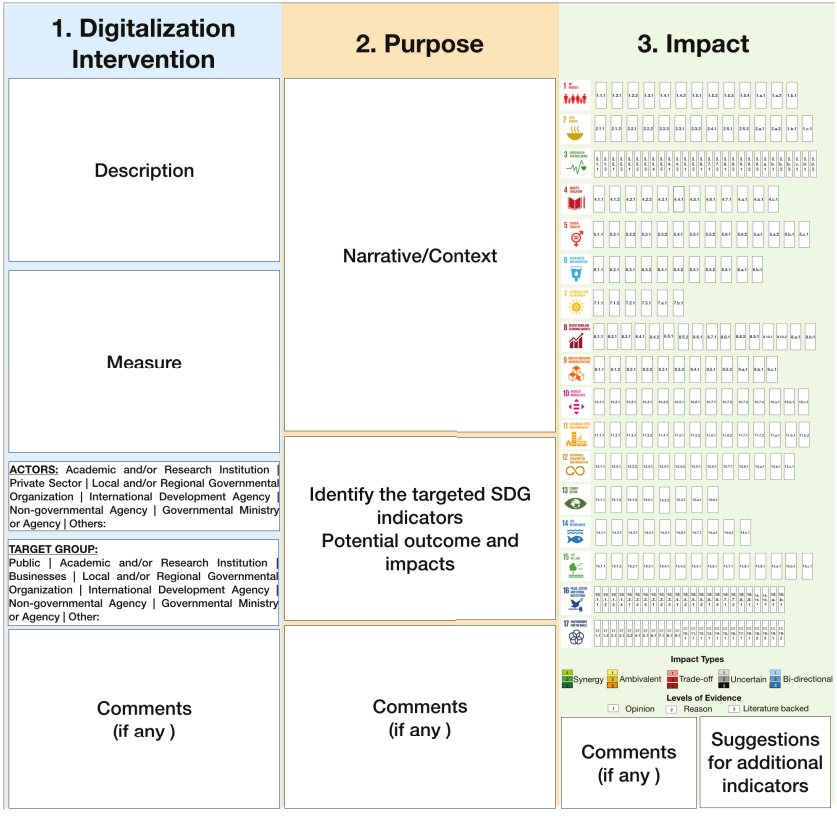


Figure 1. Digitainability Assessment Framework (DAF) overview.

3.1. Digitalization Intervention

The first segment of the DAF requires context-based information about a DI under consideration. It contains the following elements:

(a) Description;
(b) Measure;
(c) Actors;
(d) Target group;
(e) Comments.

The first four elements capture the basic description of the intervention, measure it will focus, along with the owner or primary actor of the intervention, and finally the target

group which a DI is serving, respectively. The fifth element of the segment (Comments) provides scope to integrate information that might be relevant and not covered in the preceding elements. Overall, the *Digitalization intervention* segment aims to gain the necessary information required to answer the questions:

1. What is a DI taken by the actor to bring change in the context of the SDGs?
2. What is the context within which the intervention is taking place?
3. Who are the initiators, and who are the intended receivers (e.g., governmental body, industry, NGO, international organization, public)?

3.2. Purpose

In the second segment of the DAF, information concerning:

(a) Narrative: defines the intended outcome from the DI.
(b) Envisaged SDG targets and indicators.
(c) Comments.

While the first element contains a narrative with some context about the intended effects of the intervention, the second element maps the narrative into the SDG targets and indicator framework [16]. As in the previous segment, the third element provides scope to integrate information that might be relevant and not covered in the preceding elements. Overall, the *Purpose* segment aims to gain the necessary information required to answer the questions:

1. What is the purpose of the DI (narrative)?
2. What are the targeted SDG indicators to be influenced by the DI?

3.3. Impact

This segment seeks information about the eventual impact of the DI beyond the envisaged *Purpose* (Segment 2). In the *Impact* segment, the user examines all the SDG indicators and evaluates the impact of the DI with respect to the indicator. We define the Impact types described below. Furthermore, for the assignment, we define three possible *Levels of Evidence* to back the claim. They are also defined below. Figure 2 represents the different *Impact Type* required to be backed with the *Level of Evidence* for particular impact type claimed. In the cases where the user feels that the existing SDGs targets and indicators are not specific enough to cover any specific aspect important for fostering sustainable development or lacking a systematized measure of progress considering the DI and SDGs, the comment section and the additional indicator suggestion section in the *Impact* segment provides the opportunity in the DAF to incorporate these vital information considering specific context and end outcomes.

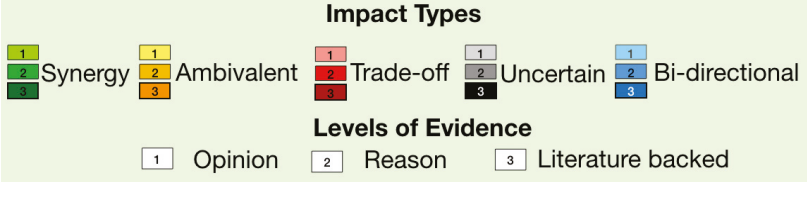


Figure 2. Legends of Impact Type required in the DAF backed with Levels of Evidence.

3.3.1. Impact Type

To incorporate the types of impact identified by the user for each SDG indicator, in the DAF, we classified *Impact types* into the following five categories:

(a) **Synergy:** implies that the DI impacts an SDG indicator in a synergistic manner. For example, the DI supporting Indicator 9.c.1 (*Proportion of population covered by a mobile network, by technology*) is likely to have a synergistic impact on the Indicator 17.8.1 (*Proportion of individuals using the Internet*).

(b) **Ambivalent:** implies that the DI impacts an SDG indicator in both a synergistic and trade-off manner. For example, the impact of AI technology on indicator 9.4.1 (*Carbon emission per unit of value-added*): while AI itself is a heavy energy consumer, on the one hand, it could also support in reducing the energy consumption if used conscientiously in energy systems, on the other hand. Often, the lack of verifiable information on the short- and projected medium-term impact is limited and suffers from a lack of systematic and accurate measurements [83].

(c) **Trade-off:** implies that, while the DI under consideration directly advances, particular indicator(s) might hinder the progress of other indicators of SDGs. For example, application of the DI for Indicator 3.8.1 (*Coverage of essential health services*) might lead to hindrances for Indicator 8.4.1 (*Material footprint, material footprint per capita, and material footprint per GDP*) or Indicator 7.3.1 (*Energy intensity measured in terms of primary energy and GDP*) because of increasing demand by digital infrastructure.

(d) **Uncertain:** implies that, while the DI might lead to an impact on the indicator, it is not ascertainable how and when (in the long term) there might be an impact. This impact type is meant to cover the situations where the logical inferences direct confidence on a particular type of impact, but evidence and rationale are not well identified. The second reason for assigning an impact to this type is if there is disagreement on the impact. Example, blockchain-based DI with demand–benefit uncertainties associated with respect to energy and finance impacting various SDG indicators related to climate change, energy demand, financial inclusiveness, and sustainable consumption.

(e) **Bi-directional:** In contrast to the previous four impact types, which are unidirectional (impact of the DI on indicators of SDGs), the bidirectional impact aims to identify the bi-directional, i.e., they are not only impacting indicators, but reversely they are (also) impacted by indicators. For example, the DI in smart grid systems might have a bi-directional impact on Indicator 7.1.1 (*Proportion of population with access to electricity*). However, for practical reasons, we will make one restriction when identifying the purpose of the *bi-directional* impact type. We will typically not include overarching aspects such as those related to peace, although they are crucial for sustainable development and are bi-directional. Unless a DI is explicitly related to peace or conflict issues, the importance of peace-related indicators is often self-evident and to not add to the risk-benefit analysis of the intervention.

The aforementioned *Impact types* must be backed by evidence in the DAF for a comprehensive assessment. Figure 2 illustrates the different Impact Types, with the numbers on the left representing different Levels of Evidence.

3.3.2. Levels of Evidence

We define three different *Levels of Evidence* to which impact on the indicators can be assessed, depending on the purpose for which the DAF is utilized.

1. **Opinion**: refers to personal opinions or beliefs, based on personal knowledge, without detailed investigation. While this level does not yet necessarily provide valid evidence, it may provide the first mapping of opinions, perceptions, and assumptions as a starting point but also collect relevant indicators, e.g., in a discussion group or a poll.
2. **Reason**: refers to judgment with a justification and, where appropriate, a discussion of the underlying displaying assumptions. This may be a next step beyond the *Opinion* level in a discussion group.
3. **Literature backed**: refers to literature and data-backed evidence from research, or practitioner knowledge, published in accredited sources.

It is encouraged that the user of the DAF provides evidence at Level 3 for legitimate inferences, but Levels 1 and 2 may represent useful precursor steps. Particularly, executing Level of Evidence 1 may reveal misperceptions. The DAF may also be used in an iterative way, refining *reason* and *opinion* based on expert level consideration of certain impacts. During the development process of the DAF, the elements of the DAF in each segment were nested or simplified to capture the most relevant information and minimize the number of inputs needed from the user. The refinement in the segments and their respective elements were based on multiple discussions between authors and reiterations to ensure that each element used in it enables the observation of critical aspects required to evaluate the impact of the DI for SDGs.

The information in the *Impact* segment helps in generating an impact profile that reflects on the following key questions:

1. What SDG indicator(s) are impacted by the DI? How?
2. How well is the DI aligned with the Agenda 2030?
3. What are the overall consequences of the DI for the SDGs as a whole, besides the intended outcome?

Inspecting the DI in a particular context requires scrutiny and in-depth evaluation of all SDGs indicators. Therefore, we encourage end-user to exercise the DAF in two phases:

- **First Phase:** identifying the relevant indicators, i.e., those indicators that logically be impacted by the DI.
- **Second Phase:** assessing the impact type for each indicator on the preferred levels of evidence.

To test the operationalization of the DAF, we use a custom-built excel sheet (Supplementary File S1), which covers both phases. The procedure prescribed in the scheme help in executing the DAF step by step. In the following section, we demonstrate the operationalization of the DAF with the help of three test case studies.

4. Results—Test Case Studies

This section describes the results of the studies we have undertaken to show how to operationalize the DAF in three different test case studies on the different levels of evidence to elucidate the application of different categories of *Impact Types* and *Levels of Evidence*.

4.1. DAF Test Case 1: Spatial Optimization for Systematic Deployment of Citizen-Driven Air Quality Monitoring Networks

According to WHO, around 92% of the world's population lives in places where air quality levels exceed prescribed limits [84]. In the recent pandemic, exposure to air pollution was one of the significant contributors, leading to an increase in COVID-19 cases [85,86]. Generally, air pollution monitoring is performed by environmental or governmental organizations using a network of fixed monitoring stations. With the significant advancements in the DIs such as the Internet of Things (IoT), cloud computing, edge computing and machine learning, citizens and environmental agencies are exploring the potential of citizen-driven air quality monitoring initiatives to enable high spatial resolution and real-time data collection on air quality in the cities [87]. However, to better understand the high-resolution spatial variability in air pollution in cities, data accuracy profoundly depends on the location *where* the data is collected. Considering the importance of the topic to public health globally, we use the work conducted by Gupta et al. [88] as a test case to understand the potential implication of the intervention for the achievement of SDGs. The method defined in this study supports identifying the "optimal" location in the city of Münster (Germany) to place the optimal number of IoT devices for air quality monitoring in a systematic manner for sustainable use of citizen's effort and IoT devices for air quality monitoring at the city level. The DI could directly support the progress toward SDG indicators:

- 3.9.1 *Mortality rate attributed to household and ambient air pollution*.
- 11.6.2 *Annual mean levels of fine particulate matter (e.g., PM2.5 and PM10) in cities (population weighted)*.

When analyzing the impact of the optimization method and IoT deployment using the DAF, we found 14 synergies and 3 potential trade-offs. We have also identified 15 indicators that can be impacted, but these impacts are unclear yet and far-fetched (Level 1 Impact Type). Most of the Level 1 impact synergies between indicators are based on the relationship rooted in the application of data gathered after the optimization method is utilized. Impact was identified for indicators: 1.a.2 *Proportion of total government spending on essential services (education, health, and social protection) for planning purposes* [89], exposure assessment in the context of 3.1.1 *Maternal mortality rate*, 3.2.2 *Neonatal mortality rate* [90], and 3.3.2 *Tuberculosis incidence per 100,000 population* [91]. An indirect relation was also found between air pollution and lifestyle-related disease in indicator 3.4.1 *Mortality rate attributed to cardiovascular disease, cancer, diabetes or chronic respiratory disease* [92]. Furthermore, increased access to high-resolution air quality data can empower medical research and health sectors, indicator (3.b.2) air pollution impacts development and growth in several forms [93]. Access to high-resolution data can decrease the impact of child physical and psychosocial development (4.2.1) along with the long-term impact on water bodies with ambient air quality (Indicator 6.3.2) [94,95]. The motive of the approach to also foster citizen participation in data collection enables many synergies, including indicator *11.3.2 Proportion of cities with direct participation structure of civil society in urban planning and management* [96], supporting education for sustainable development (Indicator 12.8.1). Active involvement of citizens could lead to improved population satisfaction with their experience of public services (Indicator 16.6.2) and population who believe that decision-making is inclusive and responsive.

Trade-offs were found with indicators linked to increased energy share in the *total final energy consumption* (Indicator 7.2.1) [97], *impact on the energy intensity of primary energy and GDP* (Indicator 7.3.1) [97], *increase in material footprint* (Indicator 8.4.1), and *impact on domestic material consumption* (Indicator 8.4.2) due to increasing demand of infrastructure for air quality data collection [98]. The bi-directional impact was also identified between Indicator 7.1.1 *Proportion of population with access to electricity and the DIs*. To install sensors for air quality measurement based on optimization method outcome, the population in the city needs to have access to electricity. The uncertain impacts include indicator 9.1.1 *Proportion of the rural population who live within 2 km of an all-season road as the optimization method uses road information data for assessing air quality*, 9.1.2 *Passenger and freight volumes as transportation is one of the significant contributors of air pollution*, 9.3.1 *Proportion of small-scale industries in total industry value-added as the industry also contributes to the air pollution*, and 9.4.1 CO_2 *emission per unit of value-added may get impacted because of access to information about air quality burden it might create*. Privacy issues are also not sufficiently covered by any of the SDG targets and associated indicators unless it leads to unjust and discriminatory outcomes and practices. We also identified the need for additional indicators to handle privacy and cybersecurity concerns while addressing sustainability with the DIs. Table 1 and Figure 3 summarize the results of test case 1. More detailed information of impact assessment for each indicator and associated comments can be assessed from the Supplementary File S2.

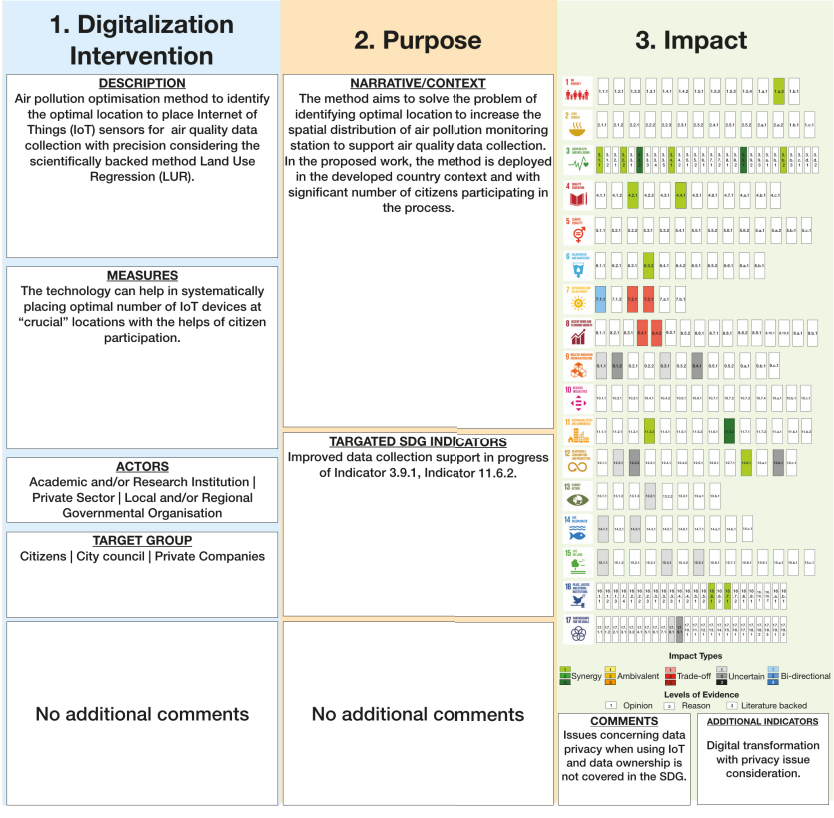


Figure 3. Filled DAF with all essential information concerning test case 1.

Table 1. The reference to publications where the *Impact Type* has been treated on Level of Evidence "Literature Backed" and the reasoning for Levels of Evidence "Reason" and "Opinion" for test case study 1 can be found in the Supplementary File S2.

DAF Test Case 1	
Impact Type	**Indicators**
Synergy	1.a.2 [89], 3.1.1 [90], 3.2.2 [90], 3.3.2 [91], 3.4.1 [92], 3.9.1 [93], 3.b.2 [93], 4.2.1, 4.4.1, 6.3.2 [94,95], 11.3.2 [96], 11.6.2 [88], 12.8.1, 16.6.1, 16.7.1
Ambivalent	NA
Trade-offs	7.2.1 [97], 7.3.1 [97], 8.4.1 [98], 8.4.2 [98]
Uncertain	9.1.1, 9.1.2, 9.3.1, 9.4.1, 12.2.1, 12.2.2, 12.b.1, 13.2.1, 14.1.1, 14.3.1, 15.1.2, 15.4.1, 15.5.1, 17.8.1, 17.9.1
Bi-Directional	7.1.1

4.2. DAF Test Case 2: Blockchain for Healthcare Service Delivery

Telehealth and telemedicine systems are supportive in remote healthcare services delivery, especially during the COVID-19 pandemic [99]. However, these systems are generally centralized and limited in providing necessary data security, integrity, transparency, preventing health records immutability, and traceability for fraud detection concerning insurance [100–102]. Blockchain is considered suitable for transactions with a limited digital footprint (decreased disclosure of sensitive digital data) alongside transparency and

immutability [103], and thus can satisfy the requirements of telehealth and telemedicine service delivery. Considering the increasing attention given to the blockchain in health care, we analyze the DI within the DAF.

The blockchain technology for healthcare delivery could help directly promote the progress towards SDG indicators:

- 3.8.1 *Coverage of essential health services*.
- 3.b.1 *Proportion of the target population covered by all vaccines included in their national programme*.
- 3.b.3 *Proportion of health facilities that have a core set of relevant essential medicines available and affordable on a sustainable basis*.

With the help of the DAF, we identified 16 possible synergies and 10 trade-offs. Five bi-directional and 11 uncertain impacts were further identified. Many of the synergies are attributed to increased access to the healthcare services and related infrastructure, such as increased access remotely in a privacy-preserving manner along with decreased digital footprint (by facilitating models that allow for minimal disclosure of sensitive digital data), which can lead to decreasing 3.1.1 *Maternal mortality ratio*, 3.2.1 *Under-5 mortality rate*, 3.4.1 *Mortality rate attributed to cardiovascular disease, cancer, diabetes or chronic respiratory disease*, and 3.4.2 *Suicide mortality rate* [104–107]. Increased trust in the system enabled by blockchain can also support in progressing 5.6.1 *Proportion of women aged 15–49 years who make their own informed decisions regarding sexual relations, contraceptive use, and reproductive health care*, 5.6.2 *Number of countries with laws and regulations that guarantee full and equal access to women and men aged 15 years and older to sexual and reproductive health care, information and education*, and 8.5.1 *Average hourly earnings of employees, by sex, age, occupation and persons with disabilities*, which can facilitate resources availability to strengthen statistical capacity in developing countries (17.19.1).

Likewise, possible trade-offs identified are related to the increased energy consumption (7.2.1, 7.3.1) and infrastructure requirements (12.2.1, 12.2.2) [108,109], resulting in environmental impacts (8.4.1, 8.4.2, 9.4.1, 13.2.2) and social impacts such as the digital divide (4.2.2). Further, the uncertain aspects mainly were identified for subsequent access to basic facilities and further resources other than healthcare (1.3.1,1.4.1,1.a.2,1.b.1, 5.b.1, 6.1.1,16.6.2,16.9.1) that the integrated approach of blockchain can empower [110]. The use of a blockchain system can not only support local-level healthcare access but can also boost the progress toward national and international cooperation and access to resources (17.14.1). However, information stored on the blockchain might be in conflict with General Data Protection Regulation (GDPR) policies (see [111]) as distributed databases hampers the allocation of responsibility and accountability, limits modification in data to comply with legal requirements, and the right to erase personal data [112–114]. This will negatively impact progress related to inclusiveness in decision making, adopting frameworks, tools, and services governing public access to the information relevant for indicators 16.7.2, 16.10.2, 17.15.1. Bi-directional impacts were linked to the available digital infrastructure (17.6.1, 17.8.1, 7.1.1) and policy frameworks to utilize the technology appropriately (9.a.1). Figure 4 presents the overall DAF populated with impacts identified considering the specific application and context of blockchain as a technology for SDGs. As can be seen from the figure, the advantage of the blockchain to decrease the digital footprint comes with a trade-off of a large ecological footprint. We also identified the lack of indicators that can concretely cover the aspect of the digital divide, ethical concerns such as respect to traditional belief while using technologies, as well as respect to cultural diversity, as commented in the impact segment. The results are summarized in Table 2 and Figure 4. Additional information can be found in the Supplementary File S3.

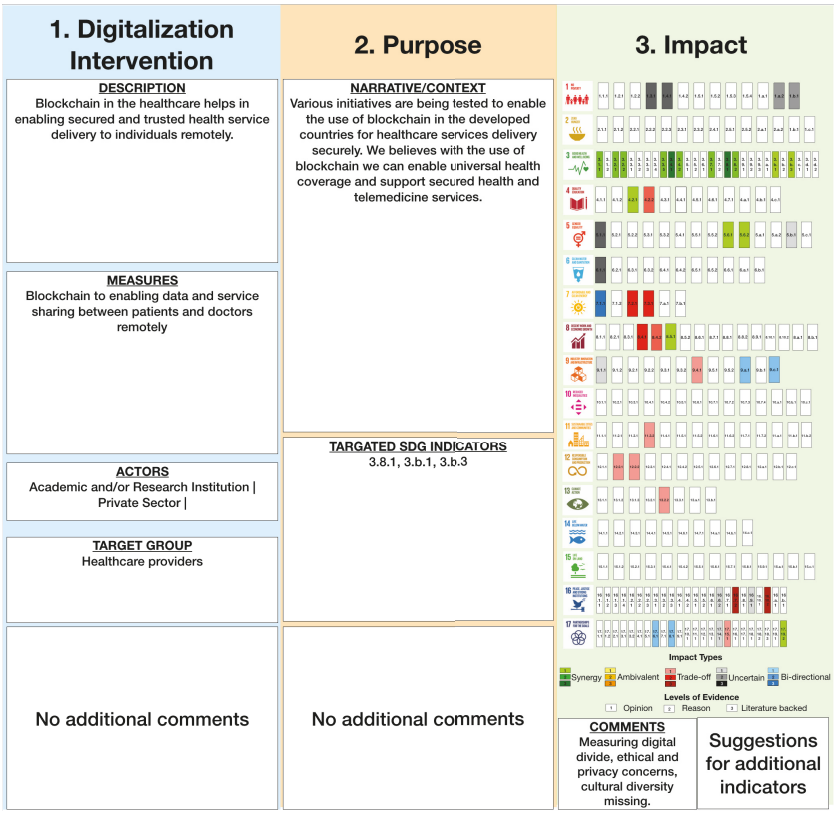


Figure 4. Filled DAF with all essential information concerning test case 2.

Table 2. The reference to publications where the *Impact Type* has been treated on Level of Evidence "Literature Backed" and the reasoning for Levels of Evidence "Reason" and "Opinion" for test case study 2 can be found in the Supplementary File S3.

DAF Test Case 2	
Impact Type	**Indicators**
Synergy	3.1.1 [104], 3.2.1 [105], 3.4.1 [106], 3.4.2 [107], 5.6.1, 5.6.2, 8.5.1, 17.19.1
Ambivalent	NA
Trade-offs	7.2.1 [108], 7.3.1 [108], 12.2.1 [109], 12.2.2 [109], 16.10.2 [112,114], 16.7.2 [112], 17.15.1
Uncertain	1.3.1, 1.4.1, 1.a.2, 1.b.1, 5.b.1, 6.1.1, 16.6.2, 16.9.1 [110], 17.14.1
Bi-Directional	9.a.1, 17.6.1, 17.8.1, 7.1.1

4.3. DAF Test Case 3: Machine Learning for Analysis of Satellite-Based Images for Disaster Risk Management

Machine learning (ML) is a subset of artificial intelligence (AI), but the two terms are sometimes used interchangeably. ML algorithms are machine programs that learn to perform particular tasks using a specific form of data inputs and rule sets. Disaster risk management and planning system relies on many different data sources and types for modeling. The application of ML approaches, such as artificial neural networks (ANN), support vector machines (SVM), and random forest (RF), and different deep-learning convolution

neural networks (CNNs) on satellite images for disaster risk management and planning is growing [115]. Regardless of the relevance of these methodological and technological advancements, little is known yet about their potential to support the achievement of Agenda 2030 as a whole. To holistically scrutinize their impact, we have considered a test case of ML on satellite images for disaster risk management and planning to explore potential impacts on SDGs with the help of the DAF. The machine learning method for disaster risk management and planning aims to directly support the progress of:

- 13.1.1/1.5.1 *Number of deaths, missing persons and directly affected persons attributed to disasters per 100,000 population*.

While the ML approach directly supports the above-listed indicator, we identified possible 25 indicators with synergies and 7 trade-offs with the DAF procedure. Many of the synergies are based on the notion that the various environmental factors considered in the disaster planning and management processes can lead to benefits for environmental indicators. Satellite image analysis include identification of land use change, type, and potential population at risk which could support in indicators: 2.1.2 *Prevalence of moderate or severe food insecurity in the population, based on the Food Insecurity Experience Scale (FIES)* [116], 6.3.2 *Proportion of bodies of water with good ambient water quality*, 6.4.2 *Level of water stress: freshwater withdrawal as a proportion of available freshwater resources* [117], 9.1.1 *Proportion of the rural population who live within 2 km of an all-season road*, 11.1.1 *Proportion of urban population living in slums, informal settlements or inadequate housing* [118], 11.3.1 *Ratio of land consumption rate to population growth rate*, 11.5.2 *Direct economic loss in relation to global GDP, damage to critical infrastructure and number of disruptions to essential services, attributed to disasters*, 11.7.1 *Average share of the built-up area of cities that is open space for public use for all, by sex, age and persons with disabilities*, 15.1.1 *Forest area as a proportion of total land area*, 15.1.2 *Proportion of important sites for terrestrial and freshwater biodiversity that are covered by protected areas, by ecosystem type*, 15.2.1 *Progress towards sustainable forest management*, 15.3.1 *Proportion of land that is degraded over total land area*, and 15.4.1 *Coverage by protected areas of important sites for mountain biodiversity*. Altogether, the advantages of the technology contribute in encouraging progress for 13.1.3 *Proportion of local governments that adopt and implement local disaster risk reduction strategies in line with national disaster risk reduction strategies*.

Significant trade-offs are identified with indicators 7.3.1 *Energy intensity measured in terms of primary energy and GDP*, 8.4.1/12.2.1 *Material footprint, material footprint per capita, and material footprint per GDP*, 8.4.2/12.2.2 *Domestic material consumption, domestic material consumption per capita, and domestic material consumption per GDP*, and 9.4.1 CO_2 *emission per unit of value-added* [119]. These trade-offs could be attributed to the energy demand and digital infrastructure required to train and implement the model. Uncertain impacts such as 1.1.1 *Proportion of the population living below the international poverty line by sex, age, employment status and geographic location (urban/rural)*, 1.4.1 *Proportion of population living in households with access to basic services*, 5.a.1 *(a) Proportion of total agricultural population with ownership or secure rights over agricultural land, by sex; and (b) share of women among owners or rights-bearers of agricultural land, by type of tenure*, 6.3.1 *Proportion of domestic and industrial wastewater flows safely treated*, 6.5.1 *Degree of integrated water resources management*, and 12.a.1 *Installed renewable energy-generating capacity in developing countries (in watts per capita)* could be benefited with the same technology if integrated processes are used in the long term. Bi-directional impacts on indicators such as 13.1.3 *Proportion of local governments that adopt and implement local disaster risk reduction strategies in line with national disaster risk reduction strategies* and 13.2.1 *Number of countries with nationally determined contributions, long-term strategies, national adaptation plans, and adaptation communications, as reported to the secretariat of the United Nations Framework Convention on Climate Change* are identified based on mutual benefits it provided to each other for sustainable development. Concern regarding remote sensing based privacy breaches as well as data divide are also not significantly covered in SDGs but might be covered when such issues lead to a lack of justice and accountability. The outcome of this test case study also indicates the need for additional indicators to address privacy concerns related to the DI, such as remote sensing. The additional indicator which

considers the digital footprint and underlying anonymity measure (or digital security) could be useful. Table 3 and Figure 5 summarize the results; additional information can be found in the Supplementary File S4.

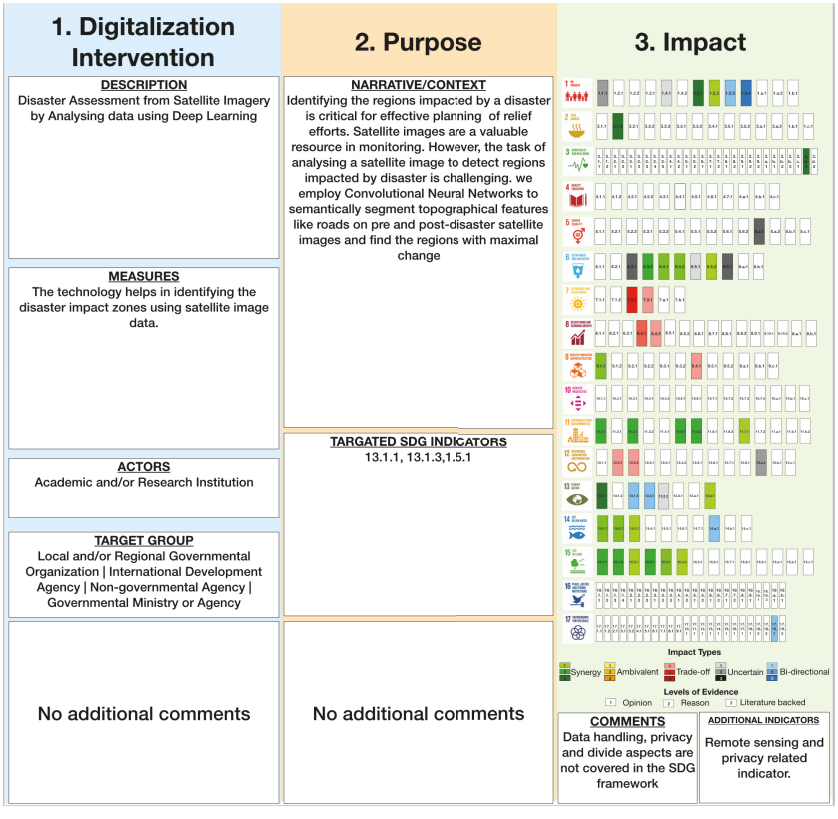


Figure 5. Filled DAF with all essential information concerning test case 3.

Table 3. The reference to publications where the *Impact Type* has been treated on Level of Evidence "Literature Backed" and the reasoning for Levels of Evidence "Reason" and "Opinion" for test case study 3 can be found in the Supplementary File S4.

DAF Test Case 3	
Impact Type	**Indicators**
Synergy	2.1.2 [116], 6.3.2 [117], 6.4.2 [117], 9.1.1, 11.1.1 [118], 11.3.1, 11.5.2, 11.7.1, 15.1.1, 15.1.2, 15.2.1, 15.3.1, 15.4.1
Ambivalent	NA
Trade-offs	7.3.1, 8.4.1/12.2.1, 8.4.2/12.2.2, 9.4.1 [119]
Uncertain	1.1.1, 1.4.1, 5.a.1, 6.3.1, 6.5.1, 12.a.1
Bi-Directional	13.1.3, 13.2.1,

The results of the three test case studies clearly demonstrate how the DAF adds value to our current knowledge on the DIs' impact on SDG. It helps to identify the potential diverse impacts of the DIs beyond the well-known synergies and trade-offs in SDGs. For test case study 1, where the focus of DI is on indicators of target 3.9 and 11.6, Anderson et al. [120]

reported synergies with targets 3.2, 3.3, 7.1, 15.1, 15.4, and 17.8, and trade-offs between indicators of 16.8, 10.6, 15.5 and target 11.6. Interestingly, the DAF acknowledges all the above interactions but also identifies several interactions beyond what is reported in recent literature. Likewise, for test case study 2, where the DI is targeted on the indicators of Target 3.8 and 3.b, Anderson et al. [120] reported trade-offs between target 3.b, 10.6, and 16.8 and synergy between 10.6, 16.8, and 15.5. Moreover, in test case study 3, where the DI addresses Target 13.1 and 1.5, synergies are reported with 11.5.

Overall, the results of the test case studies demonstrate that not only does the DAF identify DI impacts within the well-known synergies and trade-offs amid SDGs, but it also goes beyond existing knowledge by identifying casual pathways of linkages with far more interactions that are yet to be explored. Consequently, the proposed framework extends our understanding of the impact of the DIs on sustainable development beyond the current status quo.

5. Discussion

This study proposed a context inclusive and actor-specific framework for evaluating the practical impacts of the DIs and the cascading effect it may have on the synergy and trade-offs across the entire SDGs. The DAF approach draws from the Theory of Change (ToC) concepts. We adopted ToC schemes in the three DAF segments: *Digitalization Intervention*, *Purpose*, and *Impact*. Each segment of the DAF is dedicated to capturing specific practical information supportive of concealing the crucial information about the DIs' impact on SDGs, such as the actors involved, the target group, the context in which intervention is planned, the specific aim of the planned intervention, and the overall impact of intervention across all the SDG indicators. The exercise helps identify the impact digitalization may have considering the multifaceted interlinks between SDG indicators that may appear due to a particular context while using a particular DI. The DAF then supports developing the impact profile of the DI considering all the SDG indicators. Anderson et al. [120] recently reported the systems model based on the outcomes of past systemic interactions among SDG indicators spatiotemporally (across countries and years aggregated) at the global level. They identified levers and hurdles based on the current development trends. The outcomes from their study do not claim identified connections as "causal" but rather a methodological attempt in approaching causality. Based on statistical analysis and expert assessment, they applied systems modeling to explore the influence of levers and hurdles in achieving the SDGs. In our analysis, we used their SDG interaction data to compare the casual linkages identified by the DAF.

Different *Impact Type* and *Levels of Evidence* are incorporated in the DAF to help integrate multifaceted information and for adjusting potential biases, which may arise because of fragmented evidence and knowledge gaps in the literature. Level of Evidence 1 supports capturing the opinions to incorporate logical relations if the user believes it to be true. In Level 2, the user is expected to go further from their hypothesis in Level 1 with more concrete evidence, which may exist but are fragmented in literature. These first two *Levels of Evidence* guide the user first to integrate their hypothesis in the DAF, which can act as a precursor for the successive iterations. The DAF encourages the users to eventually fill the evidence at Level 3 so that the final outcome is less subjective and well-grounded in scientific evidence, where possible. In the cases where the user finds it challenging to reach Level of Evidence 3, the knowledge at Level 1 and Level 2 might act as gap identifiers, evidencing the need to further research for sustainable development through novel science-based methodologies with complexities across contexts and actors [121]. Proper grounding in reality with evidence will also ensure that the DAF outcomes are usable or doable, meaning that the intervention, actors, and context necessary for identifying impacts are explicitly identified for aligning the DIs for sustainable development [122]. As reflected in Section 3, the DAF can also be designated as an iterative process, intended to be an evolving tool, with a set of evidence-backed theories relevant to a specific context that could be articulated, tested, and improved over time [74]. In this sense, the DAF supports

in realizing the impacts not only by initial articulation but also as a lesson for long-term planning and practices. The DAF can also aid in understanding whether the new DIs can be transferred across the regions, especially taking into consideration the diverse requirements of each region's development processes.

The test case studies discussed in this paper showcase how the DAF helps identify potential risks, unintended consequences, and identification of potential missed opportunities by using digitalization. The framework presents an integrative approach along with the evidence-based one since both are crucial because finding concrete evidence for all the potential long and short-term impacts related to the DIs across all the indicators of SDGs is challenging. As can be identified from three different test case studies, the DAF also helps expand the sectorial and isolated goal level approaches existing in literature (Gupta et al. [8], Vinuesa et al. [12]) by recognizing the DIs' role in a certain context and the overarching effects on the SDG framework as a whole. The DAF also supports factoring in the indirect and uncertain impacts of the DIs for SDGs, which could help avoid uncritical optimism or systematic pessimism by accounting for all the indicators of SDGs while reiterating the DAF in different phases grounded in a particular context. Anticipating the role of the DI in context also supports comprehending that technological intervention is just part of a more extensive configuration and utilizing that may have closely tied trade-offs to other SDGs indicators and beyond, encouraging *mindfulness*. The DAF is designed to be improved further in future work with other actors as applied to various DIs.

The impact profile generated by filling potential impacts on all the SDG indicators for the three test case studies suggests that each of them have some aspect aligned in synergy (in Green), whereas some are not yet known to be aligned (in Yellow, Gray, and Blue) and some are trade-offs in their alignment (in Red). These colorful contrasting impacts suggest the need to balance them for sustainable development. The DAF supports revealing these contrasting impacts more systematically. Impact profiles of the three distinct DIs also provide fine-grained insights about diverse impacts on some SDGs over others. Test case 1 impact profile highlights more trade-offs for SDG 7 and 8, whereas in test case 2, the trade-offs increased for more SDGs such as in 9, 11, 12, and 13, and in test case 3, the trade-offs are approximately close to that of test case 2 but the increased synergies compared to test case 2 with a higher level of evidence reflect the balancing impact. Although the contexts for the three case studies and their purposes are different, the comparison above reflects the benefit the DAF provides to uncover these complex and diverse effects that need to be regarded for sustainable development.

Overall, the DAF makes it possible to systematically gather the fragmented scientific knowledge using the practical theories and knowledge to assess the vital gaps and collaborations required for fostering Agenda 2030 and beyond. For example, considering test case 2, we were able to identify potential trade-offs in Indicator 4.2.2, which could lead to a social divide; however, if utilized mindfully in collaboration with partners fostering digital learning, the trade-offs can serve as synergy and could support in the progress of many other indicators from Goal 3, 1, and 11 by encouraging participation and inclusiveness. We want to highlight that the DAF also helps identify aspects that are not covered by the SDG indicator (see, e.g., test cases 1 and 3 about privacy and cybersecurity), which may hint where a possible Post-Agenda 2030 might need to be extended. Some of the findings in the test cases have already been well known. For example, the recent studies prove how the DI paradigms, predominantly Big Data and Artificial Intelligence, are crucial for advancing the progress of SDGs [12,123,124]. However, the DAF makes it possible to systematically use the fragmented scientific literature, identify the crucial gaps, and observe potential collaborations required for fostering Agenda 2030 and beyond, comprehending tensions across space and time. Digitalization should be mindfully harnessed to monitor and achieve progress, then augment and scale up for releasing the full potential for sustainable development. Given the exceptional prospects that digitalization brings for progressing SDGs, the DAF serves as a novel analytical tool to explore the role of the

DIs by facilitating further assessments, supporting to cope with uncertainties, and context inclusive systemic actions.

It is important to note that we recognize that the DAF presented here has limitations, particularly in incorporating the highly complex interpretations, multidimensional nature, tensions among goals, and various assumptions about the impact of the DI on SDGs in the DAF and the degree of subjectivity it may carry. Nonetheless, incorporating the evidence within the DAF at Level of Evidence 3 as much as possible can minimize such limitations. Further limitations exist due to the inherent limits of the Agenda 2030 in its localization for specific contexts because of the heterogeneity due to political, societal, and economic structures that require reorientation for encouraging mindful digitalization for sustainable development [64,125]. Finally, we remark that the context and the stakeholder structures can be explored in more detail than we did in the three test case studies, which was not a primary goal of this research. This paper aims to provide a thorough description of the principles of the DAF and its significance for the digitainability assessment. Future work could also incorporate diverse actors and target groups, particularly those with contrasting contexts and interventions, to further knowledge generation towards inclusive action for sustainability.

6. Conclusions and Outlook

The goal of the DAF introduced in this article is to provide a systematic assessment of the impacts and potential, but also possible trade-offs of digitalization on sustainable development (the "digitainability"). The DAF should make it possible to provide a common umbrella for various partial and sectorial approaches in the literature by relating the analysis to the widely accepted SDG indicator framework. The DAF serves as a starting point for realizing the mindful choices considering the synergy, trade-offs, and critical aspects of the DIs for SDGs as a whole. The context-inclusive nature of the framework helps assess the impacts the DIs offers for SDGs at the indicator level considering local to global contemplation, assisting in a more targeted analysis of the impact DIs may have on critical interlinkages of SDGs. The DAF supports the step-by-step impact assessment of the DI across all SDG indicators and beyond, fostering evaluation across a broad spectrum of aspects such as ethics, environment, economy, and society in the digital age. Considering the emerging discussions on sustainability in the digital age, the outcome of the present study serves as an important means for key stakeholders to mindfully articulate, collaborate, and deliver the systems change required to harness the opportunities offered by digitalization for sustainable development. It offers stakeholders a strategic tool for spotting the potential opportunities and risks while maintaining its responsibility for sustainable development. The integration of practical insights with a theoretical outlook helps predict the future, which is not accurate science. The DAF could help in identifying potential pathways through the complex digitalization–sustainability practices. The framework also helps identify traits that are not covered by the current indicator framework of Agenda 2030 for enriching the discussion about the mindful use of digitalization for future sustainable development in the digital age. Beyond that, to enhance the perceived link between digitalization and sustainable development, the DAF guides in gathering crucial scientific pieces of evidence about how various actors and stakeholders can undertake effort mindfully. Altogether, the DAF helps in identifying the gaps and asking the right question, pinpointing potential impacts digitalization may have to shape the future more sustainably.

Future work will focus on automating some of the DAF procedures with the help of machine learning approaches, including topic-modeling and automatic text extraction, that have helped investigate the critical topics in a set of documents, such as articles published in academic journals, political texts, and in data-driven journalism [126,127]. Automating the topic extraction process in the future might support decreasing subjectivity biases and enhance re-iteration of the evidence analysis. In the long run, particularly in a context where literature on the DIs for SDGs is growing, the automated process might allow assessing key considerations and criteria from diverse domains simultaneously. However, it will be

important to consider that the automated processes should be grounded in the context of the intervention and the role of the actors in the process. The scope of the DAF presented in this paper is to stimulate the assessment of the impact of DI on sustainable development, devising methodological agenda dedicated to evaluating the technical improvements required in the DI for being aligned with SDG indicators. Future empirical studies could utilize approaches such as design science research methodology [59] for going a step ahead and developing systematic evaluation frameworks for the improvement of DIs.The problem-solving paradigm of design science research inculcates the application of crucial dimensions discussed in this paper, such as context and stakeholders within design theories to realize the innovative artifacts for addressing real-world problems [128], offering guidelines for evaluation and iteration to improve the DIs for sustainable development.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/su14053114/s1, Supplementary File S1: S1_EmptyDAFExcel.xlsx; Supplementary File S2: S2_DAF_AQ_optimisation.xlsx; Supplementary File S3: S3_DAF_Blockchain _updated.xlsx; Supplementary File S4: S4_DAF_ML.xlsx.

Author Contributions: S.G., conceptualization, developed methodology, performed formal analysis, investigated, prepared the written original draft. S.G. and J.R. reviewed and edited the draft along with essential feedback. All authors have read and agreed to the published version of the manuscript.

Funding: This research was carried out within the project "digitainable", funded by the German Federal Ministry for Education and Research (BMBF).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data can be assessed from supplementary files for all the test case studies.

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

References

1. United Nations Department of Economic and Social Affairs. *Transforming Our World: The 2030 Agenda for Sustainable Development*; United Nations Department of Economic and Social Affairs: New York, NY, USA, 2015.
2. Chakraborty, C. *Artificial Intelligence and the Fourth Industrial Revolution*; Jenny Stanford Publishing: Singapore, 2021.
3. Tegmark, M. *Life 3.0: Being Human in the Age of Artificial Intelligence*; Knopf: New York, NY, USA, 2017.
4. Katz, R.L.; Koutroumpis, P. Measuring digitization: A growth and welfare multiplier. *Technovation* **2013**, *33*, 314–319. [CrossRef]
5. Vial, G. Understanding digital transformation: A review and a research agenda. *J. Strateg. Inf. Syst.* **2019**, *28*, 118–144. [CrossRef]
6. Verhoef, P.C.; Broekhuizen, T.; Bart, Y.; Bhattacharya, A.; Dong, J.Q.; Fabian, N.; Haenlein, M. Digital transformation: A multidisciplinary reflection and research agenda. *J. Bus. Res.* **2021**, *122*, 889–901. [CrossRef]
7. Van der Velden, M. Digitalisation and the UN Sustainable Development Goals: What role for design. *ID&A Interact. Des. Archit.* **2018**, *37*, 160–174.
8. Gupta, S.; Motlagh, M.; Rhyner, J. The digitalization sustainability matrix: A participatory research tool for investigating digitainability. *Sustainability* **2020**, *12*, 9283. [CrossRef]
9. Goralski, M.A.; Tan, T.K. Artificial intelligence and sustainable development. *Int. J. Manag. Educ.* **2020**, *18*, 100330. [CrossRef]
10. Zhao, Z.; Cai, M.; Wang, F.; Winkler, J.A.; Connor, T.; Chung, M.G.; Zhang, J.; Yang, H.; Xu, Z.; Tang, Y.; et al. Synergies and tradeoffs among Sustainable Development Goals across boundaries in a metacoupled world. *Sci. Total Environ.* **2021**, *751*, 141749. [CrossRef]
11. van Wynsberghe, A. Sustainable AI: AI for sustainability and the sustainability of AI. *AI Ethics* **2021**, *1*, 213–218. [CrossRef]
12. Vinuesa, R.; Azizpour, H.; Leite, I.; Balaam, M.; Dignum, V.; Domisch, S.; Felländer, A.; Langhans, S.D.; Tegmark, M.; Nerini, F.F. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nat. Commun.* **2020**, *11*, 233. [CrossRef]
13. Kuntsman, A.; Rattle, I. Towards a paradigmatic shift in sustainability studies: A systematic review of peer reviewed literature and future agenda setting to consider environmental (Un) sustainability of digital communication. *Environ. Commun.* **2019**, *13*, 567–581. [CrossRef]
14. Yoon, B.; Shin, J.; Lee, S. Technology assessment model for sustainable development of LNG terminals. *J. Clean. Prod.* **2018**, *172*, 927–937. [CrossRef]
15. Andries, A.; Morse, S.; Murphy, R.; Lynch, J.; Woolliams, E.; Fonweban, J. Translation of Earth observation data into sustainable development indicators: An analytical framework. *Sustain. Dev.* **2019**, *27*, 366–376. [CrossRef]

16. United Nations Department of Economic and Social Affairs. Global Indicator Framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development. Available online: https://unstats.un.org/sdgs/indicators/Global%20Indicator%20Framework%20after%202021%20refinement_Eng.pdf (accessed on 1 September 2021).
17. Sacco, P.; Gargano, E.R.; Cornella, A. Sustainable digitalization: A systematic literature review to identify how to make digitalization more sustainable. In Proceedings of the International TRIZ Future Conference, Bolzano, Italy, 22–24 September 2021; pp. 14–29.
18. Pappas, I.O.; Mikalef, P.; Giannakos, M.N.; Krogstie, J.; Lekakos, G. Big data and business analytics ecosystems: Paving the way towards digital transformation and sustainable societies. *Inf. Syst. Bus. Manag.* **2018**, *16*, 479–491. [CrossRef]
19. E-Participation Index. Available online:https://publicadministration.un.org/egovkb/en-us/About/Overview/E-Participation-Index (accessed on 1 June 2021).
20. Kostoska, O.; Kocarev, L. A novel ICT framework for sustainable development goals. *Sustainability* **2019**, *11*, 1961. [CrossRef]
21. Chalmers, G. The SDG Impact Assessment Tool-a free online tool for self-assessments of impacts on Agenda 2030. *Policy* **2019**, *1*, 150–167.
22. Cowls, J.; Tsamados, A.; Taddeo, M.; Floridi, L. A definition, benchmark and database of AI for social good initiatives. *Nat. Mach. Intell.* **2021**, *3*, 111–115. [CrossRef]
23. del Río Castro, G.; Fernández, M.C.G.; Colsa, Á.U. Unleashing the convergence amid digitalization and sustainability towards pursuing the Sustainable Development Goals (SDGs): A holistic review. *J. Clean. Prod.* **2020**, *208*, 122204. [CrossRef]
24. Lu, Y.; Nakicenovic, N.; Visbeck, M.; Stevance, A.S. Policy: Five priorities for the UN sustainable development goals. *Nat. News* **2015**, *520*, 432. [CrossRef]
25. Schmidt, H.; Gostin, L.O.; Emanuel, E.J. Public health, universal health coverage, and Sustainable Development Goals: Can they coexist? *Lancet* **2015**, *386*, 928–930. [CrossRef]
26. Adshead, D.; Thacker, S.; Fuldauer, L.I.; Hall, J.W. Delivering on the Sustainable Development Goals through long-term infrastructure planning. *Glob. Environ. Change* **2019**, *59*, 101975. [CrossRef]
27. Schroeder, P.; Anggraeni, K.; Weber, U. The relevance of circular economy practices to the sustainable development goals. *J. Ind. Ecol.* **2019**, *23*, 77–95. [CrossRef]
28. Khalili, N.R.; Cheng, W.; McWilliams, A. A methodological approach for the design of sustainability initiatives: In pursuit of sustainable transition in China. *Sustain. Sci.* **2017**, *12*, 933–956. [CrossRef]
29. Collste, D.; Pedercini, M.; Cornell, S.E. Policy coherence to achieve the SDGs: Using integrated simulation models to assess effective policies. *Sustain. Sci.* **2017**, *12*, 921–931. [CrossRef] [PubMed]
30. Nerini, F.F.; Tomei, J.; To, L.S.; Bisaga, I.; Parikh, P.; Black, M.; Borrion, A.; Spataru, C.; Broto, V.C.; Anandarajah, G.; et al. Mapping synergies and trade-offs between energy and the Sustainable Development Goals. *Nat. Energy* **2018**, *3*, 10–15. [CrossRef]
31. Velis, M.; Conti, K.I.; Biermann, F. Groundwater and human development: Synergies and trade-offs within the context of the sustainable development goals. *Sustain. Sci.* **2017**, *12*, 1007–1017. [CrossRef]
32. Bisaga, I.; Parikh, P.; Tomei, J.; To, L.S. Mapping synergies and trade-offs between energy and the sustainable development goals: A case study of off-grid solar energy in Rwanda. *Energy Policy* **2021**, *149*, 112028. [CrossRef]
33. Singh, G.G.; Cisneros-Montemayor, A.M.; Swartz, W.; Cheung, W.; Guy, J.A.; Kenny, T.A.; McOwen, C.J.; Asch, R.; Geffert, J.L.; Wabnitz, C.C.; et al. A rapid assessment of co-benefits and trade-offs among Sustainable Development Goals. *Mar. Policy* **2018**, *93*, 223–231. [CrossRef]
34. Pradhan, P.; Costa, L.; Rybski, D.; Lucht, W.; Kropp, J.P. A systematic study of sustainable development goal (SDG) interactions. *Earth's Future* **2017**, *5*, 1169–1179. [CrossRef]
35. Scherer, L.; Behrens, P.; de Koning, A.; Heijungs, R.; Sprecher, B.; Tukker, A. Trade-offs between social and environmental Sustainable Development Goals. *Environ. Sci. Policy* **2018**, *90*, 65–72. [CrossRef]
36. Mainali, B.; Luukkanen, J.; Silveira, S.; Kaivo-Oja, J. Evaluating synergies and trade-offs among Sustainable Development Goals (SDGs): Explorative analyses of development paths in South Asia and Sub-Saharan Africa. *Sustainability* **2018**, *10*, 815. [CrossRef]
37. von Stechow, C.; Minx, J.C.; Riahi, K.; Jewell, J.; McCollum, D.L.; Callaghan, M.W.; Bertram, C.; Luderer, G.; Baiocchi, G. 2 °C and SDGs: United they stand, divided they fall? *Environ. Res. Lett.* **2016**, *11*, 034022. [CrossRef]
38. Sorrell, S. Jevons' Paradox revisited: The evidence for backfire from improved energy efficiency. *Energy Policy* **2009**, *37*, 1456–1469. [CrossRef]
39. Nishant, R.; Teo, T.S.; Goh, M. Energy efficiency benefits: Is technophilic optimism justified? *IEEE Trans. Eng. Manag.* **2014**, *61*, 476–487. [CrossRef]
40. Hidalgo, A.; Gabaly, S.; Morales-Alonso, G.; Urueña, A. The digital divide in light of sustainable development: An approach through advanced machine learning techniques. *Technol. Forecast. Soc. Change* **2020**, *150*, 119754. [CrossRef]
41. Kopnina, H. Education for the future? Critical evaluation of education for sustainable development goals. *J. Environ. Educ.* **2020**, *51*, 280–291. [CrossRef]
42. Sanchez, D.O.M. Sustainable Development Challenges and Risks of Industry 4.0: A literature review. In Proceedings of the 2019 Global IoT Summit (GIoTS), Aarhus, Denmark, 17–21 June 2019; pp. 1–6.
43. Dawes, J.H. Are the Sustainable Development Goals self-consistent and mutually achievable? *Sustain. Dev.* **2020**, *28*, 101–117. [CrossRef]
44. Fukuda-Parr, S.; McNeill, D. Knowledge and Politics in Setting and Measuring the SDGs. *Glob. Policy* **2019**, *10*, 5–15. [CrossRef]

45. Letouzé, E.; Pentland, A. Towards a human artificial intelligence for human development. *ITU J. ICT Discov.* **2018**, *2*, 1–8.
46. Weitz, N.; Carlsen, H.; Nilsson, M.; Skånberg, K. Towards systemic and contextual priority setting for implementing the 2030 Agenda. *Sustain. Sci.* **2018**, *13*, 531–548. [CrossRef]
47. Tsamados, A.; Aggarwal, N.; Cowls, J.; Morley, J.; Roberts, H.; Taddeo, M.; Floridi, L. The ethics of algorithms: Key problems and solutions. *AI Soc.* **2021**, *37*, 215–230. [CrossRef]
48. Breuer, A.; Janetschek, H.; Malerba, D. Translating sustainable development goal (SDG) interdependencies into policy advice. *Sustainability* **2019**, *11*, 2092. [CrossRef]
49. Kroll, C.; Warchold, A.; Pradhan, P. Sustainable Development Goals (SDGs): Are we successful in turning trade-offs into synergies? *Palgrave Commun.* **2019**, *5*, 1–11. [CrossRef]
50. Schneider, F.; Kläy, A.; Zimmermann, A.B.; Buser, T.; Ingalls, M.; Messerli, P. How can science support the 2030 Agenda for Sustainable Development? Four tasks to tackle the normative dimension of sustainability. *Sustain. Sci.* **2019**, *14*, 1593–1604. [CrossRef]
51. Edwards, N.; Barker, P.M. The importance of context in implementation research. *JAIDS J. Acquir. Immune Defic. Syndr.* **2014**, *67*, S157–S162. [CrossRef] [PubMed]
52. May, C.R.; Johnson, M.; Finch, T. Implementation, context and complexity. *Implement. Sci.* **2016**, *11*, 1–12. [CrossRef] [PubMed]
53. Nilsen, P. Making sense of implementation theories, models, and frameworks. In *Implementation Science 3.0*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 53–79.
54. Gabora, L.; Aerts, D. Evolution as context-driven actualisation of potential: Toward an interdisciplinary theory of change of state. *Interdiscip. Sci. Rev.* **2005**, *30*, 69–88. [CrossRef]
55. Tamilmani, K.; Rana, N.P.; Wamba, S.F.; Dwivedi, R. The extended Unified Theory of Acceptance and Use of Technology (UTAUT2): A systematic literature review and theory evaluation. *Int. J. Inf. Manag.* **2021**, *57*, 102269. [CrossRef]
56. Pfadenhauer, L.M.; Gerhardus, A.; Mozygemba, K.; Lysdahl, K.B.; Booth, A.; Hofmann, B.; Wahlster, P.; Polus, S.; Burns, J.; Brereton, L.; et al. Making sense of complexity in context and implementation: The Context and Implementation of Complex Interventions (CICI) framework. *Implement. Sci.* **2017**, *12*, 1–17. [CrossRef]
57. Dang, H.A.H.; Serajuddin, U. Tracking the Sustainable Development Goals: Emerging Measurement Challenges and Further Reflections. *World Dev.* **2019**, *127*, 104570. [CrossRef]
58. Barbier, E.B.; Burgess, J.C. Sustainable development goal indicators: Analyzing trade-offs and complementarities. *World Dev.* **2019**, *122*, 295–305. [CrossRef]
59. vom Brocke, J.; Hevner, A.; Maedche, A. Introduction to Design Science Research. In *Design Science Research Cases*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–13.
60. Grainger-Brown, J.; Malekpour, S. Implementing the sustainable development goals: A review of strategic tools and frameworks available to organisations. *Sustainability* **2019**, *11*, 1381. [CrossRef]
61. Laurian, L.; Walker, M.; Crawford, J. Implementing environmental sustainability in local government: The impacts of framing, agency culture, and structure in US cities and counties. *Int. J. Public Adm.* **2017**, *40*, 270–283. [CrossRef]
62. Krellenberg, K.; Bergsträßer, H.; Bykova, D.; Kress, N.; Tyndall, K. Urban sustainability strategies guided by the SDGs—A tale of four cities. *Sustainability* **2019**, *11*, 1116. [CrossRef]
63. Köhler, J.; Geels, F.W.; Kern, F.; Markard, J.; Onsongo, E.; Wieczorek, A.; Alkemade, F.; Avelino, F.; Bergek, A.; Boons, F.; et al. An agenda for sustainability transitions research: State of the art and future directions. *Environ. Innov. Soc. Transit.* **2019**, *31*, 1–32. [CrossRef]
64. Walsh, P.P.; Murphy, E.; Horan, D. The role of science, technology and innovation in the UN 2030 agenda. *Technol. Forecast. Soc. Change* **2020**, *154*, 119957. [CrossRef]
65. Senit, C.A. Leaving no one behind? The influence of civil society participation on the Sustainable Development Goals. *Environ. Plan. C Politics Space* **2020**, *38*, 693–712. [CrossRef]
66. Guan, T.; Meng, K.; Liu, W.; Xue, L. Public attitudes toward sustainable development goals: Evidence from five Chinese cities. *Sustainability* **2019**, *11*, 5793. [CrossRef]
67. Messerli, P.; Kim, E.M.; Lutz, W.; Moatti, J.P.; Richardson, K.; Saidam, M.; Smith, D.; Eloundou-Enyegue, P.; Foli, E.; Glassman, A.; et al. Expansion of sustainability science needed for the SDGs. *Nat. Sustain.* **2019**, *2*, 892–894. [CrossRef]
68. Dalby, S.; Horton, S.; Mahon, R.; Thomaz, D. *Achieving the Sustainable Development Goals: Global Governance Challenges*; Routledge: Abingdon-on-Thames, UK, 2019.
69. Kurz, R. UN SDGs: Disruptive for companies and for universities? In *The Future of the UN Sustainable Development Goals*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 279–290.
70. Bergman, Z.; Bergman, M.M.; Fernandes, K.; Grossrieder, D.; Schneider, L. The contribution of UNESCO chairs toward achieving the UN sustainable development goals. *Sustainability* **2018**, *10*, 4471. [CrossRef]
71. Lopez, B. Connecting business and sustainable development goals in Spain. *Mark. Intell. Plan.* **2020**, *38*, 573–585. [CrossRef]
72. Pimonenko, T.; Bilan, Y.; Horák, J.; Starchenko, L.; Gajda, W. Green brand of companies and greenwashing under sustainable development goals. *Sustainability* **2020**, *12*, 1679. [CrossRef]
73. Allen, W.; Cruz, J.; Warburton, B. How decision support systems can benefit from a theory of change approach. *Environ. Manag.* **2017**, *59*, 956–965. [CrossRef] [PubMed]

74. Anderson, A.A. The Community Builder's Approach to Theory of Change. 2006. Available online: https://www.theoryofchange.org/pdf/TOC_fac_guide.pdf (accessed on 14 September 2021).
75. James, C. *Theory of Change Review*; Comic Relief: London, UK, 2011.
76. Bonell, C.; Melendez-Torres, G.; Viner, R.M.; Rogers, M.B.; Whitworth, M.; Rutter, H.; Rubin, G.J.; Patton, G. An evidence-based theory of change for reducing SARS-CoV-2 transmission in reopened schools. *Health Place* **2020**, *64*, 102398. [CrossRef] [PubMed]
77. Mayne, J. Sustainability Analysis of Intervention Benefits: A Theory of Change Approach. *Can. J. Program Eval.* **2020**, *35* , 204-221. [CrossRef]
78. Li, Y.; Thomas, M.A. Adopting a theory of change approach for ict4d project impact assessment-the case of cmes project. In Proceedings of the International Conference on Social Implications of Computers in Developing Countries, Dar es Salaam, Tanzania, 1–3 May 2019; pp. 95–109.
79. Taplin, D.H.; Clark, H. *Theory of Change Basics: A Primer on Theory of Change*; ActKnowledge: New York, NY, USA, 2012.
80. Stein, D.; Valters, C. Understanding Theory of Change in International Development. 2012. Available online: https://www.theoryofchange.org/wp-content/uploads/toco_library/pdf/UNDERSTANDINGTHEORYOFChangeSteinValtersPN.pdf (accessed on 12 October 2021).
81. Van Stolk, C.; Ling, T.; Reding, A.; Bassford, M. Monitoring and Evaluation in Stabilisation Interventions. 2011. Available online: https://www.rand.org/content/dam/rand/pubs/technical_reports/2011/RAND_TR962.pdf (accessed on 18 October 2021).
82. Valters, C. Theories of change in international development: Communication, learning, or accountability. *JSRP Pap.* **2014**, *17*, 1–21.
83. Cowls, J.; Tsamados, A.; Taddeo, M.; Floridi, L. The AI Gambit—Leveraging Artificial Intelligence to Combat Climate Change: Opportunities, Challenges, and Recommendations. *AI Soc.* **2021**, 1–25. [CrossRef]
84. WHO. *WHO Releases Country Estimates on Air Pollution Exposure and Health Impact*; WHO: Geneva, Switzerland, 2016.
85. Travaglio, M.; Yu, Y.; Popovic, R.; Selley, L.; Leal, N.S.; Martins, L.M. Links between air pollution and COVID-19 in England. *Environ. Pollut.* **2021**, *268*, 115859. [CrossRef]
86. Fattorini, D.; Regoli, F. Role of the chronic air pollution levels in the Covid-19 outbreak risk in Italy. *Environ. Pollut.* **2020**, *264*, 114732. [CrossRef]
87. Jiao, W.; Hagler, G.; Williams, R.; Sharpe, R.; Brown, R.; Garver, D.; Judge, R.; Caudill, M.; Rickard, J.; Davis, M.; et al. Community Air Sensor Network (CAIRSENSE) project: Evaluation of low-cost sensor performance in a suburban environment in the southeastern United States. *Atmos. Meas. Tech.* **2016**, *9*, 5281–5292. [CrossRef]
88. Gupta, S.; Pebesma, E.; Degbelo, A.; Costa, A.C. Optimising Citizen-Driven Air Quality Monitoring Networks for Cities. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 468. [CrossRef]
89. Chen, F.; Chen, Z. Cost of economic growth: Air pollution and health expenditure. *Sci. Total Environ.* **2021**, *755*, 142543. [CrossRef] [PubMed]
90. Owili, P.O.; Lien, W.H.; Muga, M.A.; Lin, T.H. The associations between types of ambient PM2.5 and under-five and maternal mortality in Africa. *Int. J. Environ. Res. Public Health* **2017**, *14*, 359. [CrossRef] [PubMed]
91. Popovic, I.; Magalhaes, R.J.S.; Ge, E.; Marks, G.B.; Dong, G.H.; Wei, X.; Knibbs, L.D. A systematic literature review and critical appraisal of epidemiological studies on outdoor air pollution and tuberculosis outcomes. *Environ. Res.* **2019**, *170*, 33–45. [CrossRef] [PubMed]
92. Rajagopalan, S.; Al-Kindi, S.G.; Brook, R.D. Air pollution and cardiovascular disease: JACC state-of-the-art review. *J. Am. Coll. Cardiol.* **2018**, *72*, 2054–2070. [CrossRef]
93. Liang, L.; Cai, Y.; Barratt, B.; Lyu, B.; Chan, Q.; Hansell, A.L.; Xie, W.; Zhang, D.; Kelly, F.J.; Tong, Z. Associations between daily air quality and hospitalisations for acute exacerbation of chronic obstructive pulmonary disease in Beijing, 2013–2017: An ecological analysis. *Lancet Planet. Health* **2019**, *3*, e270–e279. [CrossRef]
94. Nissilä, J.J.; Savelieva, K.; Lampi, J.; Ung-Lanki, S.; Elovainio, M.; Pekkanen, J. Parental worry about indoor air quality and student symptom reporting in primary schools with or without indoor air quality problems. *Indoor Air* **2019**, *29*, 865–873. [CrossRef]
95. Casazza, M.; Maraga, F.; Liu, G.; Lega, M.; Turconi, L.; Ulgiati, S. River water quality and its relation with air quality: A long-term case study in a remote and pristine NW Italian headwater catchment. *J. Environ. Account. Manag.* **2017**, *5*, 35–47. [CrossRef]
96. Camprodon, G.; González, Ó.; Barberán, V.; Pérez, M.; Smári, V.; de Heras, M.Á.; Bizzotto, A. Smart Citizen Kit and Station: An open environmental monitoring system for citizen participation and scientific experimentation. *HardwareX* **2019**, *6*, e00070. [CrossRef]
97. Misic, J.; Misic, V.B.; Banaie, F. Reliable and scalable data acquisition from IoT domains. In Proceedings of the GLOBECOM 2017–2017 IEEE Global Communications Conference, Singapore, 4–8 December 2017; pp. 1–6.
98. Tu, M.; Chung, W.H.; Chiu, C.K.; Chung, W.; Tzeng, Y. A novel IoT-based dynamic carbon footprint approach to reducing uncertainties in carbon footprint assessment of a solar PV supply chain. In Proceedings of the 2017 4th International Conference on Industrial Engineering and Applications (ICIEA), Nagoya, Japan, 21–23 April 2017; pp. 249–254.
99. Chamola, V.; Hassija, V.; Gupta, V.; Guizani, M. A comprehensive review of the COVID-19 pandemic and the role of IoT, drones, AI, blockchain, and 5G in managing its impact. *IEEE Access* **2020**, *8*, 90225–90265. [CrossRef]
100. Jin, Z.; Chen, Y. Telemedicine in the cloud era: Prospects and challenges. *IEEE Pervasive Comput.* **2015**, *14*, 54–61. [CrossRef]
101. Ekeland, A.G.; Bowes, A.; Flottorp, S. Effectiveness of telemedicine: A systematic review of reviews. *Int. J. Med. Inform.* **2010**, *79*, 736–771. [CrossRef] [PubMed]

102. Margheri, A.; Masi, M.; Miladi, A.; Sassone, V.; Rosenzweig, J. Decentralised provenance for healthcare data. *Int. J. Med. Informatics* **2020**, *141*, 104197. [CrossRef] [PubMed]
103. Series, B.P. Opportunities and Challenges of Blockchain Technologies in Health Care. 2020. Available online: https://www.oecd.org/finance/Opportunities-and-Challenges-of-Blockchain-Technologies-in-Health-Care.pdf (accessed on 21 November 2021).
104. Musabi, A.G.; Thiga, M.M.; Karume, S.M. Enabling Secure Maternal Health Information Exchange using Blockchain. Available online: https://conf.kabarak.ac.ke/event/4/contributions/107/contribution.pdf (accessed on 21 November 2021).
105. Resiere, D.; Resiere, D.; Kallel, H. Implementation of medical and scientific cooperation in the Caribbean using blockchain technology in coronavirus (Covid-19) pandemics. *J. Med. Syst.* **2020**, *44*, 123. [CrossRef] [PubMed]
106. Krittanawong, C.; Rogers, A.J.; Aydar, M.; Choi, E.; Johnson, K.W.; Wang, Z.; Narayan, S.M. Integrating blockchain technology with artificial intelligence for cardiovascular medicine. *Nat. Rev. Cardiol.* **2020**, *17*, 1–3. [CrossRef]
107. Bell, L.; Buchanan, W.J.; Cameron, J.; Lo, O. Applications of blockchain within healthcare. *Blockchain Healthc. Today* **2018**, *1*, 1–7. [CrossRef]
108. Sedlmeir, J.; Buhl, H.U.; Fridgen, G.; Keller, R. The energy consumption of blockchain technology: Beyond myth. *Bus. Inf. Syst. Eng.* **2020**, *62*, 599–608. [CrossRef]
109. Stoll, C.; Klaaßen, L.; Gallersdörfer, U. The carbon footprint of bitcoin. *Joule* **2019**, *3*, 1647–1661. [CrossRef]
110. Mainelli, M. Blockchain will help us prove our identities in a digital world. In *Blockchain: The Insights You Need From Harvard Business Review (HBR Insights Series)*; Harvard Business Review Press: Brighton, MA, USA, 2017.
111. Parliament, E. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Off. J. Eur. Union L* **2016**, *119*, 1.
112. Al-Zaben, N.; Onik, M.M.H.; Yang, J.; Lee, N.Y.; Kim, C.S. General data protection regulation complied blockchain architecture for personally identifiable information management. In Proceedings of the 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), Southend, UK, 16–17 August 2018; pp. 77–82.
113. Fink, M. *Blockchain and the General Data Protection Regulation*; European Parliament: Bruxelles, Brussel, 2019.
114. Hasselgren, A.; Wan, P.K.; Horn, M.; Kralevska, K.; Gligoroski, D.; Faxvaag, A. GDPR Compliance for Blockchain Applications in Healthcare. *arXiv* **2020**, arXiv:2009.12913.
115. Deparday, V.; Gevaert, C.M.; Molinario, G.; Soden, R.; Balog-Way, S. *Machine Learning for Disaster Risk Management*; World Bank Group: Washington, DC, USA, 2019.
116. Biffis, E.; Chavez, E. Satellite data and machine learning for weather risk management and food security. *Risk Anal.* **2017**, *37*, 1508–1521. [CrossRef]
117. Tatar, N.; Saadatseresht, M.; Arefi, H.; Hadavand, A. A robust object-based shadow detection method for cloud-free high resolution satellite images over urban areas and water bodies. *Adv. Space Res.* **2018**, *61*, 2787–2800. [CrossRef]
118. Santos, L.B.L.; Londe, L.R.; de Carvalho, T.J.; Menasché, D.S.; Vega-Oliveros, D.A. About interfaces between machine learning, complex networks, survivability analysis, and disaster risk reduction. In *Towards Mathematics, Computers and Environment: A Disasters Perspective*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 185–215.
119. García-Martín, E.; Rodrigues, C.F.; Riley, G.; Grahn, H. Estimation of energy consumption in machine learning. *J. Parallel Distrib. Comput.* **2019**, *134*, 75–88. [CrossRef]
120. Anderson, C.C.; Denich, M.; Warchold, A.; Kropp, J.P.; Pradhan, P. A systems model of SDG target influence on the 2030 Agenda for Sustainable Development. *Sustain. Sci.* **2021**, 1–14. [CrossRef] [PubMed]
121. Sachs, J.D.; Schmidt-Traub, G.; Mazzucato, M.; Messner, D.; Nakicenovic, N.; Rockström, J. Six transformations to achieve the sustainable development goals. *Nat. Sustain.* **2019**, *2*, 805–814. [CrossRef]
122. Bester, A. Results-Based Management in the United Nations Development System: Progress and Challenges. A Report Prepared for the United Nations Department of Economic and Social Affairs, for the Quadrennial Comprehensive Policy Review. 2012. Available online: https://www.un.org/en/ecosoc/qcpr/pdf/rbm_report_2012.pdf (accessed on 27 November 2021).
123. GeSI, Deloitte. Digital with Purpose: Delivering a SMARTer2030. 2019. Available online: https://digitalwithpurpose.org/ (accessed on 5 October 2021).
124. Malhotra, C.; Anand, R.; Singh, S. Applying big data analytics in governance to achieve sustainable development goals (SDGs) in India. In *Data Science Landscape*; Springer: Singapore, 2018; pp. 273–291.
125. Centobelli, P.; Cerchione, R.; Esposito, E. Pursuing supply chain sustainable development goals through the adoption of green practices and enabling technologies: A cross-country analysis of LSPs. *Technol. Forecast. Soc. Change* **2020**, *153*, 119920. [CrossRef]
126. Mustak, M.; Salminen, J.; Plé, L.; Wirtz, J. Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda. *J. Bus. Res.* **2021**, *124*, 389–404. [CrossRef]
127. Sekiya, T.; Matsuda, Y.; Yamaguchi, K. Analysis of computer science related curriculum on LDA and Isomap. In Proceedings of the Fifteenth Annual Conference on Innovation and Technology in Computer Science Education, Ankara, Turkey, 26–30 June 2010; pp. 48–52.
128. Apiola, M.; Sutinen, E. Design science research for learning software engineering and computational thinking: Four cases. *Comput. Appl. Eng. Educ.* **2021**, *29*, 83–101. [CrossRef]

sustainability

MDPI

Perspective

A Framework for Evaluating and Disclosing the ESG Related Impacts of AI with the SDGs

Henrik Skaug Sætra

Faculty of Business, Languages, the Social Sciences, Østfold University College, N-1757 Halden, Norway; Henrik.satra@hiof.no

Abstract: Artificial intelligence (AI) now permeates all aspects of modern society, and we are simultaneously seeing an increased focus on issues of sustainability in all human activities. All major corporations are now expected to account for their environmental and social footprint and to disclose and report on their activities. This is carried out through a diverse set of standards, frameworks, and metrics related to what is referred to as ESG (environment, social, governance), which is now, increasingly often, replacing the older term CSR (corporate social responsibility). The challenge addressed in this article is that none of these frameworks sufficiently capture the nature of the sustainability related impacts of AI. This creates a situation in which companies are not incentivised to properly analyse such impacts. Simultaneously, it allows the companies that are aware of negative impacts to not disclose them. This article proposes a framework for evaluating and disclosing ESG related AI impacts based on the United Nation's Sustainable Development Goals (SDG). The core of the framework is here presented, with examples of how it forces an examination of micro, meso, and macro level impacts, a consideration of both negative and positive impacts, and accounting for ripple effects and interlinkages between the different impacts. Such a framework helps make analyses of AI related ESG impacts more structured and systematic, more transparent, and it allows companies to draw on research in AI ethics in such evaluations. In the closing section, Microsoft's sustainability reporting from 2018 and 2019 is used as an example of how sustainability reporting is currently carried out, and how it might be improved by using the approach here advocated.

Keywords: artificial intelligence; Sustainable Development Goals; ESG; CSR; reporting; disclosure

Citation: Sætra, H.S. A Framework for Evaluating and Disclosing the ESG Related Impacts of AI with the SDGs. *Sustainability* **2021**, *13*, 8503. https://doi.org/10.3390/su13158503

Academic Editors: Aimee van Wynsberghe, Larissa Bolte and Jamila Nachid

Received: 29 May 2021
Accepted: 28 July 2021
Published: 29 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Artificial intelligence (AI) permeates all aspects of modern society, and the forthcoming artificial intelligence revolution is arguably already here [1]. Both businesses and governments now implement AI systems on a large scale, both to proactively gain benefits and in fear of being left behind as others do so [2–4].

Simultaneously, businesses, civil society, politicians, and regulators are increasingly focusing on the sustainability of all human activity [5]. All major corporations are now expected to understand their environmental and social footprint and to disclose and report on related activities. This is carried out through a diverse set of standards, frameworks, and metrics related to what is referred to as ESG (environment, social, governance), which is now, increasingly often, replacing the older term CSR (corporate social responsibility) [6].

The world of AI and big tech is not exempt from these expectations, and the challenge addressed in this article is that none of the existing ESG standards, frameworks, or metrics sufficiently capture the nature of the sustainability related impacts of AI. This creates a situation in which companies are not incentivised to properly analyse and evaluate such impacts. Simultaneously, it allows the companies that are aware of negative impacts to not disclose them.

This article focuses on the need to implement insights from AI ethics research in ESG reporting, and emphasizes how the Sustainable Development Goals (SDGs) can be used to

evaluate and categorize the potential sustainability related benefits and harms related to AI [7]. The SDGs encompass environmental goals, goals related to social justice, and goals related to economic growth, health and work. This makes the SDGs useful for analyzing the sustainability of AI, which includes both how AI is used for sustainable activities, and how the use of AI might simultaneously have, for example, negative environmental impacts [8].

The outlines of a framework and a process for evaluating and disclosing ESG related AI impacts by using the SDGs are here presented, with a particular emphasis on how businesses are increasingly both expected to and required to report on and disclose such information. Examples of how the framework can force an examination of micro, meso, and macro level impacts are provided. In addition, it is shown how the framework is used to promote a consideration of both negative and positive impacts, while also accounting for ripple effects and interlinkages between the different impacts. Such a framework would make analyses of AI related ESG impacts more structured and systematic, more transparent, and allow companies to draw on research in AI ethics in such evaluations. The framework can be used by large businesses—whose primary business is the production and application of AI systems—and all others that use AI systems in parts of their activities. A comprehensive evaluation of the specific impacts of AI systems on all goals and subgoals is beyond the scope of this article, but the general approach to using the SDGs will be demonstrated through a set of examples and an examination of how Microsoft describes their AI related ESG impacts.

First of all, the concept of sustainability is examined in Section 2, with a particular focus on how it is embodied in the SDGs and how various ESG standards, frameworks, and metrics have been developed in order to help companies understand and communicate the sustainability of their activities. Secondly, in Section 3, I provide a basic account of the potential linkages between AI and environmental, social, and governance related risks and impacts in order to establish why these impacts should be taken more seriously and why it is a problem that they are partly neglected in major ESG related frameworks. Finally, the use of the SDGs in working with sustainability evaluation and disclosure is discussed in Section 4, along with the general outline of the proposed framework. In closing, selected sustainability reports from Microsoft are used to demonstrate how AI related impacts have been communicated historically, and how the framework here proposed might improve the situation.

2. Sustainability, the SDGs, and Efforts to Tackle the Sustainability of Corporations

The concept of sustainability might appear to be both intuitive and relatively straightforward. Sustainability, and the idea of sustainable development, is usually traced back to the 1987 report *Our Common Future* produced by a United Nations (UN) commission headed by Gro Harlem Brundtland. Today, sustainability in the context of business and politics is increasingly often aligned with the SDGs. These goals were presented by the UN in the document *Transforming our World: The 2030 Agenda for Sustainable Development* [7]. The framework consists of the 17 goals shown in Figure 1. With the SDGs, the notion of sustainability and a broad perspective of societal and human development on a global scale are thoroughly intertwined. This is in line with the early work on sustainable development carried out by Brundtland et al. [9]. While the SDGs are clearly related to a range of basic human rights, the two frameworks are clearly distinct, as the SDGs entail a broader focus on what is referred to as the five P's: people, planet, prosperity, peace, and partnership [7].

Figure 1. The Sustainable Development Goals [7].

While development and environmental protection might be thought of as a task for governments, the world of business and finance is increasingly playing an active role in reaching the SDGs. In part through coercion in the form of law and regulation, but also in part due to more informal processes related to the need to secure a social license to operate [6,10]. Investors, business partners, and customers now tend to demand more in terms of corporate responsibility than that which is required by law alone. This leads to a situation in which sustainability is part compliance (a "hygiene factor") and part proactive strategic business development, as sustainable business models are increasingly shown to provide businesses with an advantage in the marketplace [6,11–13]. The responsibilities of corporations to contribute to sustainable development, or at the very least to disclose how their activities have sustainability impacts, is a key part of the question examined in this article.

Corporate social responsibility (CSR) is nothing new [11], and a 40 year old quote describes fairly well the current challenges that have led to an increased focus on sustainability:

> There are several reasons for the current concern with corporate social responsibility. In recent years, the level of criticism of the business system has risen sharply. Not only has the performance of business been called into question, but so too have the power and privilege associated with large corporations. Some critics have even questioned the corporate system's ability to cope with future problems. [14] (p. 59)

Today, issues of privilege, social justice, and the societal and environmental consequences of businesses are more relevant than ever before. As a consequence, the term CSR now refers to a much broader set of issues than it did in the early ages of the concept [6]. Environmental challenges related to a range of issues, such as climate change, biodiversity loss, pollution, etc., are now universally recognized as a key challenge for humanity. In addition, as this article emphasizes the impact of AI, the unprecedented power and influence of the major technology companies attracts both attention and scrutiny. *Big Tech* describes the major tech companies, and GAFAM is one acronym describing the major US players: Google, Amazon, Facebook, Apple, and Microsoft (MS) [15]. A host of issues related to the activities of these companies are debated, and amongst the most prominent are their close to monopoly power [15], and their use of surveillance and data to monitor and increasingly exert influence over individuals and society [16]. Ethical issues more closely related to AI systems are discussed in Section 4, where it is shown how both environmental, social, and governance related risks must be mapped in order to fully understand the impact of AI.

While many focus on the three dimensions of sustainability (society, economy, and environment) when discussing the SDGs [7,17], this article focuses specifically on the

evaluation and disclosure of AI impacts, and in this context it can be useful to align the SDGs with the three aspects of ESG. Figure 2 shows a figure inspired by Berenberg [18]. The framework here presented is primarily intended as a tool for evaluating and fostering understanding of the actual impacts of AI systems, and the results from using the framework can be presented in a number of ways, as will be shown in Section 4.

Figure 2. SDG through the lens of ESG [18].

While the proposed framework might be useful for analyzing AI impacts in general, this article focuses on how corporations can analyze such impacts in order to both understand and communicate them. AI ethics research more generally deals with the same kinds of impacts, but the main challenges addressed here are the problems of making use of the insight created by such research when analyzing and reporting on corporate sustainability. It is, in part, a framework that helps translate research to a business setting.

Different types of companies in different regions all face different demands for disclosure and reporting, and in relation to disclosure requirements, ESG is the term now most often used [19]. In terms of financial regulation, for example, the Securities and Exchange Commission (SEC) in the US leaves it up to companies to determine what is material information to be disclosed, while European authorities are implementing mandatory obligations enforced by individual countries [20]. The European Union's (EU) Green Deal [21] with the sustainable finance initiative and the related taxonomy [22], which is a classification system for determining the sustainability of various economic activities, serves to illustrate how seriously regulators in Europe now take ESG related disclosure and risks. While the formal requirements for the financial sector are most developed, the trend in all sectors and businesses is clear: stakeholders of all kinds demand information about the sustainability impact and risks of business activities. Efforts to streamline and make the disclosure of such information comparable, universal, and accessible are thus increasingly relevant for all types of businesses.

As a consequence of these developments, a range of different standards and frameworks for disclosure and reporting on ESG have been developed, and this is now a landscape often described as an "alphabet soup," marred by a dizzying array of choices and few clear guidelines from regulators. Some of the major standards and frameworks mentioned in this article are the Global Reporting Initiative (GRI), the Sustainability Accounting Standards Board (SASB), the World Economic Forum's (WEF) Stakeholder Capitalism Metrics (SCM), and the SDGs. A detailed examination of the full range of standards and frameworks is beyond the scope of this chapter. Other important standards and frameworks such as the Carbon Disclosure Project (CDP) and Task Force on Climate-Related Financial Disclosures (TCFD), and more detailed examinations of GRI, SASB and the SDGs are found in [23]. These are mentioned because they are currently amongst the most popular choices

for businesses. There are many others as well, but the specifics of all these frameworks are not what is important here. What is important is that they are all insufficient in terms of incentivizing businesses to evaluate and disclose AI produced sustainability impacts.

GRI is the most widely adopted standard for preparing nonfinancial disclosures, often in the form of sustainability reports, and it can be used in combination with other frameworks [23]. While the GRI focuses on a wide range of stakeholders, the SASB focuses more specifically on investors as the target audience [23]. The WEF stakeholder metrics was designed as a unifying minimal framework that unites and draws upon a range of other frameworks [24].

Regarding the SDGs, they were, at the outset, not intended as an ESG reporting framework, but they are now increasingly used for this purpose [23]. A basic idea behind using the SDGs in such a manner is that they highlight the power of investors and businesses to engender change. The GRI, the UN Global Compact and the World Business Council for Sustainable Development have together developed the *SDG Compass* to help with the use of the SDGs for reporting (and other business oriented purposes) [25].

The SDGs are emphasized in the following, but mostly related to how they allow for improved analyses and understanding of the ESG related impacts of AI, and not as a replacement for frameworks such as GRI and SASB. This article is thus not a criticism of any particular framework or standard, but rather a call for them to be complemented by a more specific framework for analyzing how the use of AI systems, which are now ubiquitous, affect the sustainability of a company. The need for such a complementary framework arises because the other frameworks are created to account for traditional economic activity, while AI systems create certain novel challenges that require specific attention.

3. The Sustainability Impacts of AI

A growing number of sources explore the relationship between AI and the SDGs. Some attempts have been made to evaluate how AI relates to all the SDGs [17,26,27]. Others have provided more focused analyses of AI and of particular topics or specific SDGs. Some examples are research emphasizing finance related issues [28], and the technological aspects of AI [29]. Others again focus on AI and various issues of sustainability in general, without connecting this to the SDGs [30,31]. Research on sustainable business models, for example, is most often related to sustainability in general [12,13,30], while some do connect it specifically to the SDGs [3]. Efforts to examine the ethical and social implications of AI, for example work on responsible AI [32] and AI4People [33], are also clearly related to the social and governance related SDGs.

Of particular interest, however, are initiatives like the UN initiated AI4Good [34], aimed at using AI to accelerate work on the SDGs. More recently, some have used the term AI for social good (AI4SG) to describe work on AI aimed at the SDGs [35]. However, the phrase is also used by the industry and others in ways that relate it to traditional CSR and general AI ethics [36,37]. These are indications of another alphabet soup in the making, by those seeking to brand new varieties of socially responsible AI.

With AI being the main focus of the article, it is necessary to define what sort of technologies are referred to as AI. A broad and nontechnical definition is beneficial for securing that all relevant impacts are accounted for when evaluating the impact of AI systems, and Vinuesa, Azizpour, Leite, Balaam, Dignum, Domisch, Felländer, Langhans, Tegmark and Nerini [17] (p. 1) provide a description of what constitutes AI:

> ... we considered as AI any software technology with at least one of the following capabilities: perception—including audio, visual, textual, and tactile (e.g., face recognition), decision-making (e.g., medical diagnosis systems), prediction (e.g., weather forecast), automatic knowledge extraction and pattern recognition from data (e.g., discovery of fake news circles in social media), interactive communication (e.g., social robots or chat bots), and logical reasoning (e.g., theory development from premises). This view encompasses a large variety of subfields, including machine learning.

There are many potential ways to approach the evaluation of AI impacts in an ESG context, and this article highlights the usefulness of using the SDGs. Different frameworks are made for different purposes, and while these differences might at times be important, at other times the frameworks are more complementary than competitive. For example, the GRI framework is widely used for sustainability reporting, and consists of different standards, with varying specificity, with regard to the detail of disclosure and reporting [38]. The purpose the GRI standards are described as follows:

> Sustainability reporting, as promoted by the GRI Standards, is an organisation's practice of reporting publicly on its economic, environmental, and/or social impacts, and hence its contributions—positive or negative—towards the goal of sustainable development. [38] (p. 3)

This is in alignment with the purpose of this article, which is to present an approach to the evaluation of the positive and negative AI impacts to sustainable development. Similarly, the purpose of the new WEF metrics is to enable the measuring of sustainable value creation, and the metrics are explicitly linked to the GRI framework as well as to other relevant frameworks and standards [24]. However, these frameworks are designed for broad and general applications, and their origin is the need of investors to understand the risks of business. These indicators are important for generating a business level snapshot of companies, and they are intended to be generic enough for readers of the result to be able to compare different businesses. While highly useful for securing some basic and comparable basis of evaluating companies, they are not fine masked enough to ensure that all relevant AI related risks are accounted for, as discussed in more detail below.

One major benefit of the SDG framework is its origin and backing by the UN, and the fact that it is aimed at a much broader audience than the more business specific standards and frameworks. While other frameworks offer more specific guidance in the form of indicators and metrics that relate to the particular demands of regulators and, for example, the financial industry, the SDGs can be seen as the broader sustainability aspirations of the global community of business, politics, and civil society. This, and the fact that it was developed in partnership with the business and finance community, has made the SDGs a well known and widely accepted tool. As one purpose of ESG reporting is to inform and create trust between stakeholders and businesses, this is a major advantage of the SDGs.

The SDGs are also very broad, which enables us to evaluate most ethical challenges and the positive potential of AI by relating them to the SDGs. This allows for analyses where the challenges and benefits of AI systems can be identified first, followed by an analysis of how these relate to, for example, specific GRI indicators. In order to embark on their ESG reporting and disclosure journey, a company could map the various ways in which their activities positively and negatively impacted the SDGs, and when accompanied by a materiality analysis, such a preprocess would provide the companies with a more complete picture of their ESG related impacts, including those that are not naturally captured by producing a WEF or GRI report, for example.

A different approach is to start with the metrics, extract the required data (and only this), and then examine whether or not the business is in compliance with various requirements and expectations. While this can also be beneficial, it is not conducive to uncovering AI related sustainability impacts, which is why I here propose a complementary framework. The first approach allows for the examination of a broader set of questions and, not least, for uncovering unexpected and novel benefits and challenges that are not necessarily related to the various indicators found in the general ESG frameworks.

Furthermore, the more specific standards are quite restrictive as they aim to provide uniform and efficient ways to report on ESG related activities. This will, at times, have unfortunate results. For example, the Sustainability Accounting Standards Board (SASB), when discussing which factors were material to technology companies, stated that "business ethics issues are not likely to be material for the technology and communications sector" [23] (p. 26). As the discussion of AI ethics below clearly shows, business ethics is at the very core of the evaluation of AI impacts. However, the SASB Materiality

Map (https://materiality.sasb.org, accessed on 15 February 2021) does indicate that issues related to energy management, customer privacy, data security, employee engagement, diversity & inclusion, product design and lifecycle management, materials sourcing and efficiency, and competitive behavior are likely to be material for business in the technology and communications sector. These are indeed material issues for businesses developing or using AI systems, but as we will see below, an even more fine masked framework is required for capturing the less obvious potential impacts of AI.

While the SDGs are proposed as a starting point for identifying the impacts, companies can easily translate and use the results in regular sustainability reports if they choose to report according to, for example, GRI or SASB. This is both encouraged and expected, and the GRI has even released a comprehensive guide that links the GRI standard to all the SDG goals and targets [39].

4. Using the SDGs to Evaluate AI System Impacts

It now remains to examine the foundations of a framework for evaluating and reporting on sustainability related AI impacts. Some of the recent attempts to link AI and the SDGs indicate that the positive potential of AI is great [17]. This is particularly the case if one examines the various use cases that seemingly connect in some way to either the 17 SDGs or some of the targets [17,26]. In the context of ESG related impacts, both the use of AI for sustainable activities and the sustainability of AI systems are relevant [8].

AI systems enable the effective analysis Big Data and this combination can be used in a variety of ways that can potentially aid in the achievement of the SDGs. A basic example would be the implementation of AI in all stages of an enterprise resource planning system (ERP), enabling more effective production, better allocation of human resources, better financing decisions, etc. For example, AI can be used to increase energy efficiency, the effective utilization of resources, and make waste management more effective [40], or to produce product life cycle assessments by predicting energy and environmental impacts [41], all potentially leading to more sustainable businesses and economic activities more in line with circular economic principles and conducive to reduced climate gas emissions. MS, for example, emphasizes the potential of its multitiered Azure, "Power" apps and 365 ecosystems to, amongst many other things, deliver better energy efficiency, better monitoring of glaciers, and using AI to better understand climate risks [42]. Google similarly emphasizes the potential of using its Cloud and broader ecosystem of services to leverage AI for social good [43], and they have also started an "AI Impact challenge" where grants are provided to entrepreneurs using AI to achieve positive impact [44].

All types of businesses might be made more effective, but also governments, cities, and civil society might benefit from insights derived from AI analysis [45]. AI is already used in a wide array of political settings [2], as the benefits of automatic classification and prediction can make a wide range of political and bureaucratic decision-making processes more effective. According to Vinuesa, Azizpour, Leite, Balaam, Dignum, Domisch, Felländer, Langhans, Tegmark and Nerini [17], AI can enable 134 SDG targets, and AI is argued to be extremely effective at enabling SDG 1, 4, 6, 7, 9, 11, 14 and 15.

However, such analyses of AI impacts do not resonate particularly well with the current debates in the field of AI ethics, in which social, governance related, and environmental challenges are analyzed. For example, the use of AI to remedy diversity challenges by removing humans from recruitment processes [46] would be met with dire warnings from all those who emphasize the endemic nature of bias and various forms of human influence over AI systems [47–49]. Attracting most attention from AI ethicists are perhaps the social challenges, including bias in AI systems, the increased use of surveillance and a lack of privacy [50] as well as using the combination of AI and Big Data to influence and manipulate [51,52]. Such issues are arguably very difficult to map to the indicators covered by frameworks such as GRI, SASB, and WEFs metrics. This is an important shortcoming, as the importance of these issues and the risks they entail for both businesses and society are increasingly understood and accepted.

Governance, broadly understood as relating both to company governance and society's governance, involves concerns about the power of Big Tech and problems associated with platforms and the abuse of monopoly power [53–55], and, for example, the polarizing effects of algorithmic filtering and social media [56,57]. Perhaps least emphasized by the AI ethics community has been the environmental dimension. However, recent developments suggest that these issues are attracting increasing attention. For example, issues related to the carbon footprint of using AI and Big Data to train large natural language models are now problematized [8,49,58].

A primary reason for using the SDGs to evaluate AI is that they can help reconcile and bridge the potential gap that arises between business focused researchers and much of the AI ethics community. The former, at times, neglect important ethical challenges, while the latter sometimes neglects potentially important societal benefits because of the focus on the ethical challenges just discussed.

The SDGs force a broad perspective, and if AI is to be usefully evaluated in a compliance and reporting context, it is necessary to force the linkages between the E, S, and the G. The beginnings of a framework that forces the linkages and encourages a balanced perspective of AI impacts is seen in Sætra [27]. In that article, the relationship between AI and the SDGs in general is being examined, and the current article builds on this, but focuses on its application in the context of business reporting and disclosure. Sætra [27] proposed that AI impacts should be evaluated in terms of direct and indirect impacts, and that an analytical framework that distinguishes between micro, meso and macro level effects is employed in order to foster both nuance and the ability to understand the broader and long term effects of AI [27]. This framework is shown Figure 3. If an AI system aimed at improving the efficiency of workers, for example, was introduced, this could conceivably have positive meso level impacts, as the profitability and growth of a particular business would be improved (SDG 8). However, a broader analysis based on the framework presented below might show that the system had negative effects on the conditions of the individual workers (negative micro level effects on SDG 8). Furthermore, such a system—if proprietary—could lead to increased inequalities (negative macro level impacts on SDG 10).

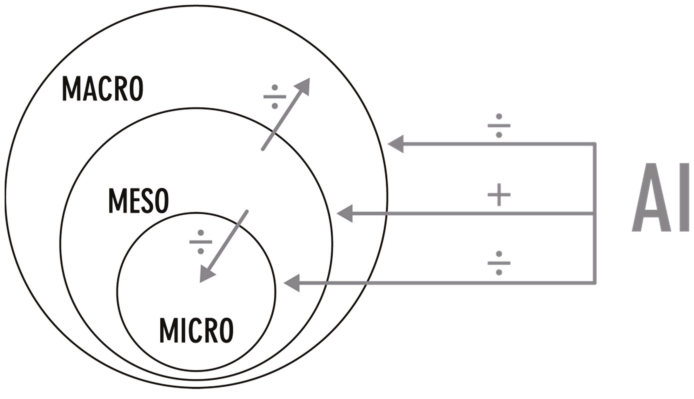


Figure 3. Three levels and the possibility of secondary effects, from Sætra [27].

4.1. The Process and the Framework

These considerations provide the basis for establishing the foundation of a framework for evaluating the sustainability related impacts of AI, and a general outline of the framework can now be established, before its potential usefulness is shown through an examination of the case of Microsoft's sustainability reporting.

Starting with one of the key elements in the framework, Figure 4 shows how the impacts for any of the SDGs can be summarized, here exemplified with SDG 8, and with sample guide questions related to this goal. The core of the framework is the complete set of such tables, in which the key guide questions and most probable interlinkages are already present in order to guide the evaluation for a specific company. The guide questions are established by mapping the issues examined in, for example, the literature of AI ethics to the various SDGs and the micro, meso, and macro levels. A wide array of sources mentioned previously in this article will be helpful in establishing these questions, and it will also be possible to develop a minimal, intermediate, and comprehensive set of questions, and also to tailor the questions to particular types of industries or applications of AI.

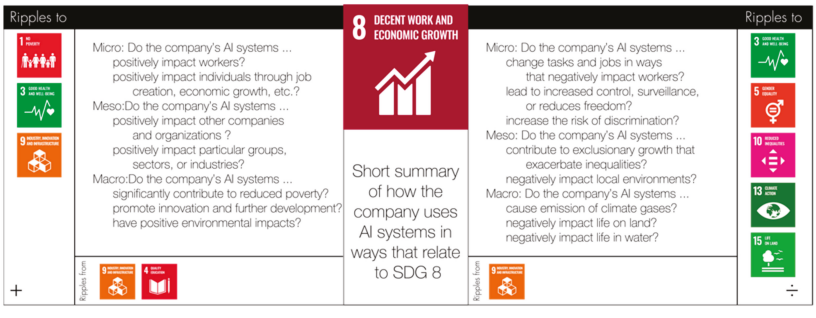


Figure 4. Table showing the key elements in the framework.

In the center, under the SDG 8 symbol, a brief summary of total impact will be provided. To the left of the center, all positive impacts are summarized and presented. In the context of SDG 8, the guide questions relate to how AI systems impact individuals and workers on the micro level, for example in terms of AI systems that monitor adherence to workers' rights and access to decent work. On the meso level, positive impacts on business growth on the business, region, or sector level are examined, but also whether particular groups of people (i.e., minorities) are positively affected by the economic growth of access to work. On the macro level, the contribution to (sustainable) economic growth is vitally important, and thus also how these impacts relate to other goals.

If the AI system contributes to economic growth that reduces poverty, provides improvements of health, innovation and infrastructure, for example, these SDGs are listed in the far left "Ripples to" column, in order to highlight interlinkages between economic growth, decent work and these other goals. In the bottom cell—"Ripples from"—the company's impact on other SDGs that have indirect effects on SDG 8 are listed (here SDG 4 and 9 are used as examples). The right side of the figure summarizes and structures the negative impacts related to SDG 8. The guide questions might here focus on increased worker surveillance and changes in work due to automation (micro), increased difference between groups and exclusionary growth (meso), and climate gas emissions resulting from the systems used (macro). These effects are related to a range of SDGs, and the potential negative indirect effects to SDGs 3, 5, 10, 13 and 15 are listed as an example.

The complete tables with guide questions are currently being developed, and the exact details of the table are not the key concern in this article. Instead, the general principle for approaching AI systems and their sustainability impacts is shown through this example. Any researcher or sustainability officer working on a sustainability report can build on this framework in order to structure their analysis and communication regarding these issues.

The complete framework can be adapted to the different needs of different businesses, and while the table above could be presented directly in the sustainability report, it could also be attached in appendices or be made available online. It need not even be published at

all, but its results could be summarized in text form. However, the benefits of a structured approach to the impacts of AI could easily be lost with such an approach.

A different question briefly broached above is what sort of process surrounds the use of this framework. Once again, the framework is amenable to a wide range of approaches, but one standard approach often used when developing a sustainability strategy and reporting approach is to start with a situation analysis followed by a materiality analysis. The situation analysis typically consists of internal and external interviews, a SWOT analysis, and a literature review, which in this case will focus specifically on AI related sustainability impacts. Next, a materiality analysis will allow the company to discover which of the SDGs (a) their stakeholders are particularly concerned about, and (b) they themselves impact most directly and to which they have the ability to adjust their impacts. Such a materiality analysis would allow the company to single out a selection of SDGs of particular importance, and thus allow them to forego a full analysis of all 17 SDGs.

4.2. AI Evaluation in Practice

In order to show how this framework can be used to improve the analysis and reporting of AI impacts, an example of actual ESG reporting is helpful. Microsoft (MS) is here used as an example of how sustainability impacts can be evaluated and disclosed. MS is selected as a case because it represents one of the larger companies in the sector and the author is familiar with their sustainability reporting. Other comparable companies (Big Tech) have been examined to ensure that MS is not particularly bad or good at sustainability reporting, and MS thus serves as a useful example of how AI impacts have been discussed, and how the framework here described might improve reporting on AI impacts. MS in itself is not of particular interest, and any comparable company could have been used to illustrate the findings here produced. The company is a key developer of AI systems, and their Azure platform is both used by MS itself and by clients in a myriad of ways. MS releases a large number of documents related to CSR and ESG related activities, but their yearly CSR reports for 2018 and 2019 are here used as the basis for the discussion. The reports are called *CSR* reports, but they could also be referred to as sustainability reports. However, these reports also contain links to and descriptions of other reports, which will be of interest for those inclined to pursue a full and comprehensive analysis of MS's efforts. For this example, the main purpose is to examine to what extent AI is discussed in relation to CSR and ESG, whether or not the SDGs are used and/or discussed, and if so: how they are discussed.

First of all, the alphabet soup discussed above is also reflected in MS's efforts to demonstrate their CSR and ESG efforts. Their online list of awards and recognitions lists 13 different sources, including FTSE4Good, The Carbon Disclosure Project, and Sustainalytics [59], and their 2018 CSR report also mentions adherence to the GRI, RAFI (UN Guiding Principles on Business and Human Rights Reporting Framework), and the UN Global Compact [60]. CSR reporting is stated to be based on a need to create trust between the company, customers, and partners [61], and these efforts are often assumed to be related to the basic idea that companies rely on such trust and a social license to operate [10].

As mentioned, there is a tendency to discuss the positive impact of AI in terms of examples of use cases, such as the use of Azure in order to improve energy efficiency in the United Arab Emirates, Marriot using AI systems to improve their water, carbon, and land change footprint, and using Azure's optical analysis functions to analyze the development of arctic glaciers. These are all examples of potential benefits from AI, and throughout the report discussing MS's contribution to the SDGs, technology is said to be an important driver of change, and examples of beneficial use are provided [42]. These examples could easily be implemented in the framework presented in Section 4.1, and instead of being presented as isolated stories, various positive and negative effects could be presented as a whole, enabling any stakeholder to more easily assess the overall AI related impacts.

MS's 2018 CSR report only superficially addresses the SDGs, and the same goes for the 2019 version, even if there, they use the SDG symbols in the chapter title pages [60,61].

However, the 2019 report refers to their sustainability website [62], and MS later released the separate report *Microsoft and the United Nations Sustainable Development Goals* [42], which deals in detail with MS's SDG related efforts. The following describes how AI impacts are discussed in these three reports.

Starting with the 2018 report [60], MS emphasizes how just about all their major products are AI infused, and thus opens the door for a broad evaluation of the impacts of AI. First of all, their products reach most people in modern societies and are an integral part of the modern workplace, in which MS 365 plays an important part in the highly visible front end, while other MS systems play important roles in the back end. Azure is the core of the back end system, and this is the core of MS's AI and the source of the infusion experienced in most or all other MS products. Dynamics 365 is key for the business applications, and it is also described as AI driven. They describe how Azure lets all their clients build AI and get the most out of their data, and that they are "democratizing data science and AI" with Azure [60].

MS states that they are leading in AI research, and that their goal is to facilitate customer adoption and innovation in all aspects of business and society. AI will, they state, create "new opportunities for education and healthcare, address poverty, and achieve a more sustainable future" [60]. They state that they champion ethical AI, and at one point they explicitly address a concern related to bias in AI systems:

> And, as we make advancements in AI, we are asking ourselves tough questions—like not only what computers can do, but what should they do. That's why we are investing in tools for detecting and addressing bias in AI systems and advocating for thoughtful government regulation. [60] (p. 12)

However, there is no description of the current status or plan for action for this work. They also state that AI "raises complex ethical questions about the impact on society including jobs, privacy, inclusiveness, and fairness" [60] (p. 41), and point to their book *The Future Computed* where AI related challenges are mentioned. In this book they discuss the role of AI in society and identify six ethical principles for the development and use of AI: fairness, reliability and safety, privacy and security, inclusivity, transparency, and accountability [63]. The above concerns and principles would benefit from being analyzed and discussed in relation to the relevant SDGs, as this would allow for a more holistic account of the impacts of MS's activities, rather than isolated discussions in which concerns and strengths are discussed without being related to the overall ESG impacts.

Similarly, they highlight the need for AI systems to be accessible, and point to work with governments, private sector and nonprofit organizations and how they donated $1,4 billion in software that year. Furthermore, they highlight their *AI for Earth* and *AI for Accessibility* programs—"putting AI tools into the hands of change-makers" [60] (p. 13). Finally, they have partnered with nonprofit organizations to provide computer science learning experiences for "millions of young people around the world" [60] (p. 13). All these issues are highly relevant for the SDGs, as many of the goals refer to the need for universal, equal, and affordable access to new technologies and innovations, and also to nondiscrimination [27]. However, the report makes no efforts to systematically evaluate the impact of AI and link these to the SDGs, and any negative impacts are mainly glimpsed through the short descriptions of philanthropic activity and vague references to investments in antibias solutions and advocacy for government regulation.

In the 2019 report, much of the same positive potential is highlighted through examples [61]. In this report, they refer to a new book named *Tools and Weapons: The Promise and the Peril of the Digital Age* [64], in which two MS directors discuss the potential challenges of AI. The six AI principles mentioned above are this year included and briefly explained, and "AI for good" is the umbrella term for a wide range of "AI for . . . " initiatives, which, in addition to the ones mentioned in the 2018 version, now includes *AI for Humanitarian Action* and *AI for Cultural Heritage*. The separate chapter on AI begins with the following statement: "We build AI responsibly, taking a principled approach to guide the development and use of artificial intelligence with people at the center of everything we do" [61].

While the CSR reports seemingly acknowledge that there are issues related to the ethical challenges of AI, there is very little in terms of discussion about the specifics of the potential harms originating in MS AI systems, and no discussion about the potential risks to MS as a company resulting from neglecting such impacts.

The report focusing specifically on MS and the SDGs is particularly telling, in that it clearly states that it is about how MS contributes to the SDGs, and that it is not an effort to evaluate MS's overall impact on the goals. Technology, it states:

> ... is a powerful force for good, and all of us here at Microsoft are working together to foster a sustainable future where everyone has access to the benefits it provides and the opportunities it creates. [42] (p. 4)

While the use of examples and sunshine stories about AI impacts, as seen in the MS reports, are not erroneous, they can be considered to be incomplete without a deeper and more comprehensive analysis of the ESG related risks of AI. For example, while the MS and SDG report highlights MS's efforts to use AI to identify and counter deepfakes, it makes no mention of how AI systems are first used to create deepfakes. Similarly, while AI systems can help conservation efforts by analyzing and tracking various animal species, for example, the same technology can be used by those who wish to hunt those same species. This is a general problem often neglected in the analysis of AI impacts. The fact that AI is a double edged sword, and that a lot of use cases demonstrating positive impacts simultaneously provides examples of how to use AI for bad, must be remembered when AI impact is described [27]. This is where the framework presented in this article helps structure and present the various effects of AI.

AI does indeed have great potential, but in order to provide stakeholders with actionable insight into the real threats and opportunities companies using AI systems face, there is a need for connecting ESG reporting more closely to AI ethics research. This, in turn, can be achieved by using the SDGs as the foundation of the analysis of AI impacts, linkages between the different SDGs, for seeing impacts at the micro, meso, and macro level in its totality, and distinguishing between direct and indirect effects [27]. The separation of effects into different levels would allow MS to show how their software has both positive and negative impacts on workers, as they link this to SDG 8 (decent work and economic growth). In the initial analysis, they might describe how their entire business suite allows for convenience and productivity for workers, and cost savings and more effective management for employers. These micro level effects can be presented as positive, and so could the meso level effects related to increased company profitability, and the macro level effects related to potential economic growth, the other aspect of SDG 8. Furthermore, their AI systems can be used to improve infrastructure and for research and innovation purposes (SDG 9), which, again, foster economic growth and demonstrate the linkages between the various goals.

However, such an analysis is incomplete without a discussion of how employers might use the 365 suite as a tool of surveillance and control, thus causing potential *negative* micro level effects for employees at the same time [65]. Similarly, their investments in anti-bias solutions indicate that they are well aware of how current AI systems are prone to bias due to a variety of sources (often related to data), and this could be accompanied by an acknowledgement that AI might have negative impacts on, for example, SDG 5, related to gender equality. However, in this context the fact that AI might also reduce bias, as humans are also prone to bias, would reasonably be included.

As has become apparent in the preceding paragraphs, AI has a wide range of potential impacts, but the overall impact is exceedingly hard to evaluate if the different impacts are presented as isolated use cases spread throughout a long sustainability report. By applying the framework here developed, we have already seen how SDG 8, for example, could be presented in the form of the table in Figure 4. We have also seen that MS has described positive AI related impacts to the monitoring of glaciers, but with the framework here proposed, such impacts would be presented along with a discussion of how the use of AI also produces emissions that ultimately leads to glacier melt and climate change. In

addition to increased transparency and honesty, which is itself beneficial, the framework is conducive to an approach in which trade offs are automatically part of any consideration of the sustainability of AI. This is because benefits and downsides are presented side by side, and in relation to each other. An example is how AI can be used to reduce human bias, while we also know that AI systems can themselves be biased in ways that are difficult to uncover. Education is another field in which MS promotes the positive impacts of AI, but we have also recently seen that the European Commission has labelled the use of AI in school a high risk application [66].

Similar considerations related to all the SDGs are presented in more detail in Sætra [27]. The next step following the work here presented is the development of the full framework built on the principles here described, based on a wide range of sources that describe various AI related sustainability impacts. By using the SDGs more actively in their ESG accounting and reporting, companies might still be able to "greenwash" and be tempted to use the SDGs more as window dressing that simply structures attempts to portray how well the company does. However, as this article has shown, ESG reporting is fundamentally about realistically communicating the positive and negative impacts of a company's activities, and the demands—both from regulators and other stakeholders—for honest and nuanced data and analyses are growing stronger by the day. The framework here presented provides an easy to use approach that forces some structure and demands a consideration of the trade offs involved in using AI systems, while it also allows the business community to make use of the knowledge produced in academia.

5. Conclusions

As AI systems permeate modern societies, a growing need to account for and understand their impacts emerge. This article has discussed various aspects related to how companies report on and disclose ESG related impacts, with a particular focus on how AI impacts might be accounted for in this context. This is increasingly important as companies of all kinds either build and develop or implement AI systems as elements in their overall activities.

One particular framework for analyzing and describing AI impacts is the SDGs. This framework has been adopted by a large number of companies, but it is usually not used in a systematic manner. This article has used MS as an example of how the SDGs are partly used in CSR and ESG reporting, but shows that there is great potential for a deeper and more comprehensive use of the SDGs. The article has also shown how to use the SDGs to analyze activities involving AI systems, with a particular emphasis on the need to go beyond simplistic analyses of how AI might relate superficially to each of the SDGs individually. For the SDGs to become a meaningful tool for analyzing AI system impacts, the interdependence between the SDGs must be factored in, and it is also beneficial to distinguish between the micro, meso, and macro level effects [27].

This article has presented the foundation of a framework to address the major challenges businesses experience related to evaluating and reporting on AI related sustainability impacts. It is not as of yet a complete framework, but while the complete framework is already under development, the core ideas here presented also allow any researcher or sustainability officer to start using the basic ideas in their own work, or to build on this and develop their own frameworks for specific (or general) purposes.

Improving the understanding of and reporting on ESG related AI impacts through the proposed framework is important for two reasons. First of all, companies might be said to have a responsibility to cause no harm, and in order to avoid causing harm they must thoroughly understand the impacts of their actions. For some companies, using the framework here proposed could lead to a deeper understanding of the sustainability of their actions. Where such an understanding has been lacking, this could in itself lead to changes in activity. Other companies might already be aware of most of their ESG related impacts, but a framework that demands more transparency and deeper analyses will potentially also affect these. While general reporting frameworks will not require

that companies disclose indirect effects and broader societal and economic effects, the framework here presented requires this. This will both deepen the understanding of the impacts and make it harder for companies to sweep the negative impacts under the rug.

Secondly, a wide range of stakeholders are becoming increasingly aware of the problems highlighted in AI ethics, and any company that does not sufficiently align their businesses with this knowledge risks jeopardizing their social license to operate. This means that even if the executives of a company themselves do not care that much about their negative impacts, stakeholders do, and will increasingly punish companies that do not perform their best to mitigate adverse impacts. Lackluster ESG performance will make capital harder to come by, it will impede the goal of attracting the most competent workers, and it will hurt the company's relation with both suppliers and customers.

One might argue that Big Tech is too big to be influenced by better frameworks for the evaluation and disclosure of AI impacts. However, the preceding considerations suggest that even such companies will be affected by an increased understanding of their negative ESG related impacts. One advantage of the framework here presented is that such an increased understanding might occur even if the larger technology companies themselves did not take it particularly seriously. The framework is generic and aimed at highlighting general ESG related impacts of AI systems, and as soon as some companies start using it, this will engender increased understanding of the impacts even of those that do not use the framework. This could lead to negative reactions from both capital markets and consumers towards those that are seen not to take these issues seriously. Even more important, perhaps, is that this will provide regulators with a better understanding of the impacts of AI systems, and it is evident from recent developments in Europe that there is an increased willingness to regulate both data and the use of AI systems more generally. Not even big tech companies are immune to the collective pressure of markets, customers, and regulators.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Makridakis, S. The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures* **2017**, *90*, 46–60. [CrossRef]
2. De Sousa, W.G.; de Melo, E.R.P.; Bermejo, P.H.D.S.; Farias, R.A.S.; Gomes, A.O. How and where is artificial intelligence in the public sector going? A literature review and research agenda. *Gov. Inf. Q.* **2019**, *36*, 101392. [CrossRef]
3. Di Vaio, A.; Palladino, R.; Hassan, R.; Escobar, O. Artificial intelligence and business models in the sustainable development goals perspective: A systematic literature review. *J. Bus. Res.* **2020**, *121*, 283–314. [CrossRef]
4. Brynjolfsson, E.; Mcafee, A. The business of artificial intelligence. *Harv. Bus. Rev.* **2017**, *7*, 3–11.
5. Walker, J.; Pekmezovic, A.; Walker, G. *Sustainable Development Goals: Harnessing Business to Achieve the SDGs through Finance, Technology and Law Reform*; John Wiley & Sons: Hoboken, NJ, USA, 2019.
6. Verbin, I. *Corporate Responsibility in the Digital Age: A Practitioner's Roadmap for Corporate Responsibility in the Digital Age*; Routledge: London, UK, 2020.
7. United Nations. *Transforming Our World: The 2030 Agenda for Sustainable Development*; Division for Sustainable Development Goals: New York, NY, USA, 2015.
8. Van Wynsberghe, A. Sustainable AI: AI for sustainability and the sustainability of AI. *AI Ethics* **2021**, 1–6. [CrossRef]
9. Brundtland, G.H.; Khalid, M.; Agnelli, S.; Al-Athel, S.; Chidzero, B. *Our Common Future*; Oxford University Press: New York, NY, USA, 1987; Volume 8.
10. Demuijnck, G.; Fasterling, B. The social license to operate. *J. Bus. Ethics* **2016**, *136*, 675–685. [CrossRef]
11. Moon, J. *Corporate Social Responsibility: A Very Short Introduction*; OUP Oxford: Oxford, UK, 2014.
12. Marczewska, M.; Kostrzewski, M. Sustainable business models: A bibliometric performance analysis. *Energies* **2020**, *13*, 6062. [CrossRef]

13. Nosratabadi, S.; Mosavi, A.; Shamshirband, S.; Kazimieras Zavadskas, E.; Rakotonirainy, A.; Chau, K.W. Sustainable business models: A review. *Sustainability* **2019**, *11*, 1663. [CrossRef]
14. Jones, T.M. Corporate social responsibility revisited, redefined. *Calif. Manag. Rev.* **1980**, *22*, 59–67. [CrossRef]
15. Petit, N. *Big Tech and the Digital Economy: The Moligopoly Scenario*; Oxford University Press: Oxford, UK, 2020; p. 11.
16. Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power: Barack Obama's Books of 2019*; PublicAffairs: New York, NY, USA, 2019.
17. Vinuesa, R.; Azizpour, H.; Leite, I.; Balaam, M.; Dignum, V.; Domisch, S.; Felländer, A.; Langhans, S.D.; Tegmark, M.; Nerini, F.F. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nat. Commun.* **2020**, *11*, 1–10. [CrossRef] [PubMed]
18. Berenberg. *Understanding the SDGs in Sustainable Investing*; Joh Berenberg, Gossler & Co. KG: Hamburg, Germany, 2018.
19. Esty, D.C.; Cort, T. (Eds.) *Values at Work: Sustainable Investing and ESG Reporting*; Palgrave McMillan: Cham, Switzerland, 2020.
20. Eckhart, M. Financial Regulations and ESG Investing: Looking Back and Forward. In *Values at Work*; Esty, D.C., Cort, T., Eds.; Palgrave McMillan: Cham, Switzerland, 2020; pp. 211–228.
21. European Commision. *The European Green Deal*; European Commision: Geneva, Switzerland, 2019.
22. EU Technical Expert Group on Sustainable Finance. *Taxonomy: Final Report of the Technical Expert Group on Sustainable Finance*; EU Technical Expert Group on Sustainable Finance: Brussels, Belgium, 2020.
23. Bose, S. Evolution of ESG Reporting Frameworks. In *Values at Work*; Esty, D.C., Cort, T., Eds.; Palgrave McMillan: Cham, Switzerland, 2020; pp. 13–33.
24. World Economic Forum. *Measuring Stakeholder Capitalism: Towards Common Metrics and Consistent Reporting of Sustainable Value Creation*; World Economic Forum: Graubunden, Switzerland, 2020.
25. SDG Compass. *SDG Compass: The Guide for Business Action on the SDGs*; Global Reporting Initiative: Amsterdam, The Netherlands, 2015.
26. Chui, M.; Harryson, M.; Manyika, J.; Roberts, R.; Chung, R.; van Heteren, A.; Nel, P. *Notes from the AI Frontier: Applying AI for Social Good*; McKinsey Global Institute: San Francisco, CA, USA, 2018.
27. Sætra, H.S. AI in context and the sustainable development goals: Factoring in the unsustainability of the sociotechnical system. *Sustainability* **2021**, *13*, 1738. [CrossRef]
28. Truby, J. Governing Artificial Intelligence to benefit the UN Sustainable Development Goals. *Sustain. Dev.* **2020**, *28*, 946–959. [CrossRef]
29. Khakurel, J.; Penzenstadler, B.; Porras, J.; Knutas, A.; Zhang, W. The rise of artificial intelligence under the lens of sustainability. *Technologies* **2018**, *6*, 100. [CrossRef]
30. Toniolo, K.; Masiero, E.; Massaro, M.; Bagnoli, C. Sustainable business models and artificial intelligence: Opportunities and challenges. In *Knowledge, People, and Digital Transformation*; Springer: Cham, Switzerland, 2020; pp. 103–117.
31. Yigitcanlar, T.; Cugurullo, F. The sustainability of artificial intelligence: An urbanistic viewpoint from the lens of smart and sustainable cities. *Sustainability* **2020**, *12*, 8548. [CrossRef]
32. Dignum, V. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*; Springer: Cham, Switzerland, 2019.
33. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef] [PubMed]
34. ITU. AI4Good Global Summit. Available online: https://aiforgood.itu.int (accessed on 31 January 2021).
35. Tomašev, N.; Cornebise, J.; Hutter, F.; Mohamed, S.; Picciariello, A.; Connelly, B.; Belgrave, D.C.; Ezer, D.; van der Haert, F.C.; Mugisha, F. AI for social good: Unlocking the opportunity for positive impact. *Nat. Commun.* **2020**, *11*, 1–6. [CrossRef]
36. Google. AI for Social Good: Applying AI to Some of the World's Biggest Challenges. Available online: https://ai.google/social-good/ (accessed on 20 February 2021).
37. Berberich, N.; Nishida, T.; Suzuki, S. Harmonizing Artificial Intelligence for Social Good. *Philos. Technol.* **2020**, *33*, 613–638. [CrossRef]
38. GRI. *Consolidated Set of GRI Sustainability Reporting Standards 2020*; GRI: Amsterdam, The Netherland, 2020.
39. GRI. *Linking the SDGs and the GRI Standards*; GRI: Amsterdam, The Netherland, 2020.
40. Nižetić, S.; Djilali, N.; Papadopoulos, A.; Rodrigues, J.J. Smart technologies for promotion of energy efficiency, utilization of sustainable resources and waste management. *J. Clean. Prod.* **2019**, *231*, 565–591. [CrossRef]
41. Kaab, A.; Sharifi, M.; Mobli, H.; Nabavi-Pelesaraei, A.; Chau, K.-w. Combined life cycle assessment and artificial intelligence for prediction of output energy and environmental impacts of sugarcane production. *Sci. Total Environ.* **2019**, *664*, 1005–1019. [CrossRef]
42. Microsoft. *Microsoft and the United Nations Sustainable Development Goals*; Microsoft: Redmond, WA, USA, 2020.
43. Google. *Environmental Report 2019*; Google: Menlo Park, CA, USA, 2019.
44. Google. Working Together to Apply AI for Social Good. Available online: https://ai.google/social-good/impact-challenge (accessed on 23 February 2021).
45. Nilashi, M.; Rupani, P.F.; Rupani, M.M.; Kamyab, H.; Shao, W.; Ahmadi, H.; Rashid, T.A.; Aljojo, N. Measuring sustainability through ecological sustainability and human sustainability: A machine learning approach. *J. Clean. Prod.* **2019**, *240*, 118162. [CrossRef]

46. Houser, K.A. Can AI Solve the Diversity Problem in the Tech Industry: Mitigating Noise and Bias in Employment Decision-Making. *Stanf. Technol. Law Rev.* **2019**, *22*, 290.
47. Noble, S.U. *Algorithms of Oppression: How Search Engines Reinforce Racism*; New York University Press: New York, NY, USA, 2018.
48. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; pp. 77–91.
49. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event Canada, 3–10 March 2021. [CrossRef]
50. Solove, D.J. Privacy and power: Computer databases and metaphors for information privacy. *Stanf. Technol. Law Rev.* **2000**, *53*, 1393. [CrossRef]
51. Yeung, K. 'Hypernudge': Big Data as a mode of regulation by design. *Inf. Commun. Soc.* **2017**, *20*, 118–136. [CrossRef]
52. Sætra, H.S. When nudge comes to shove: Liberty and nudging in the era of big data. *Technol. Soc.* **2019**, *59*, 101130. [CrossRef]
53. Culpepper, P.D.; Thelen, K. Are we all amazon primed? consumers and the politics of platform power. *Comp. Political Stud.* **2020**, *53*, 288–318. [CrossRef]
54. Gillespie, T. The politics of 'platforms'. *New Media Soc.* **2010**, *12*, 347–364. [CrossRef]
55. Sagers, C. Antitrust and Tech Monopoly: A General Introduction to Competition Problems in Big Data Platforms: Testimony Before the Committee on the Judiciary of the Ohio Senate. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3471823 (accessed on 17 October 2019).
56. Sætra, H.S. The tyranny of perceived opinion: Freedom and information in the era of big data. *Technol. Soc.* **2019**, *59*, 101155. [CrossRef]
57. Sunstein, C.R. # *Republic: Divided Democracy in the Age of Social Media*; Princeton University Press: Princeton, NJ, USA, 2018.
58. Brevini, B. Black boxes, not green: Mythologizing artificial intelligence and omitting the environment. *Big Data Soc.* **2020**, *7*, 2053951720935141. [CrossRef]
59. Microsoft. Awards & Recognition. Available online: https://www.microsoft.com/en-us/corporate-responsibility/recognition (accessed on 15 February 2021).
60. Microsoft. *Microsoft 2018: Corporate Social Responsibility Report*; Microsoft: Redmond, WA, USA, 2018.
61. Microsoft. *Microsoft 2019: Corporate Social Responsibility Report*; Microsoft: Redmond, WA, USA, 2019.
62. Microsoft. Our Commitment to Sustainable Development. Available online: https://www.microsoft.com/en-us/corporate-responsibility/un-sustainable-development-goals (accessed on 20 February 2021).
63. Microsoft. *The Future Computed: Artificial Intelligence and Its Role in Soviety*; Microsoft: Redmond, WA, USA, 2018.
64. Smith, B.; Browne, C.A. *Tools and Weapons: The Promise and the Peril of the Digital Age*; Penguin: New York, NY, USA, 2019.
65. Manokha, I. The Implications of Digital Employee Monitoring and People Analytics for Power Relations in the Workplace. *Surveill. Soc.* **2020**, *18*, 540–554. [CrossRef]
66. Europan Commision. *Europe Fit for the Digital Age: Commission Proposes New Rules and Actions for Excellence and Trust in Artificial Intelligence*; Europan Commision: Geneva, Switzerland, 2021.

sustainability

MDPI

Article

Sustainability of AI: The Case of Provision of Information to Consumers

Iakovina Kindylidi [1,2,3,*] and Tiago Sérgio Cabral [4,*]

[1] Tilburg Law School, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
[2] NOVA School of Law, 1099-032 Lisbon, Portugal
[3] Vieira de Almeida & Associados, 1200-151 Lisbon, Portugal
[4] JUSGOV-CEDU Research Centre, School of Law, University of Minho, 4710-057 Braga, Portugal
* Correspondence: imk@vda.pt (I.K.); tc@tiagosergiocabral.com (T.S.C.)

Abstract: The potential of artificial intelligence (AI) and its manifold applications have fueled the discussion around how AI can be used to facilitate sustainable objectives. However, the technical, ethical, and legal literature on how AI, including its design, training, implementation, and use can be sustainable, is rather limited. At the same time, consumers incrementally pay more attention to sustainability information, whereas businesses are increasingly engaging in greenwashing practices, especially in relation to digital products and services, raising concerns about the efficiency of the existing consumer protection framework in this regard. The objective of this paper is to contribute to the discussion toward sustainable AI from a legal and consumer protection standpoint while focusing on the environmental and societal pillar of sustainability. After analyzing the multidisciplinary literature available on the topic of the environmentally sustainable AI lifecycle, as well as the latest EU policies and initiatives regarding consumer protection and sustainability, we will examine whether the current consumer protection framework is sufficient to promote sharing and substantiation of sustainability information in B2C contracts involving AI products and services. Moreover, we will assess whether AI-related AI initiatives can promote a sustainable AI development. Finally, we will propose a set of recommendations capable of encouraging a sustainable and environmentally-conscious AI lifecycle while enhancing information transparency among stakeholders, aligning the various EU policies and initiatives, and ultimately empowering consumers.

Keywords: sustainability; artificial intelligence; sustainable AI; greenwashing; unfair commercial practices; AI Act

Citation: Kindylidi, I.; Cabral, T.S. Sustainability of AI: The Case of Provision of Information to Consumers. *Sustainability* **2021**, *13*, 12064. https://doi.org/10.3390/su132112064

Academic Editors: Tijs Vandemeulebroucke, Aimee van Wynsberghe, Larissa Bolte and Jamila Nachid

Received: 1 September 2021
Accepted: 13 October 2021
Published: 1 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The disruptive capabilities of artificial intelligence (AI) are unquestionable. Over the last years, regulators, academics, and various stakeholders have been investigating the impacts of AI on the economy, society, and the personal lives of individuals. Although the main discussion revolves around the inherent problems of machine learning and deep learning techniques, such as explainability (for an analysis of this matter from a data protection perspective see amongst others Cabral, T.S. [1]). AI biases, and privacy infringement, there is an increasing discussion not only about how AI can be used for sustainable purposes [2], but also about how the development of AI products and services can be sustainable [3].

In November 2020, the European Commission launched its New Consumer Agenda [4]. The New Consumer Agenda outlines the EU consumer policy vision until 2025 and is based on five strategic areas, among which are the *Green Transition* and the *Digital Transformation*. Both initiatives set the EU consumers in the epicenter of a green and technology-based EU-wide plan, while they predominantly focus on the importance of information to ensure higher levels of consumer protection and engagement, as well as transparency in digital

practices and environmental sustainability claims. In this regard, the Green Transition legislative proposal is expected during 2021.

Starting from these EU initiatives, we will identify the need for sustainable AI, and we will examine whether the current consumer protection regime and the latest EU policies and proposals regarding consumer protection are capable of fostering an information obligation regarding environmental sustainability in B2C contracts when AI is involved. Furthermore, we will analyze how the requirement of societal and environmental wellbeing, identified by the AI High-Level Expert Group (AI HLEG) in the Guidelines for Trustworthy AI, can promote sustainability. Considering a recent proposal for an AI Act [5], we will assess whether this new initiative can assist in materializing the sustainability objectives of the New Consumer Agenda. Lastly, we will propose a set of recommendations aiming to link the *Green Transition* objectives with the design, development, and use of sustainable AI.

In the present analysis, we will focus on consumer protection laws horizontally applicable and, therefore, we will not analyze the obligation to provide information to consumers based on sector-specific regulations applicable to particular AI products and services, for instance, medical devices in the healthcare sector.

For clarity, it should be noted that when we refer to artificial intelligence, we refer only to machine learning (including its subset deep learning AI) either embedded in hardware devices or software-based, and the terms AI or algorithm are used interchangeably. Furthermore, note that since the difference between an algorithm and a model is purely technical and not necessarily consistent across works in this field, and as it does not affect the scope of our paper to avoid constant sentences like 'the AI model created by the AI algorithm', we will frequently use the term 'algorithm' to encompass the entire procedure from learning to result

2. Materials and Methods

This paper is based on a legal doctrinal and interdisciplinary analysis, comprising theoretical and descriptive material from a legal and technological point of view. Writing of this work includes a small comparative analysis for the enforceability of the Unfair Commercial Practices Directive. All materials used are available in the references section.

3. Sustainable and Non-Sustainable AI

In 2019, a well-cited study carried out by Strubell et al. concluded that the training process of a single, deep learning, natural language processing model can lead to approximately 600,000 lb of carbon dioxide emissions [6], which roughly amounts to as much carbon dioxide emissions as the ones produced by five cars in their lifetime [7]. These problematic emissions only increase when deep neural networks are deployed on hardware platforms [6]. Considering the manifold applications of AI and the different models being developed, the environmental costs are significant.

At the same time, an increasing discussion has started in academia not only about how AI can be used for sustainable purposes, but also how the development of AI can be sustainable. For instance, it was recently reported that MIT researchers have developed a new method of deep learning training capable of reducing costs, as well as the AI training carbon footprint [8].

From the perspective of AI Ethics, Aimee van Wynsberghe defined the term sustainable AI as " ... a field of research that applies to the technology of AI (...) while addressing issues of AI sustainability and/or sustainable development" [3]. In other words, the term of sustainable AI takes into consideration the entire AI lifecycle, from training to its implementation and use. The author goes on to distinguish between two sub-concepts included under the umbrella term of sustainable AI: AI for sustainability and sustainability of AI.

AI for sustainability has been explored more over the past years. From private non-profit organizations [9] such as *AI4Good* to elaborate academic work as the AI4People ethical framework developed by Floridi et al. [2], the potential of AI to solve complicated

environmental and societal issues and help meet the United Nations Sustainable Development Goals (SDGs) and the 2030 Agenda is being promoted. On a European level, the European Commission in its White Paper on AI clearly referred to the value of AI in achieving sustainable economic growth and societal wellbeing [10], attaining the Green Deal goals [11], as well as promoting circularity in the single market in its Circular Economy Action Plan [12,13].

At the same time, the Commission, also in the AI White Paper, made a small, but nonetheless important reference to the *sustainability of AI* by referring to the importance of assessing the environmental impact of AI throughout its lifecycle and its supply chain. Interestingly, the Commission included specifically the example of "*resources usage for the training of algorithm and the storage of data*", showcasing that the European regulator is aware of the intrinsic environmental problems of AI training (the same example is mentioned in the AI HLEG Assessment List for Trustworthy Artificial Intelligence [14]). It is further suggested in the conclusion, without being defined, that AI can benefit citizens, companies, and society, when, among others, it is sustainable [10].

For the purposes of our analysis, when we use the term *sustainable AI*, we will refer only to the *sustainability of AI*, meaning the sustainable development and use of the technology, taking into consideration its environmental impact, and not when it is used to meet sustainable objectives.

It should be noted, however, that sustainability as such is not defined. In 1987, the Brundtland Commission, or the World Commission on Environment and Development (WCED), in its report, *Our Common Future*, defined sustainable development as "*development that meets the needs of the present generation without compromising the ability of future generations to meet their needs*" [15]. This definition has since been specified as to be anchored in three pillars: (i) environment, (ii) economy, and (iii) society [16].

Translating this definition, for AI to be considered sustainable throughout its lifecycle, it should not harm or otherwise impair these areas. In other words, the stakeholders involved in the designing, training, validation, verification processes, and implementation and use of AI should ensure that it serves the needs of environment, economy, and society. This is also in line with the United Nations Guidelines for Consumer Protection, and in particular regarding promoting sustainable consumption, paragraph 52 which suggests when designing, developing and using products their energy and resource efficiency should take into consideration their full lifecycle [17].

In relation to the environment (i), although not referring to AI in particular, one of the three streams of action identified in the Commission's Communication of Shaping Europe's Digital Future [18] is the promotion of an open, democratic, and sustainable society. One of the goals of this action is to reduce the carbon emissions in the digital sector. This objective will certainly impact the development of AI, considering that one of the key upcoming initiatives is achieving high energy-efficient and sustainable data centers by 2030.

Moreover, it should be noted that the scope of the Ecodesign Directive [19] covers AI-powered products. In particular, considering the broad definition of "*energy-related product*" (see article 2(1))—products which are AI-embedded—as long as they have an impact on energy consumption during their use, before entering the Single Market, they should comply with the Community ecodesign requirements set in the Directive, including bearing the CE marking. However, for the time being, this involves only devices that can function autonomously in a limited manner, as for instance, robot vacuums [20]. Although the definition of AI, as proposed by the AI HLEG and developing the 2018 Commission's definition, [10,21] and ultimately included in the Proposal for an AI Act (article 3(1) [5]), is very broad to include "*not-smart*" robots, the training and designing of such devices has limited environmental impact compared to advanced AI systems that are using machine learning and deep learning techniques. Additionally, considering the scope of the present, such products exceed the objective of our analysis.

Notwithstanding, the Sustainable Products Initiative, already announced in the Circular Economy Action Plan [12] in March 2020, will aim to address durability, reusability,

repairability, recyclability, and energy efficiency issues of various products, revising the Ecodesign Directive and extending its scope beyond energy-related products [19]. In relation to digital products, a "*digital passport*" will be developed (an initiative also announced in the EU Data Strategy). It should be noted, however, that sustainability issues of AI development are not currently contemplated in the Circular Economy Action Plan.

4. The Growing Trend towards the Green Consumer

In a survey on the *Attitudes of European citizens towards the Environment*, requested by the European Commission and the Directorate-General for Environment and published in March 2020, it was found that 53% of Europeans recognize protecting the environment as very important to them personally, and 41% as fairly important, while 68% agree that their consumption habits adversely affect the environment [22]. At the same time, the majority believes that neither the companies nor the citizens themselves are doing enough to protect the environment. Moreover, a 2018 study for the European Commission [23], aiming to provide insights on consumers' engagement in circular economy, including to sustainable consumption, found that when consumers receive adequate information on the durability of products, they also focus more on their environmental characteristics, while sufficient information on durability can almost triple the sales of products [23].

These two surveys showcase that the EU consumer is interested in participating in the circular economy, however, in doing so, it needs more information and more opportunities to actively engage in this *green transition*. Additionally, reflecting this tendency, green or environmental claims are used more and more as a marketing and advertising tool for the promotion of products and services [24].

Furthermore, the COVID-19 pandemic changed the consumption and behavioral patterns of consumers around the world and brought to the surface the need to reinforce the current consumer protection regime, especially in the context of digital transformation. It would be of interest for future policies in the area of consumer protection and circular economy to investigate the long-term effects of the pandemic in the consumption behavioral patterns of consumers to assess not only the environmental impact, such as the increase on single-use packaging waste, or the incremental online purchases, but whether following the pandemic, the interest of consumers in sustainable claims of products and services has increased or decreased. In this regard, it is also interesting to note that the Commission has announced that it plans during 2022 to explore the impact of COVID-19 on the consumption patterns of EU citizen [23].

The above were taken into consideration by the European Commission, which published in November 2020 its New Consumer Agenda following a public consultation [4]. The New Consumer Agenda outlines the EU consumer policy vision until 2025, and is based on five strategic areas: (i) the *Green Transition*, (ii) *the Digital Transformation*, (iii) redress and enforcement of consumer rights, (iv) specific needs of certain consumer groups, and (v) international cooperation. As it is understood from the context of each strategy, the agenda is following a holistic approach, addressing various existing consumer protection policies or the consumer protection aspects of other initiatives of the European Commission, as for instance, the EU Digital Strategy [15], aiming to align their objectives. For the purpose of this paper, we will focus on the first two priority areas of the Agenda.

One of the core findings of the public consultation of the New Consumer Agenda was identifying the need of consumers for "*better and more reliable information on sustainability . . . while avoiding information overload*" [4]. As it was highlighted, such information either is not available to consumers, or there is little to no reliability of the various existing environmental claims. In the screening of websites carried out by the Commission, in "*37% of cases, the claim included vague and general statements such as "conscious", "ecofriendly", "sustainable" which aimed to convey the unsubstantiated impression to consumers that a product had no negative impact on the environment*". Moreover, "*in 59% of cases the trader had not provided easily accessible evidence to support its claim*" [25]. Evidently, the obscurity of sustainability

information, or the lack of it, creates greater barriers in the decision-making process of consumers and, concomitantly, in their engagement in the circular economy.

With the Sustainable Products Initiative, the Commission wishes to increase the access of consumers to information regarding products' environmental characteristics, their durability, reparability, and reusability (as stated above). Nonetheless, to effectively change the current situation, the issues of reliability and the substantiation of sustainability claims should be tackled.

Currently, there are no specific rules on the substantiation of environmental claims. The only available tool is the Directive on Unfair Commercial Practices, which prohibits any environmental claims that are found to be misleading vis-à-vis the consumer as we will see below. However, it does not contain any specific rules on environmental claims. Nonetheless, in general, under the Unfair Commercial Practices Directive, a commercial claim is not misleading if it is presented in a clear, specific, unambiguous, and accurate manner, while the traders need to have scientific evidence available to substantiate their claims, if challenged. These criteria should be assessed on a case-by-case basis. Considering that the transposition and enforcement levels of the directive vary among member states, its application to greenwashing practices is limited and fragmented across the EU. Interestingly, already in 2013, the Commission identified in the first Impact Assessment of the Directive that "further regulation of environmental claims can only be achieved through a revision of the UCPD or the adoption of other (specific) EU legislation" [24].

As greenwashing practices increase across industries, especially in relation to digital products and services [4], the contribution of the second strategic area of the New Consumer Agenda—*Digital Transformation*—is essential. By tackling consumer protection challenges related to the use of platforms, such as fraudulent commercial practices, misinformation, and fake consumer reviews, the (truthful) information exchange and accessibility to environmental and sustainability characteristics of digital (or not) products and services can be promoted. Ultimately, the transparency of information will empower consumers, allowing them to carry out informed decisions, and increase the value and impact of sustainability claims in consumer consumption patterns.

5. Sustainability of AI and Sustainability Claims

5.1. AI HLEG Ethics Guidelines for Trustworthy AI

In April 2019, the High-Level Expert Group on Artificial Intelligence (AI-HLEG) set up by the European Commission published its final version of the *Ethics Guidelines for Trustworthy AI* following a public consultation [21]. According to the guidelines, an AI-system will be considered trustworthy when, throughout its lifecycle, it meets the following components cumulatively: (i) it is *lawful*, meaning that it complies with the applicable laws and regulations; (ii) it is *ethical*, meaning that it observes ethical principles and values, and (iii) it is technically and socially *robust*. For these components to materialize, a set of core ethical principles, as well as seven requirements based on technical and non-technical methods, should be met.

The ethical principles and requirements are identified in the Table 1 below:

Table 1. Trustworthy AI ethical principles and requirements.

Ethical Principles	Requirements
Respect for human autonomy	Human agency and oversight
Prevention of harm	Technical robustness and safety
Fairness	Privacy and data governance
Explicability	Transparency
	Diversity, non-discrimination, and fairness
	Environmental and societal well-being
	Accountability

It should be noted that AI HLEG advises that when implementing these ethical principles or "*ethical imperatives*" identified throughout the lifecycle of the technology, especially in adherence to the principle of prevention of harm, vulnerable groups and relationships where there are information asymmetries, as for instance, between businesses and consumers, should be taken into account. At the same time, one of the proposed non-technical means to facilitate meeting these requirements is information transparency. In particular, the AI HLEG suggests that providing information to stakeholders in a clear and proactive manner about the capabilities and limitations of AI, as well as of the means used to implement the seven requirements, is essential. The objective of this measure is to ensure that the stakeholders have realistic expectations about the technology.

More specifically, in order to meet the requirement of diversity, non-discrimination, and fairness, which is closely related to the ethical principle of fairness itself, aside from avoiding unfair bias and promoting stakeholder participation in AI development, accessibility and universal design is pivotal. Under this sub-requirement, it is advised that AI products and services are accessible to consumers, irrespective of their own abilities. To this, we add that the accessibility requirement does not necessarily involve only the functionality of the product or service, but also the information provided about the product or service. Information, including sustainability information and claims, should be put in a clear, legible, and accessible manner for the consumer. Therefore, overly technical and specialized vocabulary should be avoided.

In addition, the requirement of environmental and societal well-being suggests that the sustainability of AI systems should be ensured and promoted throughout the AI value chain and lifecycle. To determine whether an AI product or service is sustainable, AI HLEG suggests a critical assessment of "*resources, energy consumption during training*" [21]. In its Assessment List for Trustworthy AI (ALTAI) [14], in order to assess the conformity with the societal and environmental well-being requirement, AI HLEG proposes the following self-assessment checklist:

- "Are there potential negative impacts of the AI system on the environment?
    - Which potential impact(s) do you identify?
- Where possible, did you establish mechanisms to evaluate the environmental impact of the AI system's development, deployment and/or use (for example, the amount of energy used and carbon emissions)?
    - Did you define measures to reduce the environmental impact of the AI system throughout its lifecycle?" [14]

Notwithstanding, examples of the possible methodology or mechanisms that can be used to assess the environmental impact, or to mitigate it, are not provided.

Furthermore, although AI HLEG does not analyze or provide recommendations in relation to the lawfulness component for a trustworthy AI, these *soft law* recommendations to some extent reflect already existing legal provisions, and may influence future legislative initiatives. Especially in relation to the information transparency method, it is clear that it reflects principles embedded in various laws addressing information asymmetries such as GDPR (Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC) for the protection of data subjects, the Prospectus Regulation (Regulation (EU) 2017/1129 of the European Parliament and of the Council of 14 June 2017 on the prospectus to be published when securities are offered to the public or admitted to trading on a regulated market, and repealing Directive 2003/71/EC) for the protection of investors (, and, of course, the panoply EU consumer protection laws. In this regard, the ethical approach proposed in the guidelines is based on the fundamental rights of the EU Treaties and the Charter of Fundamental Rights of the EU ("EU Charter"). In relation to consumer protection, Article 38 of the EU Charter and Article 169 of the Treaty on the Functioning of the European Union (TFEU) are of relevance.

The guidelines and the assessment list are non-binding, and therefore non-enforceable by administrative authorities or courts. However, they can provide guidance to the various stakeholders involved in the AI lifecycle. From a consumer protection and sustainability claims standpoint, although the obligation to provide the information to the consumer lies with the trader, a transparent and facilitated flow of information between the stakeholders of the AI value chain is essential to ensure the substantiation of such claims. Especially when the provider of information is not the same as the designer, developer, or manufacturer of the AI, it is advised that information regarding the sustainable features of the product or service are addressed and supported after the product or service is put on the market and given to the trader in order to meet their own obligations vis-à-vis consumers. Such practice can be enforced contractually. Although the contractual enforcement of such an obligation to the third-party designer, developer, or manufacturer can help the trader to demonstrate, if requested by a supervisory authority or court, that substantiated information was provided to the consumers, a formal contractual relationships with the third-party that will permit contract negotiation and, secondly, a certain level of market power of the AI-trader over the third-party developer are presupposed [26]. Therefore, the scope of application of this measure may be limited in the AI-field.

5.2. Could Sustainability Information Be Included in the "Main Characteristics" of AI Products and Services?

As it was briefly mentioned in Section 4, under Article 6 of the Unfair Commercial Practices Directive, the trader cannot provide misleading information to the consumer [27]. The requirement to not mislead the consumer through untrue environmental or sustainability related claims is, of course, included in this obligation. In other words, the trader cannot engage in greenwashing practices. For instance, this is the case when a trader states that "*due to its composition, how it has been manufactured or produced, how it can be disposed of and the reduction in energy or pollution expected from its use*" a product or service will have a positive impact in the environment or a less negative impact than its competitors, without such claim being true or, at least, verifiable [28]. According to a screening conducted by the European Commission and national consumer authorities a percentage as high as 42% of market players may actually be engaging in some type of greenwashing [10].

For example, for AI, this would mean that it would not be possible to advertise a certain algorithm as trained using 100% renewable energy if, in fact, the energy had come from non-renewable sources, or if there is no adequate manner to ensure that the sources were indeed renewable. In the same manner, a trader using an AI algorithm to offer predictive maintenance to the consumer should be able to adequately substantiate any sustainability benefits (for example, related to energy consumption and waste) that they claimed to achieve, if requested.

Nonetheless, it is important to go one step further and understand that if there is margin in the current legislation to argue that in certain cases, there can be a proactive requirement to offer sustainability-related information for AI-based products and services. In this regard, both the Consumer Rights Directive [Directive 2011/83/EU of the European Parliament and of the Council of 25 October 2011 on consumer rights (as amended)], in its Articles 5 and 6, and the Unfair Commercial Practices Directive through Article 7(4) are clearly establishing that the consumer should be informed about "*the main characteristics of the goods or services, to the extent appropriate to the medium and to the goods or services*". This obligation should be interpreted in a coherent manner between both legal instruments. This should mean that a trader complying with these obligations under the Consumer Rights Directive is also complying with the obligations under the Unfair Commercial Practices Directive, as sustained by the European Commission in the Directorate-General for Justice and Consumers' Guidance Document for the application of the Consumer's Rights Directive [29] (both this Guidance document and the one related to the Unfair Commercial Practices Directive are to be updated by 2022, as announced in the New Consumer Agenda, to take into account the changes brought by the Omnibus Directive).

The question here is in fact whether in a world where consumers overwhelmingly find the environment to be important, and 57% of consumers are willing to change their purchasing habits based on sustainability considerations, 72% are willing to pay a premium for brands that promote sustainable or environmentally responsible behaviors, and 71% for brands that are fully transparent, information related to sustainability can be considered in determined cases as being part of the main characteristics of the product or service [30]. The question can be particularly relevant for AI-based products and services because as we have seen, AI can be both an engine for sustainability, but also a drain on natural resources, and consumers may want to know in which category the good or service they are buying falls in before completing the transaction. Trustworthy AI, according to the commission and AI HLEG, means both transparent and sustainable AI.

The answer to the abovementioned question is that under the instruments analyzed in this section, there is no clear legally binding requirement to provide sustainability-related information specifically for AI-based products and services. While it is certainly true that consumers are aware of sustainability in general, it would be relevant to know if they value it more when it relates to AI comparing to other characteristics to understand if it should be considered as one of AI's "main characteristics". Of course, conclusions can differ based on particular applications of the technology. For instance, if an autonomous vehicle can reduce emissions and fuel/electricity consumption by 40% due to the algorithm used for autonomous driving, one can certainly argue that this is very important information (maybe even a main characteristic of the product). On the opposite side, the fact that Gaming Console A consumes 5% less electricity than Gaming Console B due to some form of AI-based technology deployed will probably not be the key factor driving the consumer's decision to purchase.

5.3. The Commission's Proposal on AI Regulation

From late February to mid-June 2020, the European Commission ran a public consultation regarding the expected proposal for a regulation on artificial intelligence and policy options proposed in the White Paper: On Artificial Intelligence—A European approach to excellence and trust [10]. Arising from this public consultation on 21 April, the European Commission presented its proposal for an AI Act putting forward a single set of rules to regulate artificial intelligence in the European Union.

The Proposal for an AI Act came four years after the European Parliament called upon the commission to frame a legislative proposal for a set of civil law rules on robotics and artificial intelligence, arguably, the starting point of the EU's path to produce a specific AI legal instrument. It opts for a risk-based approach to AI regulation with most of its obligations being reserved for high-risk AI, and it possesses the makings of a potentially effective legal instrument with extraterritorial scope, detailed rules on market surveillance, and extremely high fines. With these characteristics, the proposal for an AI Act could have been designed in a manner that could contribute to further promoting transparency and sustainability and reinforcing consumer protection, information transparency, and fundamental rights enforceability regarding these matters. In this regard, it should be noted that the proposal for an AI Act imposes certain transparency obligations to the producer, as well as specific transparency obligations for certain AI uses. However, this information obligation focuses on the use and consequences of an AI system, and not on sustainability. Therefore, the issue of sustainability is mostly ignored in the proposal, with the exception of the possible integration of requirements related to environmental sustainability in voluntary codes of conduct (Article 69(2)) (for a more detailed assessment of the Proposal for an AI Act see Cabral and Kindylidi [31] and Cabral [32]).

In a proposal establishing requirements from risk management to data governance, and where transparency takes a central role, one cannot avoid thinking that it would be easy to go further. In fact, at least for high-risk AI systems, establishing an obligation to detail sustainability impacts in the technical documentation, and to disclose said impacts to the consumer, would not be difficult, nor would it appear out of context in the current

proposal. In addition, an important aspect that should not be disregarded is that the inclusion of rules or principles regarding environment and sustainability in the final text of the regulation, even if through a light touch principle-based approach, would mean that the EU's fundamental rights standard, based on Article 37 of the EU Charter, along with Article 3(3) TEU and 191 TFEU, will then be unquestionably applicable [33,34]. Concomitantly, this will make the likelihood of action by the Court of Justice of the European Union (ECJ) to protect and develop the "European standard" more likely.

6. Recommendations

Following our analysis, it is evident that to a large extent, AI sustainability is ignored, both in AI specific and in consumer protection legislations. Thus, it is evident that as the legislators contemplate upon the AI regulation, they should also take into consideration its impact on sustainability. Equally, new environment and sustainability legislation should take into account the specific challenges of AI. To adequately do so, a profound understanding of the various stages of the technology's lifecycle is necessary, while at the same time, incentives should be created to promote the sustainable development of the technology. For instance, measures such as requirements of environmental footprint reporting, tax benefits, and funding to entities developing and using Sustainable AI can be established in a national and European level.

Furthermore, van Wynsberghe proposes the elaboration of a "proportionality framework" at a European level in order to assess whether the environmental impact related to the training of particular AI applications is proportionate [3]. We believe that such proposal has merit, provided that larger scale studies are carried out to clearly map the environmental impact of AI and, concomitantly, the acceptable levels of energy consumption and carbon dioxide emissions per algorithmic training and use. Moreover, this also implies that the best practices and sustainability guidelines should be shared with the stakeholders, which should not burden excessively startups, SMEs, and smaller AI developers or hamper small scale training whose impact to sustainability will probably be lower.

In addition, the proportionality framework should be objective. In her opinion paper, van Wynsberghe contemplates whether the increased costs and environmental impact of AI training and *tuning* (i.e., AI repurposing or refining) may justify policy decisions that will limit certain AI-development practices for "*ethically charged tasks like recruitment of new employees or prediction of employees who may be on the verge of quitting*" [3]. The merit of the question from an ethical and societal standpoint is undeniable. However, considering the stage of the technology, we believe that such a policy decision will hinder its development and uptake. At the same time, it will prove particularly complicated for the regulator to define the criteria and design the balancing exercise of assessing which AI applications are "worthy" of their environmental impact. Furthermore, a certain AI technology that is originally developed for an ethically charged task can be later used to achieve a sustainable objective and vice-versa. Thus, such classification could end up being artificial at most. This is also true for any efforts to limit certain high-risk AI development, as defined in the proposal for an AI Act, to avoid the environmental impact. Therefore, we propose that any proportionality framework introduced should be detached from the particular application of AI and objective, especially in the training phase of its lifecycle. Notwithstanding, at the use stage, it is reasonable that different levels of energy consumption can be justified. For instance, a healthcare AI used in a hospital is expected to function for more hours, and therefore to consume more energy and have a more significant carbon footprint.

Courts can also have a role in developing their understanding related to sustainability and applying it to AI. For example, taking into consideration the growing trend of the green consumer, the ECJ could decide that, for certain AI applications, providing sustainability-related information could be required by identifying sustainability as one of the main characteristics of the goods or services. Of course, the same could be achieved through a legislative change, but that takes more time and is less efficient, as the consumer protection enforcement is fragmented in the EU and may not be even necessary in this case.

For any regulatory solutions to produce the desired effect, their uptake by entities developing and using AI should be at a scale. This requires incentivizing the industry and promoting the monitoring and ultimately the reduction of AI-energy consumption and carbon emissions as a best practice. In this regard, to the extent that sustainable development of AI cannot be achieved fully, entities developing AI should carry a cost-benefit analysis of the various available algorithms and training methods available prior to selecting one. To facilitate this process, the sharing of information between the different stakeholders about the training time [6], computational power necessary in the deployment and implementation stages, as well as the energy consumption and carbon footprint of the AI is essential. In this regard, as van Wynsberghe mentions, there are two available technical tools that can support monitoring and real-time tracking respectively of energy consumption and carbon dioxide emissions of machine and deep learning algorithms [3]: (i) the "*Machine Learning Emissions Calculator*" by Lacoste et al. [35] and the "*Carbontracker*" as suggested by Anthony et al. [36] (note that carbontracker allows the user to stop the training process "*if the predicted environmental cost is exceeded*"), and (ii) the "*Experiment-impact-tracker framework*" introduced by Henderson et al. [37].

Furthermore, we should note that interoperability is essential to achieve the sustainable use of AI, as it can promote the reusability of algorithms. Interoperability and reusability have already been identified as crucial challenges of AI-based systems by academia [38]. For interoperability and concomitantly reusability to materialize, interoperable standards should be developed. Considering that interoperability has been one of the priorities of the European Commission for the development of the Single Digital Market in Europe, and that the circular electronics initiative, addressed in the Circular Economy Actional Plan, is aiming to extend the lifecycle of electronic devices also through reusability, it is safe to assume that interoperability and reusability of algorithms will be addressed in the near future.

Nonetheless, in the European Commission's White Paper on AI, the issues of reusability and interoperability are addressed in relation to the data used in AI development [10], and a reference to this issue also appears within the Proposal for an AI Act (Recital 81), pursuant to the FAIR data management principles [39]. In addition, although not expressly referring to AI applications, the principles of the EU Interoperability Framework, including the principle of reusability, should apply to AI-powered digital public services in the EU [40]. Moreover, from a practical standpoint, it should be noted that the *once-for-all* model developed by MIT researchers builds not only on the idea of reduction in energy consumption, but also in reusability [8], since it allows the training and development of an algorithm that can be later adapted to the "*diverse hardware platforms without retraining*" [41].

7. Conclusions

As more consumers are interested in and require more information on the sustainability features of products and services, and as the literature around Sustainability of AI increases, there is a need to promote a sustainable AI lifecycle.

Considering that at the time of writing, the sustainability of AI is overlooked in (i) AI-specific legislation and initiatives, such as the recent proposal for an AI Act, (ii) Consumer Protection legislations and initiatives, including the latest New Consumer Agenda, (iii) as well as in sustainability focused laws and initiatives such as the Circular Economy Action Plan, we believe that some targeted and carefully outlined rules, setting general requirements, regulatory guidance, and codes of conduct when it comes to Sustainable AI may be needed. Although a horizontal set of rules will be the first step, considering the unique characteristics of certain industries, products, and services, sector-specific guidance will be necessary. Similarly, although we suggest that AI-powered products and services are contemplated in the commission's holistic initiatives, guidance focused on AI is necessary. As the proposal for an AI Act is under discussion, there is still time to address AI sustainability therein.

Notwithstanding, it should be noted that the uniqueness and particularities of AI do not exclude the application of existing tools that can help promote AI sustainability as well as providing sufficient information to consumers. In this regard, an update of the existing labelling system pursuant to the Ecodesign Directive and the Regulation for Energy Labelling Framework (Regulation (EU) 2017/1369 of the European Parliament and of the Council of 4 July 2017 setting a framework for energy labelling and repealing Directive 2010/30/EU) to cover more complex AI-powered products and services is advised. Parallelly, following sufficient information exchange with the industry, specific criteria can be outlined for AI products and services to be awarded the EU Ecolabel (Regulation (EC) No 66/2010 of the European Parliament and of the Council of 25 November 2009 on the EU Ecolabel). The EU Ecolabel can be awarded to any goods and services distributed, consumed or used in the EU market that have a lower environmental impact than other products in the same group. Note, however, that medical devices are excluded from the EU Ecolabel system and as such AI medical devices cannot bear the label. To increase the efficiency of the labelling systems as well as its adoption by the industry, any initiatives, horizontal or sector-specific, should be aligned.

Lastly, as it was highlighted throughout the present, any efficient policy initiative on sustainable AI requires the support and collaboration of the AI ecosystem. As a starting point, and to ensure that there is sufficient information exchange amongst the various stakeholders and that the information asymmetries between the industry and the EU regulators are bridged, further multidisciplinary research in the area of AI Sustainability should be carried. In this regard, firstly, larger scale studies to clearly map the environmental impact of AI and the acceptable levels of energy consumption and carbon dioxide emissions per algorithmic training and use are necessary. Additionally, from a consumer protection perspective, an EU-wide study to assess whether consumers consider environmental sustainability information of AI products as material is essential to encourage and accelerate possible regulatory actions towards Sustainability of AI.

In the Table 2 below, we provide a systematic overview of the main issues identified, the recommendations proposed in this paper, and the legal instruments that can facilitate such recommendations, where relevant.

Table 2. Overview of issues, recommendations and relevant regulations and policies.

Issues	Recommendations	Relevant Regulations and Policies
Address the current regulatory gap on Sustainability of AI in AI regulation and policy initiatives	• Including environmental and societal sustainability obligations in the final wording of the AI Act • Development of an objective proportionality framework for the training phase and specific standards for the use phase of AI	Proposal for AI Act
	• Promote sustainability using the AI HLEG requirement of societal and environmental wellbeing • AI accessibility should also pertain to accessibility to information, including sustainability information • Providing guidance to stakeholders and influencing future legislative initiatives	• AI HLEG Ethics Guidelines for Trustworthy AI • AI HLEG Assessment List for Trustworthy AI
	• Reducing carbon emissions in the EU digital sector • High energy-efficient and sustainable data centres by 2030	Shaping Europe's Digital Future
Address the current regulatory gap on Sustainability of AI in Consumer Protection Laws		• New Consumer Agenda • Green Transition Proposal

Table 2. *Cont.*

Issues	Recommendations	Relevant Regulations and Policies
Information obligation regarding environmental sustainability in B2C AI products	Providing further guidance and development of case law on substantiation of environmental sustainability claims in AI products to ensure harmonization and effective enforceability	Unfair Commercial Practices Directive
	Providing further guidance and development of case law on whether environmental sustainability claims consist main characteristics of AI products and services	Consumer Rights Directive
	EU-wide study to assess whether consumers consider environmental sustainability information of AI products as material	
Address the current regulatory gap on Sustainability of AI in Environmental Sustainability Laws	• Extending scope beyond energy-related products • Broadening scope of digital products to include machine and deep learning AI	• Circular Economy Action Plan • Ecodesign Directive • Sustainable Products Initiative • Regulation for Energy Labelling Framework
Promotion of Sustainability of AI in the AI ecosystem	• Incentives (e.g., environmental footprint reporting, tax benefits and funding) • New case law on AI Sustainability • AI Sustainability guidelines and best practices • Larger scale studies to clearly map the environmental impact of AI and the acceptable levels of energy consumption and carbon dioxide emissions per algorithmic training and use	
	Interoperable standards	• Single Digital Market • Circular Economy Actional Plan • White Paper on AI • Proposal for AI Act • EU Interoperability Framework

Author Contributions: Both authors contributed to every section of this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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.

References

1. Cabral, T.S. AI and the Right to Explanation: Three Legal Bases under the GDPR. In *Computers, Privacy and Data Protection Conference 2020*; Hart Publishing: Oxford, UK, 2020; pp. 29–55.
2. Floridi, L.C. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef] [PubMed]
3. van Wynsberghe, A. Sustainable AI: AI for Sustainability and the Sustainability of AI. *AI Ethics* **2021**, *1*, 213–218. [CrossRef]
4. European Commission. New Consumer Agenda Strengthening Consumer Resilience for Sustainable Recovery. In *Communication from the Commission to the European Parliament and the Council*; 52020DC0696; European Commission: Brussels, Belgium, 2020.
5. European Commission. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. In *Proposal for a Regulation of the European Parliament and of the Council*; 52021PC0206; European Commission: Brussels, Belgium, 2021.

6. Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning. In Proceedings of the NLP; 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July 2019.
7. Hao, K. MIT Technology Review, Training a Single AI Model Can Emit as Much Carbon as Five Cars in Their Lifetimes. Available online: https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/#:~{}:text=They%20found%20that%20the%20process,suspected%20for%20a%20long%20time (accessed on 10 October 2021).
8. Cai, H.; Gan, G.; Zhang, Z.; Han, S. *Once-For-All: Train One Network and Specialize It for Efficient Deployment*; ICLR: New Orleans, LA, USA, 2019.
9. Ellen MacArthur Foundation Artificial Intelligence and the Circular Economy—AI as a Tool to Accelerate the Transition. Available online: https://www.mckinsey.com/~{}/media/mckinsey/business%20functions/sustainability/our%20insights/artificial%20intelligence%20and%20the%20circular%20economy%20ai%20as%20a%20tool%20to%20accelerate%20the%20transition/artificial-intelligence-and-the-circular-economy.pdf (accessed on 5 October 2021).
10. European Commission. *White Paper On Artificial Intelligence—A European Approach to Excellence and Trust*; European Commission: Brussels, Belgium, 2020.
11. European Commission. The European Green Deal. In *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and The Committee of the Regions*; 52019DC0640; European Commission: Brussels, Belgium, 2019.
12. European Commission. A New Circular Economy Action Plan—For a Cleaner and More Competitive Europe. In *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and The Committee of the Regions*; 52020DC0098; European Commission: Brussels, Belgium, 2020.
13. Vinuesa, R.A.; Azizpour, H.; Leite, I.; Balaam, M.; Dignum, V.; Domisch, S.; Felländer, A.; Langhans, S.D.; Tegmark, M.; Nerini, F.F. The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. *Nat. Commun.* **2020**, *11*, 1–10. [CrossRef] [PubMed]
14. High-Level Expert Group on Artificial Intelligence, The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment. 2020. Available online: https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence (accessed on 10 October 2021).
15. World Commission on Environment and Development. *Our Common Future*; Oxford University Press: Oxford, UK, 1987.
16. Mensah, J. Sustainable Development: Meaning, History, Principles, Pillars, and Implications for Human Action: Literature Review. *Cogent Soc. Sci.* **2019**, *5*, 1653531. [CrossRef]
17. United Nations Guidelines for Consumer Protection. 2016. Available online: https://unctad.org/system/files/official-document/ditccplpmisc2016d1_en.pdf (accessed on 5 October 2021).
18. European Commission. *Shaping Europe's Digital Future*; European Commission: Brussels, Belgium, 2020.
19. The European Parliament and the Council of the European Union. *Directive 2009/125/EC of the European Parliament and of the Council of 21 October 2009 Establishing a Framework for the Setting of Ecodesign Requirements for Energy-Related Products*; 32009L0125; The European Parliament and the Council of the European Union: Strasbourg, France, 2009.
20. European Commission. *Commission Regulation (EU) No 666/2013 of 8 July 2013 Implementing Directive 2009/125/EC with Regard to Ecodesign Requirements for Vacuum Cleaners*; 32013R0666; European Commission: Brussels, Belgium, 2013.
21. High-Level Expert Group on Artificial Intelligence, AI Ethics Guidelines for Trustworthy AI. 2019. Available online: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed on 5 October 2021).
22. Eurobarometer. Special Eurobarometer 501, Attitudes of European Citizen towards the Environment. 2020. Available online: https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/survey/getSurveydetail/instruments/special/surveyky/2257 (accessed on 3 October 2021). otential impact(s) d.
23. European Commission. *A Behavioral Study on Consumer's Engagement in the Circular Economy*; European Commission: Brussels, Belgium, 2018.
24. European Commission. *First Report on the Application of Directive 2005/29/EC Concerning Unfair Business-to-Consumer Commercial Practices in the Internal Market 2013*; 52013DC0139; European Commission: Brussels, Belgium, 2013.
25. European Commission. Screening of Websites for 'Greenwashing': Half of Green Claims Lack Evidence. 2021. Available online: https://ec.europa.eu/commission/presscorner/detail/en/IP_21_269 (accessed on 1 October 2021).
26. Kindylidi, I.; Antas de Barros, I. AI Training Datasets & Article 14 GDPR: A risk assessment for the proportionality exemption of the obligation to provide information. *Law State Telecommun. Rev.* **2021**, *13*, 1–27. [CrossRef]
27. Carvalho, J.M. *Direito do Consumo*, 7th ed.; Almedina: Coimbra, Portugal, 2020; ISBN 9789724088921.
28. European Commission. Commission Staff Working Document Guidance on the Implementation/Application of Directive 2005/29/EC on Unfair Commercial Practices Accompanying the Document Communication from the Commission to the European Parliament, the Council, the European Economic. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52016SC0163&from=EN (accessed on 25 May 2016).
29. European Commission. DG JUSTICE Guidance Document Concerning Directive 2011/83/EU of the European Parliament and of the Council of 25 October 2011 on Consumer Rights, Amending Council Directive 93/13/EEC and Directive 1999/44/EC of the European Parliament and of the Council a. 2014. Available online: https://ec.europa.eu/info/sites/info/files/crd_guidance_en_0_updated_0.pdf (accessed on 1 October 2021).

30. Haller, K.; Lee, J.; Cheung, J. Meet the 2020 Consumers Driving Change: Why Brands Must Deliver on Omnipresence, Agility, and Sustainability. Available online: https://www.ibm.com/downloads/cas/EXK4XKX8 (accessed on 3 October 2021).
31. Cabral, T.S.; Kindylidi, I. WhatNext.Law, Proposal for a Regulation on a European Approach for Artificial Intelligence: An Overview. Available online: https://whatnext.law/2021/05/05/proposal-for-a-regulation-on-a-european-approach-for-artificial-intelligence-an-overview-pt/ (accessed on 3 October 2021).
32. Cabral, T. EU Law Live, The Proposal for an AI Regulation: Preliminary Assessment. Available online: https://eulawlive.com/oped-the-proposal-for-an-ai-regulation-preliminary-assessment-by-tiago-sergio-cabral/ (accessed on 14 October 2021).
33. Cabral, T.S.; Silveira, A.; Abreu, J. UNIO EU Law Journal, The "mandatory" contact-tracing App "StayAway COVID"—A matter of European Union Law. Available online: https://officialblogofunio.com/2020/10/20/the-mandatory-contact-tracing-app-stayaway-covid-a-matter-of-european-union-law/ (accessed on 2 October 2021).
34. Vilaça, J.L.C.; Silveira, A. The European federalisation process and the dynamics of fundamental rights. In *Citizenship within the EU Federal Context*; Cambridge University Press: Cambridge, UK, 2017; pp. 125–146.
35. Lacoste, A.; Luccioni, A.; Schmidt, V.; Dandres, T. Quantifying the Carbon Emissions of Machine. 2019. Available online: https://arxiv.org/pdf/1910.09700.pdf (accessed on 7 October 2021).
36. Anthony, L.F.W.; Kandig, B.; Selvan, R. *Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models*; ICML: Vienna, Austria, 2020. Available online: https://arxiv.org/pdf/2007.03051.pdf (accessed on 5 October 2021).
37. Henderson, P.; Hu, J.; Romoff, J.; Brunskill, E.; Jurafsky, D.; Pineau, J. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *J. Mach. Learn. Res.* **2020**, *21*, 1–43.
38. Alonso, E. Actions and Agents. In *The Cambridge Handbook of Artificial Intelligence*; Frankish, K., Ramsey, W.M., Eds.; Cambridge University Press: Cambridge, UK, 2014; pp. 232–246. [CrossRef]
39. European Commission Expert Group on Fair Data, Turning Fair into Reality. Available online: https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf (accessed on 8 October 2021).
40. European Commission. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, European Interoperability Framework—Implementation Strategy, COM(2017) 134 final. Available online: https://eur-lex.europa.eu/resource.html?uri=cellar:2c2f2554-0faf-11e7-8a3501aa75ed71a1.0017.02/DOC_1&format=PDF (accessed on 23 March 2017).
41. Matheson, R. MIT News, Reducing the Carbon Footprint of Artificial Intelligence. Available online: https://news.mit.edu/2020/artificial-intelligence-ai-carbon-footprint-0423 (accessed on 6 October 2021).

sustainability

MDPI

Article

Unraveling the Hidden Environmental Impacts of AI Solutions for Environment Life Cycle Assessment of AI Solutions

Anne-Laure Ligozat [1,*], Julien Lefevre [2], Aurélie Bugeau [3] and Jacques Combaz [4]

[1] Université Paris-Saclay, CNRS, ENSIIE, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400 Orsay, France
[2] Aix Marseille Univ., CNRS, INT, Inst Neurosci Timone, 13005 Marseille, France; julien.lefevre@univ-amu.fr
[3] Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR5800, 33400 Talence, France; aurelie.bugeau@u-bordeaux.fr
[4] Univ. Grenoble Alpes, CNRS, Grenoble INP, VERIMAG, 38000 Grenoble, France; jacques.combaz@univ-grenoble-alpes.fr
* Correspondence: anne-laure.ligozat@lisn.upsaclay.fr

Abstract: In the past ten years, artificial intelligence has encountered such dramatic progress that it is now seen as a tool of choice to solve environmental issues and, in the first place, greenhouse gas emissions (GHG). At the same time, the deep learning community began to realize that training models with more and more parameters require a lot of energy and, as a consequence, GHG emissions. To our knowledge, questioning the complete net environmental impacts of AI solutions for the environment (AI for Green) and not only GHG, has never been addressed directly. In this article, we propose to study the possible negative impacts of AI for Green. First, we review the different types of AI impacts; then, we present the different methodologies used to assess those impacts and show how to apply life cycle assessment to AI services. Finally, we discuss how to assess the environmental usefulness of a general AI service and point out the limitations of existing work in AI for Green.

Keywords: artificial intelligence; sustainability; carbon footprint; LCA

Citation: Ligozat, A.-L.; Lefevre, J.; Bugeau, A.; Combaz, J. Unraveling the Hidden Environmental Impacts of AI Solutions for Environment Life Cycle Assessment of AI Solutions. *Sustainability* **2022**, *14*, 5172. https://doi.org/10.3390/su14095172

Academic Editors: Aimee van Wynsberghe, Larissa Bolte, Jamila Nachid and Tijs Vandemeulebroucke

Received: 28 February 2022
Accepted: 15 April 2022
Published: 25 April 2022

1. Introduction

In the past few years, the AI community has begun to address the environmental impacts of deep learning programs: Ref. [1] highlighted the impacts of training NLP models in terms of energy consumption and in terms of carbon footprint, Ref. [2] proposed the concept of Green AI, and the AI community created several tools to evaluate machine learning energy consumption [3–6].

These impacts are mainly expressed in terms of energy consumption and associated greenhouse gas (GHG) emissions. Yet, as we will discuss later, this energy consumption represents only a part of the complete environmental impacts of such methods. Ref. [7], for example, states that "it is in terms of their indirect effects on the global digital sector that AI systems will have a major impact on the environment". In the same spirit, Ref. [8] warns that "optimising actions for a restricted set of parameters (profit, job security, etc) without consideration of these wider impacts can lead to consequences for others, including one's future self as well as future generations".

Evaluating the impacts of an AI service is not fundamentally different from doing it for another digital service. However, AI presents specificities that must be taken into account because they increase its environmental impacts.

First, AI—and in particular, deep learning—methods usually require large quantities of data. These data have to be acquired, transferred, stored, and processed. All these steps require equipment and energy and have environmental impacts. In the case of a surveillance satellite, the data will probably be in large quantities, but the number of

acquisition devices may be limited; in the case of a smart building infrastructure, the data may be in smaller quantities, but many devices will be required.

Training deep neural models also takes a lot of computation time and resources, partly because the model itself learns a comprehensive representation that enables it to better analyze the data. Whereas with other models, a human will provide part of this information, often in the form of a handcrafted solution. The computation cost can be even higher if the model does continuous learning.

At the same time, AI's popularity is increasing, and AI is often presented as a solution to environmental problems with AI for Green proposals [9–11]. The negative environmental impacts can be briefly evoked—and in particular, rebound effects [9,12] where unitary efficiency gains can lead to global GHG increase—but no quantification of all AI's environmental costs is proposed to close the loop between AI for Green and Green AI. That is why it is even more important to be able to assess the actual impacts, taking into account both positive and negative effects.

Incidentally, those works often use the term AI to actually refer to deep learning methods, even though AI has a much wider scope with at least two major historical trends [13]. In this paper, we will also focus on deep learning methods, which pose specific environmental issues and, as we have seen, are often presented as possible solutions to environmental problems. We describe these impacts and discuss how to take them into account.

Our contributions are the following:

- We review the existing work to assess the environmental impacts of AI and show their limitations (Sections 2.1 and 2.2).
- We present life cycle assessment (Section 2.3) and examine how it can comprehensively evaluate the direct environmental impacts of an AI service (Section 3).
- We discuss how to assess the environmental value of an AI service designed for environmental purposes (Section 4).
- We argue that although improving the state of the art, the proposed methodology can only show the technical potential of a service, which may not fully realize in a real-life context (Section 5).

2. Related Work

This section reviews existing tools for evaluating environmental impacts of AI as well as green applications of AI. It ends with an introduction to life cycle assessment, a well-founded methodology for environmental impact assessment but still not used for AI services.

2.1. Carbon Footprint of AI

Strubell et al. [1] has received much attention because it revealed a dramatic impact of NLP algorithms in the training phase—the authors found GHG emissions to be equivalent to 300 flights between New York and San Francisco. Premises of such an approach were already present in [14] for CNN with less meaningful metrics (energy per image or power with no indications on the global duration).

In [2], the authors observed a more general exponential evolution in deep learning architecture parameters. Therefore, they promoted "Green AI" to consider energy efficiency at the same level as accuracy in training models and recommend, in particular, to report floating-point operations. Other authors [15] have also reviewed all the methods to estimate energy consumption from computer architecture. They distinguish between different levels of description, software/hardware level, instruction/application level, and they consider how those methods can be applied to monitor training and inference phases in machine learning.

In the continuity of [1,2], several tools have been proposed to make the impacts of training models more visible. They can be schematically divided into

- Integrated tools, such as Experiment Impact Tracker (https://github.com/Breakend/experiment-impact-tracker, accessed on 27 February 2022) [4], Carbon Tracker (https://github.com/lfwa/carbontracker, accessed on 27 February 2022) [3], and CodeCarbon (https://codecarbon.io/, accessed on 27 February 2022), which are all Python packages reporting measured energy consumption and the associated carbon footprint.
- Online tools, such as Green Algorithms (http://www.green-algorithms.org/, accessed on 27 February 2022) [6] and ML CO2 impact (https://mlco2.github.io/impact/#compute, accessed on 27 February 2022) [5], which require only a few parameters, such as the training duration, the material, and the location but are less accurate.

AI literature mostly addresses a small part of direct impacts and neglects production and end of life, thus not following recommendations such as [16]. In [12,17], the authors point out the methodological gaps of the previous studies, focusing on the use phase. In particular, manufacturing would account for about 75% of the total emissions of Apple or of an iPhone 5, just to give examples of various scales. Their study is based on a life cycle methodology, relying on sustainability reports with the GHG protocol standard. Ref. [18] provides a list of the carbon emission sources of an AI service, which gives a more comprehensive view of the direct impacts in terms of carbon footprint only. Ref. [19] also advocates the need for taking indirect impacts (e.g., behavioral or societal changes due to AI) into account when evaluating AI services.

Some works focus on optimizing the AI processes regarding runtime, energy consumption, or carbon footprint. For example, in [20], the authors update the results from [1] and reveal a considerable reduction of the GHG impact—by a factor of 100—if one considers the location of the data center used for training (low-carbon energy) and the architecture of the deep network (sparsity). Nevertheless, as they recognize, their study evaluates the GHG emissions of operating computers and data centers only and limits the perimeter by excluding the production and the end-of-life phases of the life cycle. Their work also considers a highly optimized use case, which may not be representative of real case scenarios. The energy efficiency of machine learning has also been the subject of dedicated workshops, such as the Workshop on Energy Efficient Machine Learning and Cognitive Computing (https://www.emc2-ai.org/virtual-21, accessed on 27 February 2022).

2.2. AI for Green Benefits

When designing an AI for Green method, i.e., a method using AI to reduce energy consumption or to benefit other environmental indicators, complete AI's impacts should also be considered to build meaningful costs/benefits assessments. Ref. [21] proposes a framework for such cost–benefit analysis of AI foundation models to evaluate environmental and societal trade-offs. We discuss this framework in Section 4. Most AI solutions for the environment lack a rigorous evaluation of the cost/benefit balance, and one of our contributions is to advance this issue.

2.3. Life Cycle Assessment

LCA is a widely recognized methodology for environmental impact assessment, with ISO standards (ISO 14040 and 14044) and a specific methodology standard for ICT from ETSI/ITU [16]. It quantifies multiple environmental criteria and covers the different life cycle phases of a target system. Ref. [22] clearly states that "to avoid the often seen problem shifting where solutions to a problem creates several new and often ignored problems, these decisions must take a systems perspective. They must consider [...] the life cycle of the solution, and they need to consider all the relevant impacts caused by the solution." The LCA theoretical approach exposed in [23] describes the system of interest as a collection of building blocks called *unit processes*, for example "Use phase of the server" on which the model is trained. The set of all unit processes is called the *technosphere*, as opposed to the *ecosphere*. Each unit process can be expressed in terms of *flows* of two kinds:

- *Economic flows* are the directed links between the unit processes or said differently exchanges inside the technosphere.

- *Environmental flows* are the links from the biosphere to the technosphere or vice versa.

The detailed description of such a system is called the life cycle inventory (LCI) and it can be formulated in terms of linear algebra. The goal of a life cycle assessment consists in computing the sum of the environmental flows of the system associated with a *functional unit*. To be concrete, if one considers a heating system in a smart building, the functional unit could be "heating 1 m^2 to 20 °C for one year".

Of course, very often, the LCI does not correspond exactly to the functional unit. The size of economic flows may not match (e.g., the functional unit may partially use shared servers and sensors), and a process may be *multifunctional*, i.e., producing flows of different types at the same time (e.g., storage capacity and computational power). Both these problems can be solved using, for instance, *allocation* methods according to a *key*. A typical allocation key for network infrastructures would be the volume of data. For a data center, it could be the economic value of storage and computational services when they cannot be physically isolated.

Even though LCA is widely used in many domains, it has rarely been applied to AI services.

3. Life Cycle Assessment of an AI Solution

When it comes to quantifying the impacts of digital technologies and, in particular, AI technologies, one faces several methodological choices that deserve a specific definition of the studied system. For instance, assessing the global impacts of the AI domain—if we could circumscribe it precisely—is not the same as assessing the impacts of an AI algorithm or service. The emerging field of AI's impacts quantification still suffers from a lack of common methodology, and, in particular, it very often focuses only on the Use phase of devices involved in an AI service. To perform meaningful quantification, we strongly suggest following the general framework of life cycle assessment (LCA, detailed in Section 3.2). We will show how it can be adapted to an AI service, i.e., in this case, a deep learning code used either alone or in a larger application.

With AI being part of the Information and Communication Technology (ICT) sector, and following the taxonomies from [24,25], its impacts can be divided into first-, second-, and third-order impacts. In this section, we focus only on first-order impacts, while we will discuss second and third orders in Sections 4 and 5.

We will use the term *AI service* for all the equipment (sensors, servers, etc.) used by the AI and the term *AI solution* for the complete application using AI. In the case of the smart building, the *AI solution* is the smart building itself, while the *AI service* is the digital equipment needed for the smart infrastructure.

3.1. First-Order Impacts of an AI Service

First-order—or *direct*—impacts of the AI service are the impacts due to the different life cycle phases of the equipment:

- *Raw material extraction*, which encompasses all the industrial processes involved in the transformation from ore to metals;
- *Manufacturing*, which includes the processes that create the equipment from the raw material;
- *Transport*, which includes all transport processes involved, including product distribution;
- *Use*, which includes mostly the energy consumption of equipment while it is being used;
- and *End of life*, which refers to the processes to dismantle, recycle, and/or dispose of the equipment.

For simplicity reasons, we will merge the first three items into a single *production* phase in the rest of the paper.

For example, an AI solution in a smart building may need sensors and servers that require resources and energy for their production, operation, and end of life.

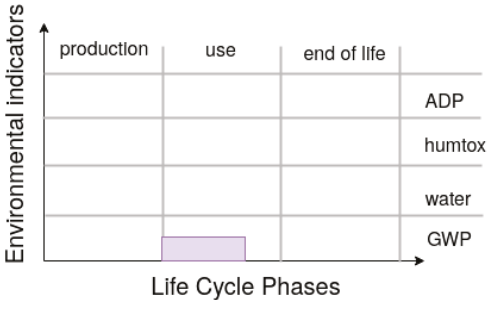


Figure 1. LCA dimensions: the first dimension corresponds to the phases of life cycle and the second one to the environmental impacts (see Section 3 for more details on this last dimension).

A second dimension is necessary to assess the impacts, with a set of environmental criteria considered. Indeed, each life cycle phase has impacts on different environmental indicators: Greenhouse Gases emissions (usually expressed as Global Warming Potential, GWP), water footprint, human toxicity, or abiotic resource depletion (ADP) for instance. In general, evaluating the environmental impact of a service requires multiple impact criteria [16]. ISO states that "the selection of impact categories shall reflect a comprehensive set of environmental issues related to the product system being studied, taking the goal and scope into consideration". Additionally, "the selection of impact categories, indicators and models shall be consistent with the goal and scope of the LCA study". Hence, the costs must take into account at least the criteria that are supposed to be tackled by the AI solution in the case of AI for Green—if the AI solution is applied to reduce energy consumption, for example, the main expected gain will probably be in terms of carbon footprint, so at least the carbon footprint of using the model should be considered. For an application monitoring biodiversity, the most relevant criterion may be natural biotic resources (and not carbon footprint), which includes wild animals, plants, etc.

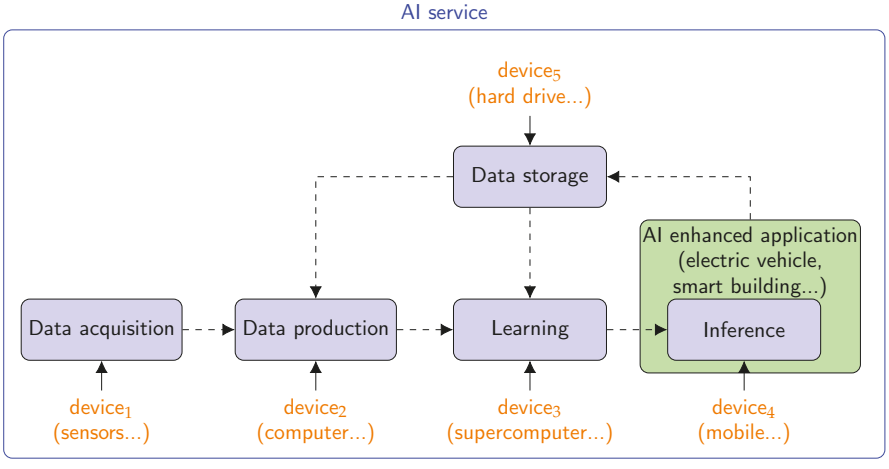
Figure 1 sums ups these two dimensions. As it has been previously stated, in the literature, only part of the global warming potential due to the use phase has generally been considered when evaluating AI, which corresponds to the shaded area in the figure.

3.2. Life Cycle Assessment Methodology for AI

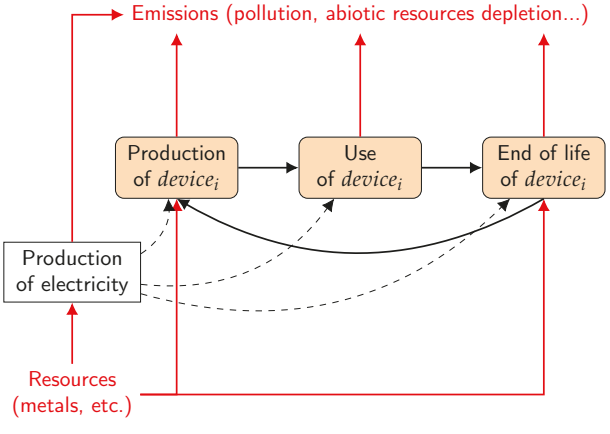
In this section, we focus on life cycle assessment of the AI solution and the associated ICT equipment. We aim at proposing a methodology for applying the general framework of LCA to AI services. For LCA of all other processes, we refer to LCA standards and [22], for example. In order to concretely apply the methodology presented for an AI service, we use the ITU recommendation [16] for environmental evaluation of ICT.

Figure 2 shows two sides of the life cycle of an AI service. The top part of this figure shows the different tasks involved in an AI service from a software point of view (data acquisition, inference, etc.). For each task, one or several devices is used. The bottom part of the figure shows the life cycle phases of each of these devices from a hardware point of view. The environmental impacts of the AI service will stem from the life cycle phases of the devices. Note that all devices involved in the AI tasks should be taken into account.

A remark on terminology: in the paper, the term "Use phase" refers to the use phase of the life cycle of equipment, corresponding to the devices provided for the AI service (box "Use of device" of the lower part in Figure 2). We call "Application phase" the inference phase of the AI service (green box of the upper part in Figure 2).



(a) Different tasks involved in an AI service



(b) Life cycle phases of each $device_i$ used by the service

Figure 2. Diagram representing the Life Cycle Inventory of an AI service. Above: an AI for Green application corresponds to the inference step that depends on other unit processes that require various devices. Below: the use of devices is located in a more global environment, including production of resources and impacts. In both schemes, colored boxes correspond to unit processes, black arrows correspond to economic flows (bold: material, dashed: energy), and red arrows to environmental flows.

Concerning the system boundaries, we refer to [26] to consider the equipment for three tiers:

- *terminals*. In the case of the smart building, this can include: user terminals used to develop, train, and use the AI service; terminals in the facility where the AI service is trained and dedicated to IT support; and smart thermostats.
- *network*. For the smart building case, this includes network equipment used for training the AI model in the facility and network equipment in the buildings where the thermostats are used.

- *data center/server*. For the smart building case, this includes servers on which the model is trained and used; training and inference can be done on the same server or not.

For each tier, all support equipment and activities may also be considered. For example, the power supply unit and HVAC of the data center should be taken into account.

The life cycle stages to consider are the ones previously mentioned: production, use, and end of life. In particular, Refs. [16,26] give classifications of unit processes according to the life cycle stages, which can be applied to AI services, as shown in Table 1.

Table 1. Application to AI services of ITU recommendation [16] regarding the evaluation of life cycle stages/unit processes.

Life Cycle Id	Life Cycle Stage and Unit Processes	Recommendation
A—Raw material acquisition		Mandatory
B—Production		
	Device production and assembly	Mandatory
	Manufacturer support activities	Recommended
	Production of support equipment	Mandatory
	ICT-specific site construction	Recommended
C—Use		
	Use of ICT equipment	Mandatory
	Use of support equipment	Mandatory
	Operator support activities	Recommended
	Service provider support activities	Recommended
D—End of life		
	Preparation of ICT goods for reuse	Mandatory
	Storage/disassembly/dismantling/crushing	Mandatory

If applied to our smart building use case, the unit processes that must be taken into account would be:

- For equipment that is dedicated to the application, such as the smart thermostats: Production, Use, and End of life.
- For the servers on which the AI service is trained and used and their environment (network devices, storage servers, backup servers, user terminal, HVAC, and other potential equipment not dedicated to the application):
  - Production and End of life with an allocation of the impacts, with respect to the execution time, for instance.
  - Part of the use phase corresponding to the dynamic energy consumption, i.e., raise of consumption due to the execution of the program.
  - Part of the use phase corresponding to the static consumption, with an allocation (for example, if n programs are run simultaneously, $1/n$ of this consumption) "since equipment is switched on in part to address the computing needs of the (Machine Learning) model" [18].

The production phase is generally important for ICT equipment in terms of global warming potential at least. Yet, when trying to assess this phase for deep learning methods, we are faced with a lack of LCAs for Graphical Processing Unit (GPUs) (or Tensor Processing Unit (TPUs) or equivalents). Ref. [27] yet showed that for a CPU-only data center in France, around 40% of the GHG emissions of the equipment were due to the production phase.

The use phase is mostly due to the energy use, so the impacts of this part are highly dependent on the server/facility efficiency and the carbon intensity of the energy sources.

The end-of-life phase is difficult to assess in ICT in general because of a lack of data concerning this phase of equipment. In particular, the end of life of many ICT equipment is poorly documented; globally, about 80% of electronic and electrical equipment is not formally collected [28].

4. Assessing the Usefulness of an AI for Green Service

Now that we have presented how the general framework of life cycle assessment can be adapted to AI solutions, we propose to use it for evaluating the complete benefits of an AI for Green service.

In this section, we will consider the following setting:

- A reference application M_1 that corresponds to the application without AI. If the application is a smart building, for example, M_1 will be the building without smart capabilities.
- An AI-enhanced application M_2 that corresponds to the application with an AI service that is supposed to have a positive impact on the environment. In the previous case, it would be the smart building.

4.1. Theoretical Aspects

When proposing an AI for Green method, one should ensure that the overall environmental impact is positive; the positive gain induced by using the AI solution should be higher than the negative impacts associated to the solution.

This requires to assess first-, second-, and third-order impacts of AI [24,25], as illustrated in Figure 3. As we detailed in the previous section, first-order impacts come from the life cycle phases of all the equipment necessary to develop and deploy the AI service.

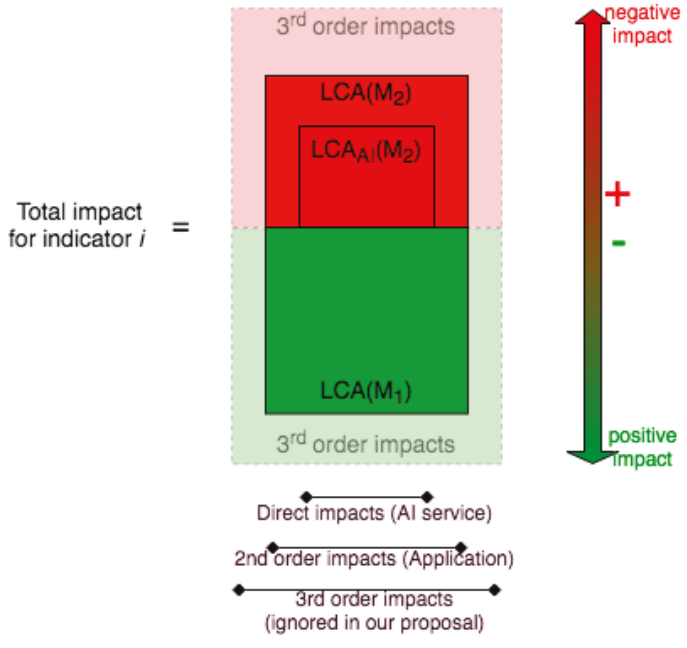


Figure 3. Overview of AI's impacts. First-order or direct impacts result from the equipment life cycle. Second-order impacts are the difference between the LCAs of the reference system and the AI-enhanced system. Third-order impacts are changes in technology or society induced by the application.

Second-order impacts correspond to the impacts due to the application of AI. AI can optimize or substitute existing systems; energy consumption in a building can be optimized using occupancy or behavior detection, energy profiling, etc.

Third-order impacts are all changes in technology or society due to the introduction of AI solutions, possibly encompassing effects of very different scales, from individual behavioral responses to systemic and societal transformations, and from short-term to long-term

effects. Rebound effects fall into this category—an increase in efficiency does not necessarily translate into a reduction of impacts of the same magnitude, and it can even lead to an increase in these impacts [29]. Rebound effects occur because potential savings (in terms of money, time, resources, etc.) are transformed into more consumption [30]. For example, due to economic savings, smart building users may decide to increase heating temperature for better comfort or to buy more flight tickets after an increase in energy efficiency.

Third-order impacts are beyond the scope of the methodology proposed here and are briefly discussed in Section 5.

According to [16], first- and second-order impacts of the AI service should be estimated based on life cycle assessment (LCA), the difference between the two being the scope—for first-order impacts, the scope is restricted to the equipment involved in the target AI service (for example, the AI involved in a smart building), while second-order impacts consider the whole solution (the smart building itself). Including second-order impacts requires extending the scope to the whole application AI is supposed to enhance. More specifically, the net environmental impacts considering both first- and second-order effects are obtained by computing:

$$\Delta(M_2|M_1) = LCA(M_2) - LCA(M_1) \in \mathbb{R}^d \quad (1)$$

with:

- M_1 the reference application without using the AI service,
- M_2 the application enhanced by AI,
- $LCA(x)$ a quantification of d types of environmental impacts (e.g., GHG emissions, water footprint, etc.). The LCA methodology is described in Section 3.2. Note that $LCA(M_2)$ includes the impacts of the AI service itself, i.e., $LCA_{AI}(M_2)$.

A previous work [21] also gave a simplified scheme for assessing the cost–benefit of deploying a foundation model, which also includes social benefits and costs but does not explicitly state the direct environmental costs of using this model. We propose to relate our methodology (Equation (1)) to their proposal. Adopting their equation but focusing on the environmental impacts only, the overall value of a model can be assessed with:

$$V(M) = S(M) - E(M) - O(M) \quad (2)$$

with:

- $V(M)$ the value of using the model, i.e., the environmental gain induced by its use in the practical application considered
- $S(M)$ the environmental benefit that can be interpreted as the difference between the initial impact of the application and its final impact (not taking into account the AI solution, i.e., the Learning and Inference task in the top part of Figure 2)
- $E(M)$ the energy cost of the model
- $O(M)$ all other impacts, including chip production, waste, risks for biodiversity, and third-order impacts (which are not discussed here).

Regarding the well-established framework of LCA, this approach suffers from several weaknesses. First, in the equation, all the values are expressed in dollars. This formally allows performing addition of several kinds of impacts but with an arbitrary consideration to the diversity of environmental issues. By definition, LCA considers multiple criteria for the impacts, previously described at the beginning of Section 3 (GHG emissions, water footprint, etc.). LCA may aggregate several impacts but with specific weights not necessarily dependent on an economic value. As noted in [22], "there is no scientific basis on which to reduce the results of an LCA to a single result or score because of the underlying ethical value-choices".

Besides, if one considers, for instance, the case of an AI service dedicated to biodiversity (see, for instance, 8.1 in [9]), one would expect to precisely quantify the positive impact of this service on biodiversity (schematically, how many species can be saved?), balanced by the negative ones (producing chips for GPUs has an impact on the biodiversity through

several sources of pollution [31]). Adopting Equation (2) will mix several impacts together and may dilute the value of interest (e.g., biodiversity) that could be burdened by negative impacts regarding energy to train the models, for instance.

Last, even if the equation is not wrong per se, the expression in terms of benefit/costs is questionable, and practical means for its computation are missing in [21].

We thus believe that Equation (1) should be used. Terms of Equation (2) can be related to the methodology proposed in our paper as follows:

$$\underbrace{V(M_2)}_{-\Delta(M_2|M_1)} \approx S(M_2|M_1)\underbrace{-E(M_2)-O(M_2)}_{LCA_{AI}(M_2)} \quad (3)$$

where $\Delta(M_2|M_1)$ and $LCA_{AI}(M_2)$ are defined in Equation (1). The negative impacts of an AI solution M_2 compared to the reference solution M_1 are not always restricted to its AI part (i.e., to $E(M_2)$ and $O(M_2)$). For example, compared to a standard vehicle, the negative impacts of an autonomous vehicle are not only due to the life cycle of (additional) ICT equipment, but also to additional aerodynamic drag due to the presence of LIDAR on the roof [32]. Hence, the nature of the impacts in $S(M_2|M_1)$ (positive or negative) cannot be stated a priori and depends on complete LCA results for both applications M_2 and M_1. It may also depend on the target environmental criteria.

4.2. Case Studies

In order to review the kind of evaluation that is usually made in the AI for Green literature, we analyzed the references for several domains of [9], which identifies potential applications of machine learning for climate change adaptation of mitigation (this review was documented in a csv file, which is given as Supplementary Material as described in the Data Availability Statement section).

We mostly chose domains that had been flagged as having a *High Leverage* and noted for each paper cited in the corresponding section the kind of environmental evaluation, with the following categories:

a. No mention of the environmental gain.
b. General mention of the environmental gain.
c. A few words about the environmental gain but no quantitative evaluation or only indirect estimation.
d. Evaluation of the energy gain without taking the AI service into account.
e. Evaluation of the energy gain taking the use phase of the AI service into account.
f. Comprehensive evaluation of the environmental gain (comparison of LCAs).

The results of the review are shown in Figure 4.

The central node is "Rolnick et al. citations". On its left are the domains of the citations. For example, the Smart building section contained 15 relevant citations.

On its right, the first flows show the partition into general machine learning applications (ML), deep learning applications (DL), and other methods (other). For example, 20 papers corresponded to deep learning applications.

The last flows on the right show the kinds of environmental evaluation. We can note that about half of the papers do not include any environmental evaluation, although the focus is on applications to tackle climate change. Many papers also give a distant proxy for evaluation, such as detailing the possible impacts without quantification or indicating the execution time of the program.

A few citations evaluate the environmental gain, mostly in terms of energy gain, but none of the papers considered took into account the AI service impacts.

It can be noted that other papers that include an evaluation of part of these impacts, can be found in the literature. Ref. [33], for example, present an intelligent control system that takes into account the expected occupancy in order to adapt the thermostat and save energy. They do not take into account learning the occupancy model, but take into account

the LCA of the smart thermostats and show that the energy needed for these devices across their whole life cycle will almost always be lower than the energy saved.

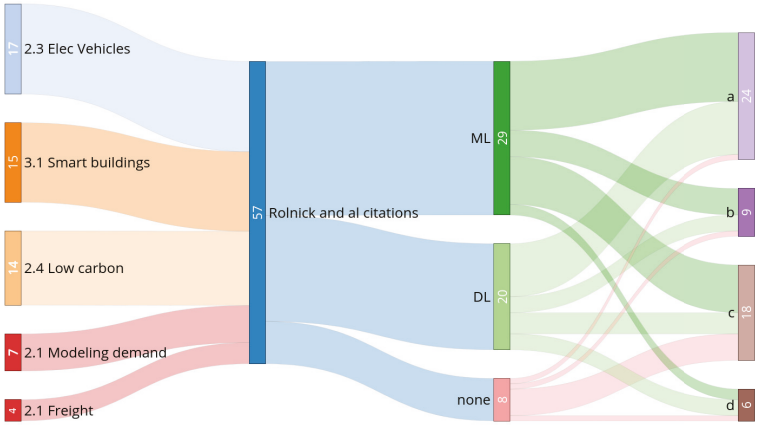


Figure 4. Sankey diagram of parts of Rolnick's paper references in terms of environmental evaluation (created with the Sankey Diagram Generator by Dénes Csala, based on the Sankey plugin for D3 by Mike Bostock; https://sankey.csaladen.es; accessed 27 February 2022).

5. Discussion

In this paper, we have analyzed the environmental impacts of AI solutions, in particular, in the case of AI for Green applications and proposed a framework to evaluate them more completely. The proposed methodology compares, through life cycle assessment, the impact of a reference solution with the AI one (1) for the appropriate types of environmental impacts. The analysis of literature on AI solutions has made the following issues/problems salient.

5.1. Current Environmental Evaluation of AI Services is Under-Estimated

We have shown that AI for Green papers only take into account a small part of the direct environmental impacts.

Several reasons can explain this under-estimation. The narratives about dematerialization that would correspond to a dramatic decrease in environmental impacts permeate AI as a part of ICT [34]. However, these narratives have proven to be false until now. Attention to AI's GHG emissions has focused on electricity consumption (energy flows). At the moment, material flows receive less attention in AI. However, it is beginning to be considered [12,17].

5.2. AI Research should Use Life Cycle Assessment to Assess the usefulness of an AI Service

Life cycle assessment is a solid methodology to evaluate not only global warming potential, but also other direct environmental impacts. LCA considers all the steps from production to use and end of life. However, it has several well-known limitations due to the complexity of processes involved in material production. Obtaining all the information to assign reliable values to each edge of the life cycle inventory also proves difficult, e.g., there is very little information on manufacturing impacts of GPU either from manufacturers or in LCA databases. To solve this problem, we could encourage the AI community to lobby companies to open a part of their data. This approach would be in the same spirit as what is happening for open science but would also require taking legal issues into account.

5.3. AI for Green Gains Are Only Potential

Even when a properly conducted LCA concludes that an AI solution is environmentally beneficial, such a result should be considered with caution. Environmental benefits computed by the LCA-based methodology proposed in this paper correspond to a technical and simplistic view of environmental problems: it assumes that AI will enhance or replace existing applications, all other things being equal. The ambition to solve societal problems using AI is praiseworthy, but it should probably be accompanied by socio-technical concerns and an evaluation of possible third-order effects. For example, autonomous vehicles are often associated with potential efficiency gains (such as helping car sharing or allowing platooning) and corresponding environmental benefits [32]. However, autonomy could also profoundly transform mobility in a non-ecological way [35].

5.4. AI Services and Large Deployment

Evaluating third-order effects is even more critical when large-scale deployment of the proposed solution(s) is projected, e.g., to maximize absolute gains. This case requires special attention even in LCA since large-scale deployment may induce societal reorganizations for producing and operating the solution(s). For example, the generalization of AI may lead to a substantial increase in demand for specific materials (such as lithium or cobalt) or energy. This increase may have non-linear environmental consequences, e.g., opening new and less performing mines, increasing the use of fossil fuel-based power plants, etc. Hence, in this case, the *attributional* LCA framework we suggest using in this paper needs to be replaced by the much more complex *consequential* one [22].

Author Contributions: Conceptualization, A.-L.L., J.L., A.B. and J.C.; methodology, A.-L.L., J.L., A.B. and J.C.; validation, A.-L.L., J.L., A.B. and J.C.; formal analysis, A.-L.L.; J.L., A.B. and J.C.; investigation, J.L. and A.-L.L.; data curation, J.L. and A.-L.L.; writing—original draft preparation, all authors; writing—review and editing, all authors; visualization, A.B. and A.-L.L.; supervision, A.-L.L.; project administration, A.-L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The file detailing the bibliographic study will be available on https://arxiv.org/abs/2110.11822 (accessed on 27 February 2022).

Acknowledgments: This work was partly supported by the CNRS EcoInfo group (https://ecoinfo.cnrs.fr/, accessed on 27 February 2022).

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DL	Deep Learning
GHG	Greenhouse Gas
GPU	Graphics Processing Unit
HVAC	Heating, Ventilation, and Air Conditioning
ICT	Information and Communications Technology
LCA	Life cycle Assessment or Analysis
LCI	Life Cycle Inventory
ML	Machine Learning
NLP	Natural Language Processing
TPU	Tensor Processing Unit

References

1. Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. *arXiv* **2019**, arXiv:1906.02243.
2. Schwartz, R.; Dodge, J.; Smith, N.A.; Etzioni, O. Green AI. *Commun. ACM* **2020**, *63*, 54–63. [CrossRef]
3. Anthony, L.F.W.; Kanding, B.; Selvan, R. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. *arXiv* **2020**, arXiv:2007.03051.
4. Henderson, P.; Hu, J.; Romoff, J.; Brunskill, E.; Jurafsky, D.; Pineau, J. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *J. Mach. Learn. Res.* **2020**, *21*, 1–43.
5. Lacoste, A.; Luccioni, A.; Schmidt, V.; Dandres, T. Quantifying the Carbon Emissions of Machine Learning. *arXiv* **2019**, arXiv:1910.09700.
6. Lannelongue, L.; Grealey, J.; Inouye, M. Green Algorithms: Quantifying the carbon emissions of computation. *arXiv* **2020**, arXiv:2007.07610.
7. Abrassart, C.; Bengio, Y.; Chicoisne, G.; De Marcellis-Warin, N.; Dilhac, M.-A.; Gambs, S.; Gautrais, V.; Gibert, M.; Langlois, L.; Laviolette, F.; et al. Montréal Declaration for a Responsible Development of Artificial Intelligence—2018 Report. Technical Report, IA Responsable. 2018. Available online: https://www.montrealdeclaration-responsibleai.com/the-declaration (accessed on 27 February 2022).
8. Walsh, T.; Evatt, A.; de Witt, C.S. *Artificial Intelligence & Climate Change: Supplementary Impact Report*; Technical Report; University of Oxford: Oxford, UK, 2020.
9. Rolnick, D.; Donti, P.L.; Kaack, L.H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A.S.; Milojevic-Dupont, N.; Jaques, N.; Waldman-Brown, A.; et al. Tackling Climate Change with Machine Learning. *arXiv* **2019**, arXiv:1906.05433.
10. Vinuesa, R.; Azizpour, H.; Leite, I.; Balaam, M.; Dignum, V.; Domisch, S.; Felländer, A.; Langhans, S.D.; Tegmark, M.; Fuso Nerini, F. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nat. Commun.* **2020**, *11*, 233. [CrossRef] [PubMed]
11. Gailhofer, P.; Herold, A.; Schemmel, J.P.; Scherf, C.U.; Köhler, A.R.; Braungardt, S. The Role of Artificial Intelligence in the European Green Deal. Technical Report, Study for the Special Committee on Artificial Intelligence in a Digital Age (AIDA), Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament. 2021. Available online: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662906/IPOL_STU(2021)662906_EN.pdf (accessed on 27 February 2022).
12. Wu, C.; Raghavendra, R.; Gupta, U.; Acun, B.; Ardalani, N.; Maeng, K.; Chang, G.; Behram, F.A.; Huang, J.; Bai, C.; et al. Sustainable AI: Environmental Implications, Challenges and Opportunities. *arXiv* **2021**, arXiv:2111.00364.
13. Cardon, D.; Cointet, J.P.; Mazières, A.; Libbrecht, E. Neurons spike back. *Reseaux* **2018**, *211*, 173–220.
14. Li, D.; Chen, X.; Becchi, M.; Zong, Z. Evaluating the Energy Efficiency of Deep Convolutional Neural Networks on CPUs and GPUs. In Proceedings of the 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom), Atlanta, GA, USA, 8–10 October 2016. [CrossRef]
15. García-Martín, E.; Rodrigues, C.F.; Riley, G.; Grahn, H. Estimation of energy consumption in machine learning. *J. Parallel Distrib. Comput.* **2019**, *134*, 75–88. [CrossRef]
16. ITU-T. Methodology for Environmental Life Cycle Assessments of Information and Communication Technology Goods, Networks and Services. Technical Report, ITU-T, 2014. Available online: https://www.itu.int/rec/T-REC-L.1410-201412-I/fr (accessed on 27 February 2022).
17. Gupta, U.; Kim, Y.G.; Lee, S.; Tse, J.; Lee, H.H.S.; Wei, G.Y.; Brooks, D.; Wu, C.J. Chasing Carbon: The Elusive Environmental Footprint of Computing. *arXiv:2011.02839* **2020**.
18. Ligozat, A.L.; Luccioni, A. A Practical Guide to Quantifying Carbon Emissions for Machine Learning Researchers and Practitioners. Technical Report, Bigscience Project, LISN and MILA. 2021. Available online: https://hal.archives-ouvertes.fr/hal-03376391/document (accessed on 27 February 2022).

19. Kaack, L.H.; Donti, P.L.; Strubell, E.; Kamiya, G.; Creutzig, F.; Rolnick, D. Aligning artificial intelligence with climate change mitigation. working paper or preprint. Available online: https://hal.archives-ouvertes.fr/hal-03368037/document (accessed on 27 February 2022).
20. Patterson, D.; Gonzalez, J.; Le, Q.; Liang, C.; Munguia, L.M.; Rothchild, D.; So, D.; Texier, M.; Dean, J. Carbon emissions and large neural network training. *arXiv* **2021**, arXiv:2104.10350.
21. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2021**, arXiv:2108.07258.
22. Hauschild, M.Z.; Rosenbaum, R.K.; Olsen, S.I. *Life Cycle Assessment*; Springer International Publishing: Cham, Switzerland, 2018; Volume 2018.
23. Heijungs, R.; Suh, S. *The Computational Structure of Life Cycle Assessment*; Springer Science & Business Media: Dordrecht, The Netherlands, 2002; Volume 11.
24. Hilty, L.M.; Hercheui, M.D. ICT and sustainable development. In *What Kind of Information Society? Governance, Virtuality, Surveillance, Sustainability, Resilience*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 227–235.
25. Horner, N.C.; Shehabi, A.; Azevedo, I.L. Known unknowns: Indirect energy effects of information and communication technology. *Environ. Res. Lett.* **2016**, *11*, 103001. [CrossRef]
26. ADEME. General Principles for the Environmental Labelling of Consumer Products, Methodological Standard for the Environmental Assessment of Digital Services. Technical Report, ADEME. 2021. Available online: http://www.base-impacts.ademe.fr/documents/Numerique.zip (accessed on 27 February 2022).
27. Berthoud, F.; Bzeznik, B.; Gibelin, N.; Laurens, M.; Bonamy, C.; Morel, M.; Schwindenhammer, X. Estimation de l'empreinte Carbone d'une Heure.coeur de Calcul. Research Report, UGA—Université Grenoble Alpes; CNRS; INP Grenoble; INRIA. 2020. Available online: https://hal.archives-ouvertes.fr/hal-02549565v4/document (accessed on 27 February 2022).
28. Baldé, C.P.; Forti, V.; Gray, V.; Kuehr, R.; Stegmann, P. *The Global e-Waste Monitor 2017: Quantities, Flows and Resources*; United Nations University (UNU), International Telecommunication Union (ITU) and International Solid Waste Association (ISWA), 2017. Available online: https://ewastemonitor.info/gem-2017/ (accessed on 27 February 2022).
29. Berkhout, P.H.; Muskens, J.C.; Velthuijsen, J.W. Defining the rebound effect. *Energy Policy* **2000**, *28*, 425–432. [CrossRef]
30. Schneider, F.; Hinterberger, F.; Mesicek, R.H.; Luks, F. ECO-INFO-SOCIETY: Strategies for an Ecological Information Society. In *Sustainability in the Information Society*; Metropolis: Marburg, Germany, 2001.
31. Villard, A.; Lelah, A.; Brissaud, D. Drawing a chip environmental profile: Environmental indicators for the semiconductor industry. *J. Clean. Prod.* **2015**, *86*, 98–109. [CrossRef]
32. Taiebat, M.; Brown, A.L.; Safford, H.R.; Qu, S.; Xu, M. A Review on Energy, Environmental, and Sustainability Implications of Connected and Automated Vehicles. *Environ. Sci. Technol.* **2018**, *52*, 11449–11465. [CrossRef] [PubMed]
33. Bracquené, E.; De Bock, Y.; Duflou, J. Sustainability impact assessment of an intelligent control system for residential heating. *Procedia Cirp Life Cycle Eng. (Lce) Conf.* **2020**, *90*, 232–237. [CrossRef]
34. Bol, D.; Pirson, T.; Dekimpe, R. Moore's Law and ICT Innovation in the Anthropocene. In Proceedings of the IEEE Design and Test in Europe Conference, Grenoble, France, 1–5 February 2021; pp. 1–5.
35. Coroamă, V.C.; Pargman, D. Skill Rebound: On an Unintended Effect of Digitalization. In Proceedings of the 7th International Conference on ICT for Sustainability, Bristol, UK, 21–26 June 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 213–219. [CrossRef]

sustainability

MDPI

Article

Sustainability Budgets: A Practical Management and Governance Method for Achieving Goal 13 of the Sustainable Development Goals for AI Development

Rebecca Raper [1,*], Jona Boeddinghaus [2], Mark Coeckelbergh [1], Wolfgang Gross [2], Paolo Campigotto [2] and Craig N. Lincoln [2]

[1] Department of Philosophy, University of Vienna, 1010 Vienna, Austria; mark.coeckelbergh@univie.ac.at
[2] Gradient Zero, 1010 Vienna, Austria; jb@gradient0.com (J.B.); wg@gradient0.com (W.G.); pc@gradient0.com (P.C.); cl@gradient0.com (C.N.L.)
* Correspondence: rebecca.raper@univie.ac.at

Abstract: Climate change is a global priority. In 2015, the United Nations (UN) outlined its Sustainable Development Goals (SDGs), which stated that taking urgent action to tackle climate change and its impacts was a key priority. The 2021 World Climate Summit finished with calls for governments to take tougher measures towards reducing their carbon footprints. However, it is not obvious how governments can make practical implementations to achieve this goal. One challenge towards achieving a reduced carbon footprint is gaining awareness of how energy exhaustive a system or mechanism is. Artificial Intelligence (AI) is increasingly being used to solve global problems, and its use could potentially solve challenges relating to climate change, but the creation of AI systems often requires vast amounts of, up front, computing power, and, thereby, it can be a significant contributor to greenhouse gas emissions. If governments are to take the SDGs and calls to reduce carbon footprints seriously, they need to find a management and governance mechanism to (i) audit how much their AI system 'costs' in terms of energy consumption and (ii) incentivise individuals to act based upon the auditing outcomes, in order to avoid or justify politically controversial restrictions that may be seen as bypassing the creativity of developers. The idea is thus to find a practical solution that can be implemented in software design that incentivises and rewards and that respects the autonomy of developers and designers to come up with smart solutions. This paper proposes such a sustainability management mechanism by introducing the notion of 'Sustainability Budgets'—akin to Privacy Budgets used in Differential Privacy—and by using these to introduce a 'Game' where participants are rewarded for designing systems that are 'energy efficient'. Participants in this game are, among others, the Machine Learning developers themselves, which is a new focus for this problem that this text introduces. The paper later expands this notion to sustainability management in general and outlines how it might fit into a wider governance framework.

Keywords: AI; artificial intelligence; sustainability; AI governance; ethics; ethical AI; differential privacy

Citation: Raper, R.; Boeddinghaus, J.; Coeckelbergh, M.; Gross, W.; Campigotto, P.; Lincoln, C.N. Sustainability Budgets: A Practical Management and Governance Method for Achieving Goal 13 of the Sustainable Development Goals for AI Development. *Sustainability* **2022**, *14*, 4019. https://doi.org/10.3390/su14074019

Academic Editors: Aimee van Wynsberghe, Larissa Bolte, Jamila Nachid and Tijs Vandemeulebroucke

Received: 24 February 2022
Accepted: 22 March 2022
Published: 29 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction: Sustainability and AI

In 2020, carbon dioxide levels in the atmosphere were measured to be 149 percent above the pre-industrial level, Nitrogen levels at 262 percent and Nitrous Oxide at 123 percent [1]. By the end of the century, it is expected that these increased levels will contribute to a global warming of 1.5–2 degrees Celsius above the pre-industrial levels. This will not only lead to a change in climate conditions, but also contribute to melting ice caps, rising sea levels and ultimately pose risks to humans due to flooding, dangerous weather and disrupted crop growth [2]. In 2021, it might be claimed that we are already beginning to see some of the impacts of global warming, with significant heatwaves, forest fires and flooding causing global disruption in 2021 [3–5].

In 2015, the United Nations (UN) set their 'Sustainable Development Goals' (SDGs) as key global priorities for a better world by 2030 [6]. One of the key goals (Goal 13) is 'to take urgent action to combat climate change and its impacts'. More recently, at the 2021 World Climate Summit (a global conference aimed at addressing climate change), a worldwide pledge was made for countries to do more to reduce their carbon footprints.

To address these issues, governments need to take action to address the reliance on infrastructure that is fuelled by carbon. As Stein [7] outlines, Artificial Intelligence has the power to deal with some of these challenges through improved efficiency, accuracy and efficacy of systems; however, it can also be a significant contributor to some of the same issues it aims to resolve, due to the large amount of computing power that is often required to train and test modern (Machine Learning-based) AI systems. Strubell et al. [8] found that training large natural language machine learning models can consume huge amounts of energy, with the average AI development projects, training many models to find the best solution, emitting as much as 78,000 pounds of Carbon Dioxide—more than five cars in their lifetime. Moreover, Strubell et al. [8] showed this to be the case especially during the last optimization phases of AI development, i.e., increasing the performance of an already well-performing AI model a few more points, consumes disproportionally high amounts of energy for the utility of the project. Therefore, focusing on the AI development perspective is as important as considering broader management and governance.

With the 'Fourth Industrial Revolution' [9] accelerating the rate at which AI is adopted in society, and in the light of the possibility that AI is going to be used as a tool to resolve issues relating to climate change, there are serious issues regarding sustainability that need addressing, namely, the amount of energy that is consumed.

One strategy for dealing with the amount of energy expended by AI systems is to exhaustively document the amount of energy that the development and operation of an AI system uses, from idea generation, through to training, re-tuning, implementation and governance [10], and to actively work to minimise these levels. However, from a governance perspective it is not easy to see how this could be managed. In the first instance, there are minimal procedures or methods to document how much energy the production of an AI system expends (or might expend), and, even if this were known, it is not known what should be done to minimise this level. From a product manager's perspective, there is no access to the development process to understand how changes could be made to reduce the amount of energy consumption, and, from a developer perspective, they do not have the broader understanding of the system to see where allowances could be made and when energy should be reduced.

Furthermore, following the documentation, usually the idea is then to achieve minimization of energy levels by constraining developers and their companies via regulation. However, this may be politically problematic and controversial, since it raises the issue of freedom [11] and does not respect the autonomy and creativity of the developers and other stakeholders such as management and end users. The focus is also on the organisation; developers are often not directly addressed. What is required, then, is a methodology to combine and integrate these insights to allow for appropriate oversight, management, development and governance in a way that respects the autonomy and creativity of stakeholders, in particular the AI developers.

The remainder of this paper traces a method towards achieving this aim: utilising 'Sustainability Budgets', in analogy with Privacy Budgets in Differential Privacy, it develops a procedure that empowers developers, allows management to have sufficient oversight and provides a governance framework towards achieving Goal 13 of the SDGs. Section 2 introduces Differential Privacy, Privacy Budgets and then 'Sustainability Budgets'; Section 3 outlines how this can fit into a wider organisational management framework, including using Gamification to incentivise more sustainable development. In Section 4, Sustainability Budgets are outlined in relation to sustainability in society and governance towards achieving the SDGs. Finally, there is reflection on the method, along with an outline of the limitations in Section 5, before a conclusion.

2. Our Proposed Method: 'Sustainability Budgets' in Analogy with Differential Privacy

2.1. Differential Privacy and Budgets

Managing the privacy of individual data is a priority in data science. Often, within the development phase, vast quantities of personal data are used to train the AI system. Owing to regulatory frameworks, such as GDPR [12], organisations and developers have a duty to protect the personal data of individuals and to ensure it is not lost or given to the unapproved individuals. Privacy becomes particularly important in settings involving sensitive data (for example, health care settings). In the wrong hands, such data could be used to manipulate (i.e., for insurance purposes) for financial gain. Therefore, in these settings, it is even more important to ensure that data privacy is maintained. Some personal data is so sensitive that it requires specialist approval or clearance before it can be seen. It can be very time consuming and difficult from a management perspective to ensure that every developer and project team member working on development of an AI system has the right degree of clearance. One approach to resolve this is to prevent the project team having access to such sensitive data at all. However, for AI development, access to the data is required to train the models in the first place.

A resolution to this problem is to keep the data on a separate system so that training occurs 'remotely', and the data scientist does not have access to the information in the data that is being used to train but can still see the analysis outcomes. However, even if information is not immediately accessible to a data scientist, through hacking techniques (i.e., querying the information in certain ways) personal information can still be retrieved. In 2006, Cynthia Dwork et al. introduced a mathematical technique which prevents this problem, known as Differential Privacy (DP) [13]. In DP, developers can interrogate (or analyse) individual data securely, while preserving its anonymity. A differentially private algorithm guarantees that its output does not change significantly (as quantified by the DP parameter ε) if one single record in the dataset is removed or modified or if a new record is inserted in the dataset. Differential Privacy protects individuals in "that the analyst knows no more about any individual in the data set after the analysis is completed than she knew before the analysis was begun." [14].

However, if not managed effectively, the more a data set is queried, the greater the chance is of retrieving potentially private information from that data. Within DP, this is what is referred to as 'Privacy Leakages', whereby unwanted revelations about individual data points are gained by repeatedly performing queries on seemingly harmless statistics. For example, I might be able to identify who an individual person is in a medical data set, by performing repeated queries on known medical diagnoses. Within DP, this is often managed using a notion known as 'Privacy Budgets'. The Privacy Budget is a direct consequence of the privacy parameter ε as introduced by Dwork [13]. As the ε parameter in this so-called epsilon-DP defines the degree of the allowed influence of individual data points on analysis results, the privacy budget is the upper limit of accumulated epsilon values for a given dataset, effectively guaranteeing the privacy of individuals. In simple terms, the Privacy Budget defines an upper limit for information that can be disclosed for a given dataset. The important point being that every time a data analytics task is performed—be it a SQL query with aggregate statistics or a Machine Learning training job—some information about the data is revealed. The 'purported ideal' would be to have a completely anonymized data set that at the same time can be used for sensible data analysis; however, as this is not achievable, the amount of 'privacy leakage' must be balanced against the utility of the outputs. Differential Privacy, with its mathematical definition of privacy, makes this fact clear and transparent.

DP is a robust concept that tries to focus on how much information is retrieved from a data query. A particularly interesting feature of this approach is that privacy issues need to be dealt with at the time of initial development, which puts responsibility at the start of development procedure.

One aspect of sustainability and, more specifically, energy efficiency in Machine Learning development, is that it is all too often considered as an organisational issue,

where an organisation that is aligned with sustainability goals, strives to achieve a lower carbon footprint by reducing the energy consumed within its data centres [15]. While this is definitely an important endeavour, the AI development perspective is often overlooked but can contribute significantly to achieving sustainability goals.

As with 'Privacy', from a legal and governance perspective, we have a duty to manage the amount of energy that is expended when AI systems are developed. Akin to 'Privacy Budgets' within DP, we introduce the concept of a "Sustainability Budget" or "Sustainability Score" to allow sufficient oversight of this.

In DP analysis, the amount of information can—and must—be specified before each data query computation. That means, in practice, before an AI developer starts to train a Machine Learning model on a dataset, they need to detail the specified amount of 'information leakage'. The high the budget that is requested, the more accurate the model will be, but the less remaining room there will be for further computations on the dataset, and vice versa. The privacy problem is put in the hands of the developer and therefore becomes a central element of the analysis itself which leads to nothing else other than a strong requirement for data protection awareness at the development level. The developer must carefully think about how they would like to use the data; once the analysis is started (or more specifically for DP: once the analysis result is published), the decision has been made. In a similar regard, a 'Sustainability Budget' can motivate developers to think about the effects of their to-be-started analyses before they hit the run button.

Let us define the Sustainability Budget (SB) as a virtual upper limit of available compute resources for Machine Learning jobs. Likewise, a 'Sustainability Score' can be defined as the inverse: the to-be-consumed compute resource deducted from a virtual upper limit.

Such a score would not constrain developers and their organisations but instead raise awareness on the side of the developer and incentivize solutions that are less resource-intensive and therefore more sustainable.

To calculate and use this new Sustainability Budget, we need a formula for the carbon footprint of a software system and metrics that can be used to calculate it. A good template comes from the Greensoftware Foundation and their "Software Carbon Intensity (SCI) Specification". An alpha version can be found in [16].

The formula for the SCI rate is defined as:

$$\text{SCI} = ((\text{E} \times \text{I}) + \text{M}) \text{ per R}, \quad (1)$$

where E is the energy consumed by a software system, I is the location-based marginal carbon emissions, M is the embodied emissions of a software system, and R is a functional unit.

We can set R to be one ML (training) job and thereby calculate the SCI per the ML process. SCI indicates how much carbon equivalent emissions are produced per kilowatt-hour (energy, E). This must be identified with the local energy provider and data centre administration team. Similarly, the embodied emissions are calculated beforehand based on the employed hardware systems. To quantify the carbon equivalent emissions per unit, one approach is to measure the actual consumption by implementing a respective monitoring at the supplier level. However, there are other 'prediction' approaches, potentially conceptual ML techniques, that could also be conceptual.

Once the initial calculations and measurements are in place, the SCI can be used to compare different ML jobs (units R) and to publish the rates in the team's sustainability reports. On a basic level, with everything else fixed, the SCI comes down to measuring the energy consumption of a ML job. With a balanced set of compute resources this can further be refined to measure the time one ML job is running.

Having such a compute-time or SCI rate measurement in place, AI development teams can now define a Sustainability Budget (or goals for sustainability scores) for different projects. Each developer will be aware of this budget. They will plan for an efficient model training process using as little energy as possible for a model that is as high performant level (accurate, precise, private, etc.) as possible under these conditions. Since model

optimization usually follows a logarithmic curve (model performance vs. iterations), a good approach would be to estimate the expected performance gain with each step, i.e., experiment, that the machine learning developer performs while finding an optimal model. This makes the sustainability score part of the development process itself.

2.2. A Practical Case Study: Using Bayesian Optimization for Experiment Selection

If we look at the technical development process at an individual level, ML development can be described by a series of smaller experiments that are executed to develop one Machine Learning model. In this sense, each 'experiment' uses computational resources that can be described with the SCI rate introduced above. All the experiments (units R) need to fit into the given Sustainability Budget for the entire development lifecycle [10]. So, the question becomes how to structure work as an ML engineer, and what the most efficient workflow is to obtain the best model, within a given budget.

The quantity of data and size of the model are the biggest drivers for the computational cost of each experiment [17]. There is work that tries to mitigate this by working with efficient networks [18,19], compression networks [20–22] or a minimum amount of data [23,24]. Not only can this reduce computational cost during development, but it can also solve other limitations, such as computational capacity on edge devices [18,19] or the availability of data. Moreover, ideas such as Transfer Learning, etc. [25] can be used to develop models more efficiently.

These are interesting approaches, but to highlight how a sole developer might drive sustainability in their individual development, we implement Sustainability Budget approach, this section focuses on Bayesian Optimization as a case study, to guide the selection of experiments and reduce the number of experiments needed to reach the desired result.

To understand how this can be done, it is important to realise that each experiment is carried out to gain information about the model's performance for a specific set of parameters. For the ML engineer, this is the main part of their work, where they want to understand, with high certainty, how high the model's error is with different settings, in order to determine the best set of settings or to conclude how the different settings influence the results for comparison. The model's error can be described by a function, with input space over the parameters and the output space over the error of the model, where low errors represent good models. This error function is unknown and cannot be optimized directly, so optimization methods such as Gradient Descent [26] are not applicable and only black box optimization methods [27] or brute force methods (systematically calculating all possible options) can be used.

Naturally, systematic testing of different input sets could be carried out to estimate the error function for different parameters. However, this would not be favourable, because it is not very efficient and grows exponentially with the number of parameters, because each parameter is tested individually. It is very likely that some parameters have a bigger impact on the model's performance than others, but we do not know which ones beforehand. If known, the most important should be tested first. However, this becomes more complicated when we consider that the parameters are not independent at all, and that randomly selecting parameters is a better option, because it is more efficient in testing more unique combinations, which has also been proven experimentally [28]. To achieve even better efficiency, the information from previous experiments should be used to determine which parameter set it evaluates next. This can be conceptualized with Bayes Formula:

$$P(H \mid E) = P(E \mid H)\, P(H)/P(E). \quad (2)$$

Bayes Formula of conditional probabilities puts the prior probability P(H) of the hypothesis H in relation with the likelihood P(E | H) of the evidence E for a given H. With the prior and the likelihood, we can calculate the posterior probability P(H | E), which is what is most interesting. The term P(E) is the marginal likelihood of all hypotheses H and can be treated as a normalisation constant for these purposes.

Applying this to the experiment selection problem of the ML engineer, this method can be used to find the optimal set of parameters for the error function of the model. Due to the fact that the function cannot be accessed directly, it is treated as a probability function. The prior P(H), expressed as a probability function over the parameter space, gives the prior belief of the form of this error function. Prior to any experiments, it can be assumed that a uniform probability density function (pdf) would represent the belief that no set of parameters is advantageous to any other set, but the ML engineer might just as well implement any other prior beliefs that may be held over the model's error function into the equation. The evidence represents the experiments performed and the likelihood of this evidence, which indicates the probability that the evidence fits together with the hypothesis. The posterior can be calculated by combining the likelihood and the prior which can be understood as an update of the prior once new evidence (i.e., information) is present. This process can be repeated when new evidence is present. The prior of the next interaction is set equal to the posterior of the latter interaction, and the new posterior is calculated, i.e., the belief is updated. This represents an update of the belief on the ML engineer (or the system) given new evidence.

To select the best experiment to perform next, the described posterior function (which is often modelled as a gaussian process) can now be used, and the next point to evaluate can be derived (see Figure 1). During the gaussian process, a density estimation is gained for each point on the error function rather than a simple point estimate. In the plot below, the line in the middle is the mean of the estimate and the area around represents the uncertainty, i.e., the density of the estimate. The red dots are points where the function has been evaluated in some experiments, and the uncertainty is therefore zero at these locations.

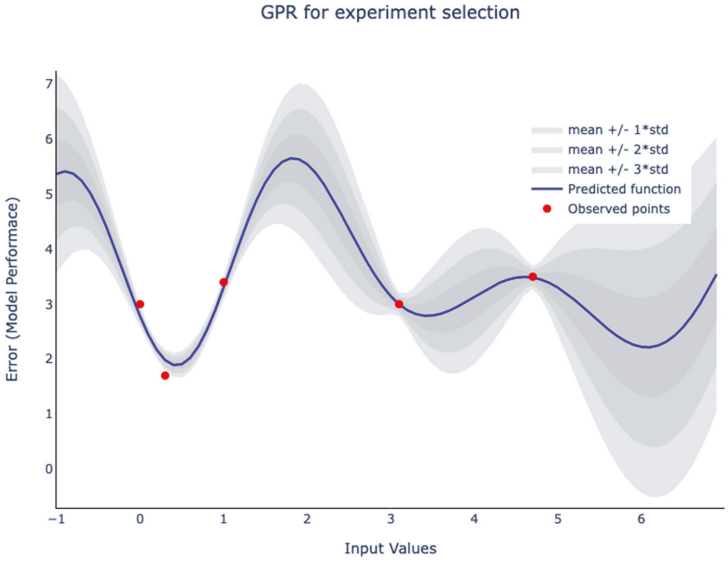


Figure 1. A Gaussian Process Regression (GPR) model that has been fitted on observed data (red dots). In the context of the case study, these observations would be the results of previous experiments. The predicted function (blue) is the result of the GPR and is simply the mean of the GPR posterior estimation. The grey area visualises the uncertainty of this function. If the parameters are desired (note that really the number of parameters is much greater than 1), in the next experiment the plot can be used to find a good trade-off between exploration and exploitation. What is seen is that the dip around 0.5 is certain, but further development would likely result in limited performance gains. Alternatively, the area around 4 or 6 could be explored, with high uncertainty, to improve the current estimation and avoid missing a better local optimum for the model performance.

Now, this provides the information to choose which point of the error function to evaluate next. Depending on the task, the current optimal point could be evaluated or a point with large uncertainty to improve the estimate there.

This strategy is an example of how to make informed decisions while optimising a machine learning algorithm. With every step, the next best experiment is decided. Therefore, the cost per unit as described above is clearly present in the form of the number of experiments to be performed. It is a good idea to keep this process a manual one (instead of trying to automate away the exploration) as it brings the desired care into the process. The ML developer is aware of the sustainability costs associated with each experiment and therefore uses a carefully chosen method to optimise the model, integrating and updating beliefs with each iteration.

3. 'Gamify!': Gamification Techniques to Manage Sustainability

We introduced Sustainability Budgets and have offered a practical way developers might use up-to-date ML techniques to achieve optimum sustainability. With these in mind, let us ask how developers might be incentivised to achieve more sustainable development. For example, while a developer might have awareness of how much energy the particular development of their system expends and the capability to minimise it, without some form of incentivisation, the whole procedure may seem pointless. Why invest time and energy in making a system more efficient if there are no benefits for doing so?

While we believe this to ultimately be a management issue, where managers decide how to maintain sustainability, Sustainability Budgets open the possibility to use 'Gamification', which has been shown to be successful in incentivising participation.

Gamification—a notion taken from the gaming industry—is the enrichment of products, services and information systems with game style elements in order to positively influence the productivity, motivation and behaviour of users [29]. An early example was the Nike+ running application [30], which encouraged individuals to engage in physical activity by making a 'game' out of running. Users were encouraged to compete against other runners for positions on leader boards, 'fastest course completions' or simply 'most runs that week'. More recently, other exercise applications such as Strava and Peloton make use of the gamification technique, and this can be extended to other areas.

Sustainability Budgets lend themselves very well to gamification techniques, with 'Sustainability Scores' or Limits usable to drive competition between developers. Therefore, a developer might be incentivised to develop their system in the most efficient way, because they are part of a 'game' with other developers, where there is a benefit for 'winning' (whether this be financial or otherwise).

As well as lending itself to competition between developers, moving one abstraction level away, we can begin to look at how this might also translate to operating at an organisational level.

As mentioned in Section 1, many organisations are motivated to reduce their carbon footprints, but this often seems to weigh against certain costs, meaning that there often has to be a compromise on sustainability in order to drive down the costs of development. However, if gamification were used as a technique to not only incentivise developers, but also to audit how systems were developed and their efficiency (through a Gamification management system), it would be easier for organisations to understand and explain the added time and costs for the delivery of projects. Ultimately, gamification systems would not only be instrumental in incentivising developers, but they could also be used by organisations to show why they are particularly sustainable

It could also be applied from an executive management perspective, and it could even be used at a national and global governance level. There could be competitions between teams of an organisation or even across different organisations or countries.

Utilising the 'Sustainability Budgets' methodology, there will be a mechanism to plan for and log the amount of energy expended per development task and also information for developers to provide explanations at a management level for different components of

the development process. For example, a developer would be able to articulate why they chose a particular training step or method over another, in terms of the balance between sustainability and effectiveness, and there would be oversight as to how much energy was expended per process. The developer would be incentivised to be more energy efficient in their development approach, and there could be incentivised competitions between different developers to find the most 'energy efficient', effective approach.

The incentivisation could be extended to include 'games' between different developers. From a management level, this would give the chance to reward more sustainable development through a points/score-based system. There would need to be a balance between sustainability and effectiveness of the algorithms, but a score-based system would allow visibility of the compromises and management accordingly. For example, as a manager, I might decide to increase my sustainability budget (i.e., the amount of energy a development project expends), because the project is particularly pertinent (from a business perspective) or because having a lower budget would make the system unsafe (i.e., in a medical context). Equally, I might decide to lower my budget because, from a resource perspective, the project allows for this. Gamification would encourage developers to create their systems in the most energy efficient ways, being rewarded for creative technical solutions.

4. Using Sustainability Budgets to Achieve the SDGs

Eventually, the proposed method might also contribute to wider political and societal goals. At the beginning of this paper, the Sustainable Development Goals (SDGs) were discussed, in particular, Goal 13 and the drive for countries to achieve carbon neutrality, or at least reduce their carbon footprints. As well as offering a mechanism for incentivisation through Gamification (or similar) approaches, at either a developer, management or governance level, what Sustainability Budgets ultimately provide is a mechanism for 'Energy Consciousness' insofar as not only do developers become more aware of the amount of energy the development of their system expends, but managers can translate this into development action, and governments can reward/recognise organisations who offer more sustainable solutions or govern organisations to be more sustainable.

Strategically, those systems that offer the greatest benefit/cost ratio would be the systems that are invested in the most. Where previously it was difficult to determine this ratio in the early phases of a project because there was no mechanism to quantify it beforehand, there can be effective management of energy expenditure at all levels of an organisation: a company, for example, but also a sector and even a nation.

For example, energy budgets could be set on an industry basis, based upon the needs of the industry. There could be regulation to ensure that industries did not go beyond their allowance, and the different levels could be managed at a political level. Organisations could be incentivised to be sustainable through energy ratings that would ensure that standards were being maintained. An organisation might receive an 'A' energy rating for particularly energy efficient processes and a lower rating ('B') for a lower energy efficiency. As with energy efficiency levels used to rate buildings and appliances (ref), the energy expended for each project would need to be measured against the size and utility of the project.

A strong contribution of the Sustainability Budget approach is to provide a way to embed the topic of environmental sustainability into ML practices and organisations from the beginning. The urgent need to address climate change is injected directly in the project development phase of machine learning and AI applications. Whereas most of the time such topics are managed purely in an abstract way with review processes that are being applied after the fact and that are followed or not, sustainability budgets put this matter at the heart of the development process itself. With the aforementioned gamification implementation, there is a direct and accessible way to address these problems.

And most of all: sustainability budgets raise awareness about this topic. Next to incentivisation, sustainability budgets and the related gamification strategies are likely to

raise awareness of energy consumption and sustainability issues, which may influence the behaviour of individuals and the organisation. Just like privacy budgets in differential privacy, sustainability budgets need to be handled before starting a training run of a machine learning task and, because of that, they force developers, project managers and policy makers to think about the impact of the project early on and continuously. It does matter if a machine learning task is solved by using the next best large network trained on all available next best data or rather designed and developed with energy consumption and sustainability in mind, carefully selecting the optimal network architecture, data subset and iterative development approach. Being aware, after all, is the first and most important step when trying to change and influence things. Sustainability budgets help raise awareness of the climate change impacts Machine Learning development can have at and before the development process itself.

5. Limits of the Methodology

It is important to note that the notion of 'budgeting' energy consumption has its limitations. As with the concept of budgeting privacy, careful calculations need to be carried out to weigh up the balance between the amount of energy that should be expended and how useful it would be to expend that energy. If the correct ratio is not established, then it becomes useless to set sustainability limits in the first place. If energy limitations are hindering innovation and project development, then the budget needs to be reassessed and project priorities need to be reassigned. This, however, is something that would need considering at the management level.

Another limitation of the presented approach is the extent to which the privacy budget metaphor extends to the case of energy budgets. It is noted that there are some dissimilarities between the two notions, namely 'privacy' as a unit of measurement is an abstract concept whereas 'energy' has real tangible values (i.e., the amount of energy that is consumed). Differential Privacy is a mathematical concept, whereas sustainability budgets stem from an organisational or management background. However, focusing on such dissimilarities distracts from the focus of the approach that has been detailed, which is an emphasis on creating an awareness of how sustainable a project might be and putting in a framework to mitigate or manage this. In terms of the methodological approach and its consequences, namely the required attention to the data protection and sustainability factors, respectively, in the development process itself, there are useful similarities that can be taken from the management of differential privacy to encouraging sustainability.

Finally, one key limitation of the approach detailed in this paper is the extent to which this is an implementable process. What has been given is a conceptual approach and a sample case study, but by no means is it ready to be placed into the software development process. Ultimately, we see this as an area for future development, and if Sustainability Budgets would be pursued, an area that would require further research. The following areas need developing further:

Calculation of the appropriate ratio for the performance vs. cost.

Investigation for how organisations ought to be governed given energy efficiency levels.

Development of a suitable gamification platform on which to record (and provide a game for) development energy consumption.

6. Conclusions

In this paper we offered a technique for managing sustainability in the development process of AI system creation, which utilises a notion we term 'Sustainability Budgets'. This both empowers the developer creating the AI system and allows the management and governments to have appropriate oversight. The notion of 'games', whereby developers can engage in competitions to achieve the most sustainable product vs. its effectiveness, is introduced as a possible incentive to encourage developers to achieve their organisational and a nation's sustainability goals. The conversation in this paper not only offers a new methodology for individuals to move closer towards achieving the SDG's, it also inspires

debate in this area, and this may lead to even more practical ways to ensure that Artificial Intelligence aids (rather than hinders) a sustainable future.

Author Contributions: Conceptualization, J.B., R.R. and M.C.; methodology, J.B., R.R., M.C. and W.G.; validation, W.G., C.N.L. and P.C.; formal analysis, R.R.; investigation, R.R. and W.G.; resources, W.G. and J.B.; data curation, W.G.; writing—original draft preparation, R.R., M.C., J.B., W.G., P.C. and C.N.L.; writing—review and editing, R.R., M.C., C.N.L., P.C. and J.B.; visualization, W.G.; supervision, R.R. and M.C.; project administration, R.R.; funding acquisition, J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partly funded by the FFG, "Österreichische Forschungsförderungsgesellschaft" (Austrian Research Funding Institution) as part of the research project "Ethische KI", funded as an FFG "Basisprogramm" and by Gradient Zero. Open Access Funding by the University of Vienna.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Open Access Funding by the University of Vienna.

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

References

1. UN News. 'No Time to Lose' Curbing Greenhouse Gases: WMO. 2021. Available online: https://news.un.org/en/story/2021/10/1103892 (accessed on 13 December 2021).
2. UK Met Office Online. Effects of Climate Change—Met Office. 2021. Available online: https://www.metoffice.gov.uk/weather/climate-change/effects-of-climate-change (accessed on 13 December 2021).
3. NBC News. Heat Wave 2021: Climate Scientists Warn about a New Normal. 2021. Available online: https://www.nbcnews.com/science/environment/heat-wave-2021-climate-scientists-warn-new-normal-rcna1664 (accessed on 21 March 2022).
4. The Guardian. Fires Rage around the World: Where Are the Worst Blazes? 2021. Available online: https://www.theguardian.com/world/2021/aug/09/fires-rage-around-the-world-where-are-the-worst-blazes (accessed on 25 March 2021).
5. BBC News. Germany Floods: Dozens Killed after Record Rain in Germany and Belgium. 2021. Available online: https://www.bbc.co.uk/news/world-europe-57846200 (accessed on 25 March 2021).
6. Globalgoals.org. The Global Goals. 2021. Available online: https://www.globalgoals.org/ (accessed on 13 December 2021).
7. Stein, A.L. Artificial Intelligence and Climate Change. *Yale J. Reg.* **2020**, *37*, 890.
8. Strubell, E.; Ganesh, A.; McCallum, A. Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 28 July–2 August 2019.
9. Schwab, K. The Fourth Industrial Revolution. 2017. Available online: https://www.weforum.org/about/the-fourth-industrial-revolution-by-klaus-schwab (accessed on 22 March 2022).
10. Wynsberghe, A.V. Sustainable AI: AI for sustainability and the sustainability of AI. *AI Ethics* **2021**, *1*, 1–6. [CrossRef]
11. Coeckelbergh, M. AI for Climate: Freedom, Justice, and other Ethical and Political Challenges. *AI Ethics* **2021**, *1*, 61–72. [CrossRef]
12. Regulation (EU) 2016/679 of the European Parliament and of the Council, General Data Protection Regulation. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02016R0679-20160504&from=EN (accessed on 23 February 2022).
13. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating noise to sensitivity in private data analysis. In Proceedings of the TCC 2006: Theory of Cryptography Conference, New York, NY, USA, 4–7 March 2006; pp. 265–284.
14. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. Found. *Trends Theor. Comput. Sci.* **2014**, *9*, 211–407. [CrossRef]
15. Gao, J. Machine Learning Applications for Data Center Optimization. 2014. Available online: https://www.google.com/url?q=https://research.google/pubs/pub42542/&sa=D&source=docs&ust=1645625964864718&usg=AOvVaw2sF6awAp8KZDficvpwLU5i (accessed on 23 February 2022).
16. Hussain, A.; Gupta, A.; Time, H.-W.; Buchanan, W.; Bergman, S.; Knight, V.; Lloyd-Jones, C.; Srinivasan; Kariya, M.; Lewis-Toakley, D. Software Carbon Intensity (SCI) Specification (v.Alpha). 2021. Available online: https://github.com/Green-Software-Foundation/software_carbon_intensity/blob/main/Software_Carbon_Intensity/Software_Carbon_Intensity_Specification.md (accessed on 2 February 2022).
17. Justus, D.; Brennan, J.; Bonner, S.; McGough, A.S. Predicting the computational cost of deep learning models. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 3873–3882.
18. Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Sandler, M.; Howard, A.; Le, Q. MnasNet: Platform-Aware Neural Architecture Search for Mobile. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1–9.

19. Howard, A.; Menglong, Z.; Chen, B.; Kalenichenko, D.; Wang, W.; Wayand, T.; Andreeto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
20. Han, S.; Mao, H.; Dally, W. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. *arXiv* **2015**, arXiv:1510.00149.
21. LeCun, Y.; Denker, J.; Solla, S. *Advances in Neural Information Processing Systems*; Massachusetts Institute of Technology Press: Cambridge, MA, USA, 1989; pp. 598–605.
22. Tanaka, H.; Kunin, D.; Yamins, D.; Ganguli, S. Pruning neural networks without any data by iteratively conserving synaptic flow. *arXiv* **2020**, arXiv:2006.05467.
23. Arora, S.; Du, S.; Li, Z.; Salakhutdinov, R.; Wang., R.; Yu, D. Harnessing the Power of Infinitely Wide Deep Nets on Small-data Tasks. *arXiv* **2019**, arXiv:1910.01663.
24. Tartaglione, E.; Barbano, C.A.; Berzovini, C.; Calandri, M.; Grangetto, M. Unveiling COVID-19 from chest X-ray with deep learning: A hurdles race with small data. *Int. J. Environ. Res. Public Health* **2020**, *17*, 6933. [CrossRef] [PubMed]
25. Donahue, J.; Yangqing, J.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; Darrell, T. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In Proceedings of the 31st International Conference on International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 647–655.
26. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
27. Audet, C.; Kokkolaras, M. Blackbox and derivative-free optimization: Theory, algorithms and applications. *Optim. Eng.* **2016**, *9*, 100011. [CrossRef]
28. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
29. Deterding, S.; Dixon, D.; Khaled, R.; Nacke, L. From game design elements to gamefulness: Defining "gamification". In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, New York, NY, USA, 28–30 September 2011; pp. 9–15.
30. Blohm, I.; Leimeister, J.M. Gamification. *Bus. Inf. Syst. Eng.* **2013**, *5*, 275–278. [CrossRef]

sustainability

MDPI

Article

The Environmental Sustainability of Digital Technologies: Stakeholder Practices and Perspectives

Gabrielle Samuel [1,*], Federica Lucivero [2] and Lucas Somavilla [3]

[1] Department of Global Health and Social Medicine, King's College London, Bush House, Strand, London WC2B 4BG, UK
[2] Ethox Centre and Wellcome Centre for Ethics and Humanities, University of Oxford, Oxford OX3 7LF, UK; federica.lucivero@ethox.ox.ac.uk
[3] Responsible Technology Institute, Department of Computer Science, University of Oxford, Oxford OX1 3LN, UK; lucas.somavilla@cs.ox.ac.uk
* Correspondence: gabbysamuel@gmail.com

Abstract: Artificial Intelligence and associated digital technologies (DTs) have environmental impacts. These include heavy carbon dioxide emissions linked to the energy consumption required to generate and process large amounts of data; extracting minerals for, and manufacturing of, technological components; and e-waste. These environmental impacts are receiving increasing policy and media attention through discourses of environmental sustainability. At the same time, 'sustainability' is a complex and nebulous term with a multiplicity of meanings and practices. This paper explores how experts working with DTs understand and utilise the concept of environmental sustainability in their practices. Our research question was how do stakeholders researching, governing or working on the environmental impacts of DTs, utilise environmental sustainability concepts? We applied a combination of bibliometric analysis and 24 interviews with key stakeholders from the digital technology sector. Findings show that, although stakeholders have broad conceptual understandings of the term sustainability and its relation to the environmental impacts of DTs, in practice, environmental sustainability tends to be associated with technology based and carboncentric approaches. While narrowing conceptual understandings of environmental sustainability was viewed to have a practical purpose, it hid broader sustainability concerns. We urge those in the field not to lose sight of the wider 'ethos of sustainability'.

Keywords: sustainability; artificial intelligence; digital technologies; qualitative research; environmental impact; sustainable development; carboncentric; technocentric

Citation: Samuel, G.; Lucivero, F.; Somavilla, L. The Environmental Sustainability of Digital Technologies: Stakeholder Practices and Perspectives. *Sustainability* **2022**, *14*, 3791. https://doi.org/10.3390/su14073791

Academic Editors: Aimee van Wynsberghe, Larissa Bolte, Jamila Nachid and Tijs Vandemeulebroucke

Received: 28 February 2022
Accepted: 18 March 2022
Published: 23 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Digital technologies (DTs) allow for the datafication of things; they gather, store and process data for various uses, including machine learning technologies and other artificial intelligence (AI) algorithms. DTs are often viewed as a driver for reducing the environmental sustainability of various sectors by providing, for example, information to reduce energy consumption (DTs for environmental sustainability) [1–3]. However, DTs have their own environmental impact. This includes heavy carbon dioxide emissions linked to the energy required to generate and process large amounts of data; the impact on the material environment (e.g., where data centres are constructed); impacts linked to unsustainable practices for extracting minerals for technological components, as well as the manufacturing of these components; and e-waste disposal [4,5]. While likely improvements in energy efficiency and the move to renewable energy will no doubt relieve at least some of these concerns [6], the pace of data driven innovation raises concerns that digital technologies could outpace the world's renewable energy sources, leading to increases in carbon emissions when other sectors are decreasing their energy use [7,8]. Furthermore, data driven solutions have rebound effects, meaning that, while digital solutions in the

near term may appear to offer environmental advantages in terms of increased efficiency, in the long run, this will lead to increases in demand for digital technologies, data storage and analyses, not a reduction [9–11].

While, over the past decade, concerns about the environmental impacts associated with DTs have been increasingly discussed in the digital sector [12,13], more recently, they have received policy and media attention [14–16]. A range of sector associated initiatives and practices are emerging bottom up [17], building awareness about sustainability issues, and helping accelerate understanding in the sector. Many hyperscalers have pledged net-zero commitments, as well as other environmental commitments, often revolving around notions of sustainability [18,19]. Similar to many other sectors, sustainability reporting associated with the environmental impacts of DTs is now becoming a feature of many companies and organisations [20]. As it has been noted that the environmental impacts of DTs should be a topic for discussion in the ethics and politics of digital data and AI literature [21], sustainability is also being increasingly called for by scholars in this field. For example, the European High Level Expert Group on AI states that responsible AI should be sustainable and environmentally friendly [22], Van Wynsberghe argues that sustainable AI is the third wave of AI ethics [23], and Tamburrini (2022) explores the responsibilities of AI scientists for the carbon footprint of their work [24].

At the same time, whilst sustainability is most notably defined as being associated with 'sustainable development' that considers not only financial profits but also social and environmental factors [25], the term 'sustainability' has a long history of being a complex and nebulous concept—it has different pillars and dimensions, and a multiplicity of meanings, which are neither stable nor fixed [26]. This is because the meaning of policy relevant terms such as 'sustainability' is not 'hardwired into social reality', but requires interpretation [27], and will be 'read' differently by various audiences [28,29]. Meaning is not universal or determinate, but depends on the context and the perception and interpretation of those who are enacting it [28]. Different scholars place different weight over whether sustainability is more about economics, ecology, or social science, as well as whether it should address technology, resources, waste, pollution and/or other issues [30–32]. Furthermore, meanings may be inferred from a philosophical, technical, or ethical perspective [30]. Many sustainability initiatives are also perceived to be an attempt at 'green-washing', and concerns have been raised about sustainability initiatives being reliant on the objectification of carbon driven by neoliberal markets [33,34]. While, for some, this heterogeneity of the sustainability concept allows it to act as an umbrella concept [35], the ambiguity, confusion and lack of clarity around how to apply the concept is highly problematic [36–39].

We explored how experts working with DTs and sustainability understand and utilise the term in their practice. In particular, our aim was to explore how those considering, researching, governing and/or working on the environmental impacts and/or sustainability of DTs, are drawing on the concept of environmental sustainability in their own work. Our research question was how do stakeholders (academics, researchers, NGOs, policymakers etc) researching, governing or working on the environmental impacts of DTs, utilise environmental sustainability concepts? We applied a combination of bibliometric analysis and 24 interviews with key stakeholders from the digital technology sector.

Findings show stakeholders have broad conceptual understandings of the term sustainability and its relation to the environmental impacts of DTs. However, in practice, environmental sustainability tended to be associated with technology based and carbon-centric approaches. These approaches have been criticised in other fields because they hide broader sustainability concerns. This was evident in our findings too. While narrowing conceptual understandings of environmental sustainability was perceived to have a practical purpose, technology based and carboncentric approaches alone cannot address sustainability. Those in the field must not lose sight of the wider ethos of sustainability, though, at the same time, to do so requires changes in the socioeconomic and political climate.

2. Materials and Methods

2.1. Interviews

The inclusion criterion for interviews was having expertise in the field of digital environmental sustainability [17]. Sampling was purposive. Participants were identified via two approaches. First, bibliometric analysis identified key academics researching in the field of digital environmental sustainability (see below). Second, snowballing that included (a) asking key stakeholders in the field known to the authors to provide a list of relevant individuals working in the sector, and (b) asking interviewees if they knew other individuals who would be useful to talk to. Seventy-three individuals were contacted via email and asked to participate in an interview. Twenty-four individuals accepted the invitation and interviews were conducted online or on the phone. Interviews were conducted primarily with individuals based in the UK and continental Europe, though also in continental North America (n = 3) and Australia (n = 1). Interviewees self-reported as being associated with a range of sectors (Table 1; note, some individuals self-reported as crossing more than one sector). Interviewees were primarily male (n = 17; see limitations). No other demographic criteria were collected because, as is often the case with expert interviews, we were more interested in exploring the different ways that stakeholders were drawing on the concept of environmental sustainability in their own work, rather than correlating this to particular demographic criteria.

Table 1. Self-reported sectors of interviewees. Some individuals self-reported as being associated with more than sector, and in the table they have been marked as both, which explains the higher total compared to the number of interviewees.

Sector	Number of Individuals Self-Reporting as Belonging to a Sector
Academic researchers (computer scientists, sustainability experts, social scientists, engineers, societies)	10
Industry (commercial, corporate, spin offs; directors, researchers, alliances/organisations)	8
Data centre representatives or consultants, or involved with the sector's markets	5
Policymaker/consultant (funding bodies, organisations associated with standards	5
NGO	1

Interviews were designed to be exploratory, and the interview schedule was broad. Interviewees were asked about their roles and work practices (job role and/or research area; how their role was relevant to DTs (e.g., their research, their interests, their industry etc)), their understanding of the term sustainability; how environmental sustainability was being incorporated into their own practices and their perceptions about how it was being drawn upon in the digital sector more generally; and the actual and perceived challenges associated with this. Interviews were semi- to unstructured. This meant that, by the end of the interview, the interview schedule was covered, but the interviewer would also let the interviewee lead the interview in other directions if they chose, and asked impromptu questions associated with new issues if they were raised. Interviews lasted 32–92 min, with most interviews being between 50–70 min.

Interviews were transcribed by an external transcriber, and these transcripts were analysed via inductive thematic analysis. Thematic analysis is one of the most well established approaches for analysing qualitative data (for example, see [40]). Our inductive analysis approach aligned with the approach taken by Braun and Clarke [41]. GS and a research assistant independently read and re-read each interview transcript to familiarize themselves with the data. Both coders made extensive memos as they proceeded through this step.

GS and the research assistant then independently coded the data. In depth meetings were held on a number of occasions to discuss relevant codes and overlaps. For this paper, codes associated with the meaning ascribed to the term sustainability were considered relevant for analysis. GS combined the codes and drew on them to develop themes.

2.2. Bibliometric Analysis

On 12 February 2021, articles in Web of Science published between 2016–2021 were searched using four separate keyword string combinations (Supplementary Materials: Table S1). Keyword strings were developed deductively and inductively through an iterative process, and combined a range of keywords relating to the environment, sustainability and the need for energy efficiency, alongside a range of keywords related to digital technologies. Specific keywords particularly 'noisy' during the inductive searches were removed from this main keyword string, and created as separate strings. In total, the combination of all the keyword strings returned 4598 articles.

Titles and abstracts of articles were reviewed for the inclusion criterion—articles that explored or discussed the environmental impacts of DTs. Initially, the inclusion criterion was independently applied to 100 articles by a research assistant, GS and FL to ensure consistency of approach. Discrepancies emerged and were discussed, with a refinement of review that included exclusion criteria. The process was repeated twice more until consistent. An exclusion criterion included articles that discussed environmental impacts, but those environmental impacts were not specifically associated to the digital aspects of the technology. The research assistant then applied the inclusion/exclusion criterion to the remaining articles. Following this, 489 articles remained.

A coding schedule and manual was deductively developed to analyse the articles. The coding schedule was applied to 30 articles by a research assistant and GS. Discrepancies were discussed and the coding schedule was inductively refined to ensure consistency. Codes included: main academic field of the research based on the keywords of Web of Science; explicit reference to climate or sustainability risk or problem as a motivation or justification for research; type of DT that was the subject of analysis; sustainability or other issue addressed; approach used to address the issue (Supplementary Materials: Table S2). GS and the research assistant duplicate coded a further 10 articles and achieved 96% similarity between all codes coded. GS and the research assistant discussed the discrepancies and slight changes were made to the coding schedule/manual. The remainder of the articles (449 articles) were coded by the research assistant with the updated coding schedule/manual.

2.3. Limitations

First, Web of Science contains bibliographic information from a set of more than 7500, primarily English language journals. Fields that publish heavily in the journal literature, such as the sciences, are better covered than those that do not, such as philosophy. Therefore, some subject areas are poorly covered, including business and education. Nonetheless, Web of Science is one of the broadest academic databases, covering a wide range of subjects. Second, our sample of 24 stakeholders did not capture the whole digital sustainability landscape. However, this was not our intention. Rather, we aimed to speak to key stakeholders who could give us a better understanding of the issue. Though we do note that none of our interviewees were residing in low to middle income countries, and further research should explore the views of such stakeholders. Furthermore, we did not have an equal gender balance of interviewees. When identifying potential participants for interview, they were mainly males, most likely reflecting the gender bias of the workforce (see, for example, [42]). Further research should aim for a higher female representation.

3. Results

3.1. Sustainability as a North Star

Interviewees considered sustainability as a universally accepted value that guides people's actions. At the same time, interviewees viewed sustainability as an ill-defined abstract concept that is hard to measure and action.

When interviewees were asked about how they would define the term sustainability, or how sustainability was defined within their sector, they framed the concept using the well-established notion of encompassing economic, environmental and social factors into the development processes: *'you have to look at it [development] systemically, for a start, you can't look for things in isolation. It must include the environmental, it must include the social, and it must include the economic'* (interviewee 13). At the same time, interviewees echoed discussions in the literature that have questioned the term's usefulness as a metric to align themselves with when considering their own development or business practices: *'I think that sustainability is . . . it's almost too broad to be useful as a term'* (interviewee 12). Participants pointed to the well established confusion about what the term 'actually means' in practice: *'I think there's still a debate on what do you actually mean with all this . . . If I talk to the economists here . . . I have a different view, and they have a different view . . . '* (interviewee 1). Interviewee 19 reflected on the different understandings of sustainability promoted by 'zealots' and 'pragmatists': *'there's a definition created by the group that I refer to as the zealots . . . people that use sustainability . . . as a religion . . . Then there's the pragmatists . . . So . . . it's . . . how you really manage this'* (interviewee 19).

This left questions about how individuals and businesses should approach the notion of sustainability, understand what this concept means and/or decide how they need to change their practices to achieve it. Furthermore, interviewees were concerned that, because the concept of sustainability is opaque, it can be used by industry as a greenwashing strategy. One of the respondents, participant 6, explained how encapsulating everything under sustainability becomes problematic because, with so many definitions of sustainability circulating, it can lead to confusion in terms of standards and practices, which businesses can play to:

> *'no one really properly defines what sustainability actually means when it comes to reporting. And then when there is reporting, it's not standardised. It's not consistent. And it is often just hidden in environmental reports that look really nice and have a lot of good photos but are very difficult to compare. So as a, as a researcher, or a consumer, or a business trying to make a decision on what is, what is the more sustainable product, if you're trying to make a comparison between different options, it's basically impossible to do'.*

Despite this, many interviewees had a general sense that sustainability, and the more specific sustainable development goals (SDGs) [17], were a 'target to reach' (interviewee 14). They were a universal value—something shared by different people—a 'North Star' (interviewee 22) that brought consensus to the field in terms of aspiring towards an ideological ethos of sustainability. This can be considered as a Kantian regulative ideal [43,44]—a goal to be approached, that we may never reach but that guides our actions:

> *'sustainability . . . [is] about people, economy and the planet . . . the common language that all of us speak are the sustainable development goals and that's really been the main basis of our work . . . it doesn't matter what sector, what region of the world. We have these, we have this North Star . . . '* (interviewee 22);

> *'there is more consensus now, also in the line of the SDGs that are another big international agreement and target to reach . . . and no one should be left behind. Besides the economy . . . there is the big issue of environment and social justice'* (interviewee 14)

However, beyond this abstract conceptualisation of sustainability as a guiding principle, in decision making and agenda setting in the DT sector, the principle of sustainability—and more specifically, environmental sustainability—was actioned in a variety of ways that often were narrowed down to a single or two dimensions.

3.2. Practices of Narrowing down Environmental Sustainability to a Single or Two Dimensions

Participants stressed that a combined approach to addressing environmental sustainability was needed—one that focused on carbon reduction, but also on decreasing the use of resources through increased efficiency, addressing water consumption, and promoting biodiversity. In the below extract, one participant (interviewee 4) discusses the need to optimise both water and energy efficiency to ensure stability in the operations of a data centre, as well as consider carbon emissions and the circular economy—*'it's not just a single dimension'* that needs focussing on, explained interviewee 14, *'but multiple dimensions'*. Participant 4 remarked:

> *'people are starting to talk about water usage effectiveness, carbon usage effectiveness. People are starting to measure efficiency in terms of how much performance you get from the energy that you use rather than it having anything to do with cooling . . . Circular economy is becoming an increasing area of focus . . . things like load balancing as well is another area of interest . . . '*

Participant 21 also described the development of a data centre through sustainable practices that included considerations of carbon emissions, efficiency and low waste:

> *'this is a great story of . . . An end-to-end sustainability offering . . . This company takes a containerized data centre that might fit between say ten or 8 to 12 racks of hardware. They sit next to a greenhouse . . . They use the heat from that container . . . so they sell the heat to the farmers while they're producing a distributary grid . . . and all of that hardware is second user decommissioned . . . that has very low Scope 3 emissions because it's already been in the ecosystem, reusing the heat for agricultural purposes and running highly efficient hardware'*.

Although most interviewees were aware that the concept of environmental sustainability had these multiple dimensions, there was a perception that this combined approach was not mainstream. They worried that the concept was often narrowed down to only one or two actions pertaining to environmental sustainability—most notably associated with either the efficiency dimension or carbon emission. For example, interviewee 8 explained that focussing on carbon dioxide/greenhouse gas (GHG) emissions was a key concern: *'if a company says it [talks about sustainability] then it usually means, how can we reduce our greenhouse gas emissions? . . . That's what they would be focussing mainly on'*.

3.3. Increasing Efficiency and/or Decreasing Carbon Emissions as a 'No Brainer' in a Business Sense

Interviewees explained that, with so many business pressures to remain profitable, it made sense for businesses to begin addressing environmental sustainability in the area that would have limited effect on finances, and the relationship between increasing efficiency and the financial goals of a business made these appealing places to start. Increasing the efficiency of DTs has been a historic 'business driver' for the ICT industry because of its inextricable link to saving money, long before the environmental sustainability movement. Interviewees explained that it was a 'no brainer' (interviewee 10) for companies to focus on efficiency gains to become (in their perceptions) more environmentally sustainable:

> *'when it comes to, you know, environmental impact or sustainability in data centres . . . business drivers behind these are . . . in terms of energy efficiency . . . Rather than spending 100 megawatts of energy on . . . my energy consumption . . . I can only spend 50, well actually that puts me in much better position . . . when you do the numbers'* (interviewee 11);

> *'green IT, sustainable IT, was originally about making data centres more efficient . . . [it was also about companies wanting] to say, "we want to sell you green data centres . . . and it's gonna save you money so buy it from us." That's . . . the kind of, sell point'* (interviewee 8)

Addressing carbon dioxide/GHG emissions was also perceived by interviewees to make appealing business sense, especially when it was—and it often was—tied to increased efficiency (increased efficiency meant less energy used, which meant fewer carbon emissions): *'there is a direct relationship between financial costs and carbon emissions. The lower the financial cost of your solution . . . the lower your carbon footprint will be'* (interviewee 13); *'there's . . . good business reasons that they're doing this . . . they're not doing it just purely out of, "Oh you know, we want to be environmentally friendly"'* (interviewee 10). Interviewees also narrated a range of other reasons they perceived individuals and companies to be pursuing environmental sustainability through a focus on carbon/GHG emissions. They were aware of the various benchmarks that focused environmental sustainability efforts on carbon emissions, and these were seen to provide a goal for businesses and organisations to 'aim towards': *'it's not easy but it's easier if you sort of give people a kind of benchmark and say, "Right, you really should aim towards that"'* (interviewee 7). Regulations in the sector were perceived to be forcing industries to assess, monitor and minimise their carbon/greenhouse gas emissions: *'[the focus on carbon has] a lot to do with government regulation forcing those standards'* (interviewee 6). Some interviewees considered how this needed to be considered in a geopolitical context, because some countries were more set up to address these issues than others, and had more accommodating regulatory environments: *'it's a little bit easier [in Europe to consider these issues]'* (interviewee 22). Finally, strong pressure to consider these issues was considered to have come from peers (*'everyone's worrying about emissions because they have to, because of law, but also because everyone else is'* (interviewee 12)), as well as from consumers:

> *'there's been growing consumer interest in this as well . . . companies . . . probably want to develop or maintain an image that they are, you know, not polluting or green and that the services people are using are powered by clean electricity'* (interviewee 10)

These factors often led to a carboncentric approach to environmental sustainability, despite an understanding of its limitations. For example, interviewee 23 explained how the global sense of urgency to reach net-zero carbon emissions meant that their work was focusing primarily on reducing carbon dioxide emissions: *'the expectation is that the EU [European Union] require that 2050 the whole of Europe is carbon neutral so we are looking mainly at the carbon emissions, the $C0_2$ emissions of electricity'*. In the extracts below, two interviewees provided further examples of where a choice was made to focus on a carboncentric construction of sustainability, despite an understanding that sustainability is a much broader issue. First, interviewee 3 explained that their reasoning for refining the scope of a large report around issues of carbon, rather than the environment as a whole, was based on the fact that there was more media attention on this issue. Second, interviewee 18, who worked with the sustainability department of an organisation, explained how, while they were looking at more than carbon issues when focusing on environmental and social sustainability, it was easier to talk about carbon to their clients because they were more familiar with the issue. In both instances, while they themselves had a broader understanding of sustainability, they were perpetuating a construction of sustainability that was carboncentric:

> *'if we took all the environmental impact, we would probably have published a 10,000-page document. So, we had to refine . . . and emissions seemed to be a really interesting area because there was quite a, a lot of controversy in the media, a lot of uncertainties . . . Something that's come up in some of our follow-on meetings is that . . . environmental sustainability is . . . more multidimensional . . . thinking about biodiversity'* (interviewee 3);

> *'I'm just looking from the carbon, because when I talk to customers it was the easiest question, because now it's like everyone wants to have a carbon index on what they are buying . . . but internally we are tracking many other indicators [of sustainability]'* (interviewee 18)

3.4. A Carboncentric and Technocentric Approach: A Very Narrow Frame That Misses the Bigger Picture

Some interviewees were concerned that a focus on benchmarks, regulations and profit was leading to the concept of environmental sustainability being viewed solely in terms of metrics, and that, together, carbon emissions and/or efficiency gains were being conflated with environmental sustainability as a normative concept. Interviewee 18 described how, in terms of efficiency *'many people think it's [sustainability is] . . . not taking all the pillars of sustainability, just more focussed on efficiency than really . . . caring for environments or caring for the planet'* (interviewee 18). This worried interviewees, who viewed sustainability as a broader concept: *'we can help them [our clients] with the energy efficiency . . . If they want to put it within . . . the sustainability banner, fine, we'll support it, but, but I'll be really reluctant to call that sustainability'* (interviewee 11). Interviewees were concerned that approaching sustainability in this narrow sense detracted attention from other elements associated with sustainability—not just water and waste, but also more hidden issues, such as those related to toxins that may be produced during electronic manufacturing processes [45]: Interviewee 7 remarked on the carboncentric drivers pushing companies to account for their carbon emissions:

> *'this is the, the danger of drivers in a way, isn't it? . . . They can almost forget about the other stuff, you know, and just because you're being presented with ways of achieving these targets and you're going to get a pat on the head for doing it, and, and it's harder to think about the other stuff . . . [for example] . . . you might have something that's very low carbon, but actually it's incredibly toxic to water or, you know, human beings . . . you need to look at the biggest picture possible . . . Some people . . . think about energy . . . and that's it'.*

Participant 17 concurred: *'certainly carbon accounting is important, but . . . once you focus only on that . . . you create a very narrow frame . . . carbon is not toxicity . . . there are always problems [with what] . . . you choose to count and what gets left'.*

However, there was also a realisation that, while addressing environmental sustainability was viewed as something relatively achievable for larger companies, trying to be environmentally sustainable posed difficulties for smaller companies. Participant 21 provided an example:

> *'[a partner] ran the numbers for a company that was looking at moving [to be more sustainable], they say "Look, I can save [you] 4% on [your] energy bill" . . . And they say "It's just not worth it . . . it needs to be 25 to 30% gains" . . . And that's a lot. And then, even if [it is that much] Dell and HP come along and say "Well I'll just cut my price, 15%" . . . And the guy goes "Oh great, I don't have to learn anything new, I'll just stick with [them]" . . . Everyone has a different measurement on their sustainability'.*

As interviewee 19 explained, to be properly environmentally sustainable required having a good sustainability plan, and this, described this interviewee, *'takes a lot of resource'*. Other interviewees concurred: *'to move towards a sustainable infrastructure, it's gonna be huge and it's gonna be costly and time consuming'* (interviewee 21); *'it's costly, it's expensive, nationwide there may not be mechanisms in place, so there's just no motivation for some companies to do that'* (interviewee 22). Interviewee 1 reflected on this cost when describing the environmental impacts of mining, remarking on how these impacts are not often addressed because of the resource required:

> *'one of the most damaging things we see in, in the supply is, is actually the left waste of metal mining . . . we need to control those waste parts, or waste for thousands of years . . . the thing we should do is put back as it was, but we, that will cost too much . . . '* (interviewee 1)

3.5. Broader Conceptualisations of Sustainability: The Economic and Social

While not the focus of our interviews, a number of our participants—experts in sustainability—discussed environmental sustainability in relation to the pillars of economic

growth and social justice (*'we like to emphasise there are other environmental impacts than just $C0_2$ emissions . . . there is also a hoard of social issues'* (interviewee 23)), because economic, environmental and global justice issues were perceived to be connected: *'they're very hard to decouple'* (interviewee 5). Participant 15, whose business was repurposing hardware as part of a circular economy, reflected on how they were trying to bring social sustainability considerations into their business to address this missing social component:

> *'we try to . . . think about how we can bring this recertified equipment into the parts of the world that are most disadvantaged . . . people whose children . . . are deep in mines mining cobalt that goes into electronic devices, under the most hazardous working conditions, . . . [or] . . . [whose] . . . children, they are making, you know, 20 cents a day picking through [piles of e-waste] to find a piece of gold'.*

Concerns were raised that a focus on environmental sustainability was deprioritising geopolitical social justice issues. Interviewee 23 explained how environmental sustainability benchmarks and metrics in Europe have led to the obscuring of sustainability issues associated with global environmental inequalities:

> *'in Europe it looks like our economy is reaching decoupling . . . Decoupling meaning that you get more profit whilst your environmental impact is lower. So, we are kind of fooling ourselves to saying that Europe is doing good . . . it only means that the environmental impacts happen elsewhere . . . where the materials have been mined or processed . . . our current metrics are not sufficient to show this inequality'* (interviewee 23)

Interviewee 1 described how they 'bumped into' social justice issues when considering environmental sustainability, but that, while sustainability measures exist for the latter, their experience led them to believe that less work had concentrated on the social justice component:

> *'social, I think that is something that, that we sort of bumped into very early on when we started [looking at environmental sustainability] . . . Where are you sourcing your metals? . . . Tantalum used to be the first one we talked about, coming from Africa, it's a large source . . . now cobalt is the main metal . . . [you cannot address this in terms of] life cycle assessment, so you have to treat that in some other way [as there is no other way to assess it] . . . especially when you're talking about the, the impact on . . . human health'* (interviewee 1)

Finally, one interviewee highlighted the tension between environmental and social justice, which were not always aligned, and their concern that the drive to be environmentally sustainable could lead to social injustices. Interviewee 17 described:

> *'those [environmental and social issues] are incommensurable, right. I want carbon mitigation, I do, but I also do not want to continue to colonise . . . the land of indigenous people [where this is occurring]. That's a major contradiction right, between carbon accounting and mitigation, and all that kind of stuff and social justice'.*

3.6. Bibliometric Analysis

Our bibliometric analysis echoed our interview findings. Nearly all 489 analysed articles framed their approaches within the broader discourse of sustainability (90%). Over four-fifths of the articles adopted a technocentric approach to addressing the environmental impacts of DTs (85%), i.e., they analysed or explained an environmental impact that was of concern and outlined a potential technology based solution ('solve and explain': Figure 1). This can be explained by the fact that most articles were classified by Web of Science as coming from a technology focused discipline: computer science (45%), engineering (19%) or science and technology (12%). Only 1% of the articles were classified by Web of Science as social science.

Within the articles adopting a technocentric based approach, about two-thirds of the authors were solely focused on trying to decrease the energy consumption of DTs through energy efficiency improvements (see Figure 1; 'energy'). Just under a quarter of the articles focused on addressing one or more environmental impacts (Figure 1; 'general'), and these

were predominantly associated with improving energy efficiency and reducing carbon emissions (not shown in diagram). This meant that, together and reflecting the interview findings, nearly all technocentric articles we analysed were trying to address either the energy efficiency of DTs and/or their carbon emissions. Environmental impacts associated with e-waste and other issues, such as biodiversity and water consumption, have received little attention in the literature.

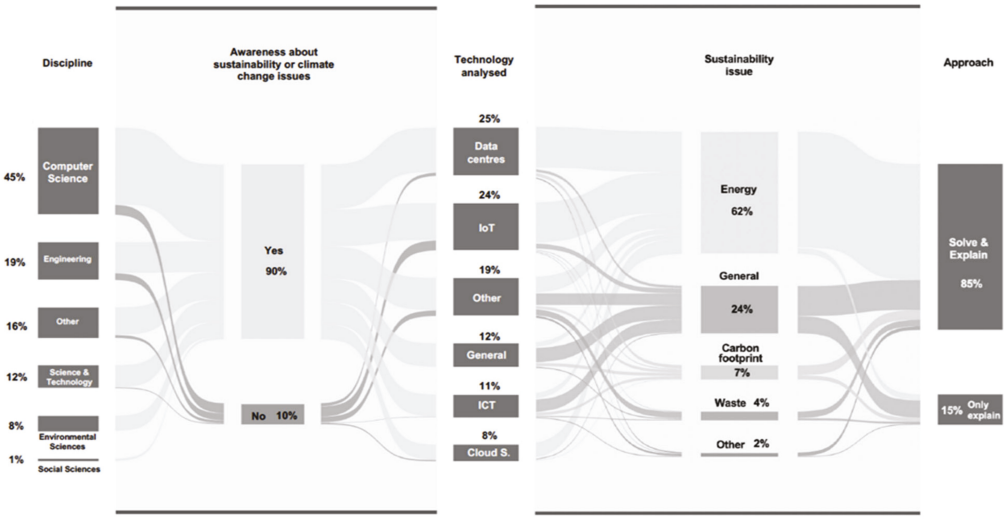


Figure 1. Sankey diagram of research articles published between 2016–2021 that explore the environmental impacts of digital technologies. Of the 489 articles retrieved from our bibliometric analysis, they were coded into five themes: main academic field of the research; explicit reference to climate or sustainability risk or problem as a motivation or justification for research; type of digital technology that was the subject of analysis; environmental/sustainability issue addressed; approach used to address the issue.

4. Discussion

Interviewees had broad understandings of the concept of sustainability and its association with the environmental impacts of DTs, however, the way the term was operated in practice varied. While, at the higher level, there was general consensus on what was meant by the term sustainability—albeit with some interviewees pointing to the vagueness of the concept; in practice, the meanings applied to the term differed. Some practices dominated more than others, with a focus on addressing carbon emissions and/or energy efficiency predominating. These interview findings were supported by our bibliometric analysis, which suggested that most published academic articles on the environmental impacts of DTs focused on improving energy efficiency and/or carbon emissions through technocentric approaches. Few articles had taken a wider approach to exploring the environmental sustainability of DTs. Together, both sets of findings suggest that, in this field/sector, sustainability is often associated with narrow, technocentric and/or carboncentric approaches.

It is understandable that a broad regulative principle such as sustainability needs to be narrowed for practical purposes in the DT sector. Moreover, it is understandable that, in academia, this narrowing will reflect those disciplines working in the field (technology driven computer science and engineering), and in the broader sector this will reflect other competing business and sociopolitical agendas. However, the shift from theory to practice has some implications. For example, it renders the concept of sustainability open to interpretation, so it can be used to justify different types of interventions (efficiency, carbon accounting, biodiversity, water consumption, etc). In the broader sector, it also permits

the concept to sometimes be a catch all phrase that allows a wide variety of changes to be categorised as sustainability as a means of justifying business as usual approaches (for example, efficiency gains in line with historical business models). This produces ambiguities and tensions with significant implications for social, political, and ecological change [34]. In light of this, we note two implications associated with the narrowing down of the sustainability concept that raise concern.

First, with a 'spotlight' on carbon emissions, and a range of metrics and regulations associated with the climate agenda, it is unsurprising that this has become an academic and stakeholder focus. Using metrics is surely important, especially when it comes to carbon metrics associated with the development and deployment of AI. In fact, there are urgent needs to generate transparency frameworks that clarify how the digital sector is addressing issues of governability and standards to guide best practice in carbon accounting, social awareness of digital technologies' environmental impacts and response measures to complex dynamics emerging from the rapid development of DTs. Furthermore, once the sustainability properties of digital technologies are identified and standardised, they have to be regulated in ways that increase transparency and accountability of environmental impacts. However, using metrics too much means that sustainability is narrowed to something that is documented as 'done' [46]. This loses an important aspect of sustainability, the fact that it guides actions as a 'regulative ideal' or, as one interviewee stated, a '*North star*'. Our interviews suggest that this guiding role of the concept of sustainability, albeit recognised by our interviewees and seen in the bibliometric analysis, has not become a widespread aspect of their practices in the sector. Vitally, relying on metrics means that, when there are no metrics, issues are invisible and therefore not considered [5,33]. This was evident in one of our respondent's remarks, who explained how the focus on environmental sustainability is hiding social justice issues, and that this also correlates to having metrics to measure the former but not the latter. This narrowing down also implies neglecting important aspects pertaining to the concept of sustainability, such as issues related to other environmental impacts besides carbon emissions [47]. This was evident in our bibliometric analysis. When the meaning of sustainability is narrowed down, the bigger picture gets lost.

A second implication of the narrowing down of the concept of sustainability in specific practices is that technocentric approaches focusing on efficiency gains are limited in scope. They fail to consider issues associated with rebound effects, that is, that efficiency gains will likely lead to increases in demand for data storage and analyses, not a reduction [9–11]. This means that if the digital sector really wants to attain sustainability it is likely to be complex, difficult and costly, and also require a shift in practices. It is unlikely that technically driven solutions will be able to carefully consider what sustainability means for society at different levels. This means that companies may not always be able to conciliate between their business needs and broader sustainability goals. As our interview findings show, in particular, a conciliation between business models and sustainability goals often happens when sustainability is understood and enacted in terms of efficiency. However, to take sustainability seriously means prioritising sustainability at least equally, if not more so, than financial drivers. Unfortunately, current competition between business and sociopolitical agendas makes this difficult for individual companies. The little academic literature outside of technocentric approaches also provides little guidance on how to address these issues. To drive sustainability in the sector, changes are needed in the sociopolitical and economic climate, and, in fact, a number of our participants pointed to the need for this. This requires viewing AI, 'not as benign or neutral but as a reflection of capitalism and an instrument of power' [48], such that to address AI sustainability requires addressing key political and economic issues tied to economic growth and a lack of regulation in the drive for power and consumption [49].

In conclusion, as the concept of environmental sustainability has been translated from a 'regulative ideal' into the practices of the digital sector, it has mutated to a technocentric and carboncentric approach that fails to consider broader sustainability issues. We have problematised this in various ways. Our goal is not to criticise those working or researching

in the sector, many of whom are doing their utmost to try and ensure their practices are sustainable, or are working to promote sustainable practices in the sector more broadly. Rather, we wish to expose the implications associated with adopting such a narrow sustainability focus and encourage stakeholders to differentiate between narrowed down activities and a broader sustainability ethos that they adopt in their practices. Stakeholders, including researchers, need to be able to zoom in and out between a narrowed approach (e.g., related to metrics) and a broader 'ethic of sustainability'. Seeing the concept and value of sustainability as having a dual role—both at the higher abstract level, as well as at the more local specific level—allows for splitting the usefulness of the concept into two by simultaneously using it as a way to drive an 'ethos of sustainability', as well as targeted interventions that can have a measurable and impactful change [50]. This two pronged approach allows culture change by instilling an ethos of sustainability in all layers of research, as well as in all layers of an organisation, creating a consistent message and support for this approach, while alongside, targeting specific interventions to provide an opportunity to create test beds at 'pinch points' where sustainability is vitally important [50]. To use an example from our interviewees (which is not discussed in our findings), a company delivering digital goods to people in lower and middle income countries is not sustainable if it does not also ensure appropriate payment of workers in the supply chain. In the research sector, a field in which sustainability is addressed using mainly techno–scientific approaches could benefit from more social science and ethics input. Across the digital technology research sector and industry, actors must build on coordination capabilities and a shared understanding of sustainability that includes the broad ethos of the concept, as well as its functions and limitations. At the same time, we must be careful not to shift responsibility too much onto those working in this area to develop meaningful practices to address environmental sustainability. Rather, meaning and practices associated with an ethos of sustainability need to be embedded within policy and regulatory decisions that are associated with this research and industry sector. To address sustainability issues in the digital technology sector is a collective issue.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/su14073791/s1, Table S1: Key-strings used for searching the academic literature on Web of Science and Table S2: Coding schedule for coding the academic literature.

Author Contributions: Funding acquisition, conceptualization, methodology: G.S. and F.L. Formal analysis, G.S. Writing—original draft preparation, review and editing: G.S., F.L. and L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the British Academy grant number: SRG20\201487 and the EPSRC grant number: EP/V042378/1.

Institutional Review Board Statement: This study received ethics approval from the Oxford University Central University Research Ethics Committee (CUREC): reference: R75723/RE001.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Empirical data is available on request, as long as it adheres to the consent received from our interviewees.

Acknowledgments: We are extremely grateful for our interviewees' time, and also grateful to our research assistant, José Resendiz Garcia, who helped us with the bibliometric analysis.

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

References

1. European Commission. 2019. Available online: https://energy.ec.europa.eu/topics/energy-system-integration/digitalisation-energy-sector_en (accessed on 27 February 2022).
2. Junge, A.L.; Straube, F. Sustainable supply chains—Digital transformation technologies' impact on the social and environmental dimension. *Procedia Manuf.* **2020**, *43*, 736–742. [CrossRef]

3. Nishant, R.; Kennedy, M.; Corbett, J. Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. *Int. J. Inf. Manag.* **2020**, *53*, 102104. [CrossRef]
4. Marks, P. Blood Minerals Are Electronics Industry's Dirty Secret. New Scientist. 2014. Available online: https://www.newscientist.com/article/mg22229734-800-blood-minerals-are-electronics-industrys-dirty-secret/ (accessed on 27 February 2022).
5. Lepawsky, J. *Reassembling Rubbish*; MIT Press: Cambridge, MA, USA, 2018.
6. Giles, M. Is AI the Next Big Climate-Change Threat? We Haven't a Clue. Available online: https://www.technologyreview.com/2019/07/29/663/ai-computing-cloud-computing-microchips/ (accessed on 28 February 2020).
7. Blair, G. *A Tale of Two Citites: Reflections on Digital Pollution*; Patterns: New York, NY, USA, 2020. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7427557/ (accessed on 27 February 2022).
8. Freitag, C.; Berners-Lee, M.; Widdicks, K.; Knowles, B.; Blair, G.S.; Friday, A. The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns* **2021**, *2*, 100340. [CrossRef] [PubMed]
9. Alcott, B. Jevons' paradox. *Ecol. Econ.* **2005**, *54*, 9–21. [CrossRef]
10. Hilty, L.M.; Köhler, A.; Schéele, F.V.; Zah, R.; Ruddy, T. Rebound effects of progress in information technology. *Poiesis Prax.* **2006**, *4*, 19–38. [CrossRef]
11. Takahashi, K.I.; Tatemichi, H.; Tanaka, T.; Nishi, S.; Kunioka, T. Environmental impact of information and communication technologies including rebound effects. In Proceedings of the IEEE International Symposium on Electronics and the Environment, Scottsdale, AZ, USA, 10–13 May 2004; Conference Record, pp. 13–16.
12. Ensmenger, N. The environmental history of computing. *Technol. Cult.* **2018**, *59*, S7–S33. [CrossRef]
13. Belkhir, L.; Elmeligi, A. Assessing ICT global emissions footprint: Trends to 2040 & recommendations. *J. Clean. Prod.* **2018**, *177*, 448–463. [CrossRef]
14. Schwartz, J. Tech's Environmental Impact and What You Can Do About It. *New York Times*, 2019. Available online: https://www.nytimes.com/2019/11/06/technology/personaltech/techs-environmental-impact-and-what-you-can-do-about-it.html (accessed on 27 February 2022).
15. Gilmore, M. Expert and Stakeholder Consultation Workshop 2018. Available online: https://www.eclab.uel.ac.uk/?p=5301 (accessed on 27 February 2022).
16. Department for Environment, Food & Rural Affairs. Greening Government ICT and Digital Services: 2019 to 2020 Annual Report. 2021. Available online: https://www.gov.uk/government/publications/greening-government-ict-and-digital-services-2019-to-2020-annual-report (accessed on 27 February 2022).
17. Sustainable Web Manifesto. Available online: https://www.sustainablewebmanifesto.com/ (accessed on 27 February 2022).
18. Microsoft. Available online: https://www.microsoft.com/en-us/corporate-responsibility/sustainability/report (accessed on 27 February 2022).
19. Amazon. Available online: https://sustainability.aboutamazon.com/pdfBuilderDownload?name=sustainability-all-in-june-2020 (accessed on 27 February 2022).
20. Armstrong, A. Ethics and ESG. *Australas. Account. Bus. Financ. J.* **2020**, *14*, 6–17. [CrossRef]
21. Lucivero, F. Big data, big waste? A reflection on the environmental sustainability of big data initiatives. *Sci. Eng. Ethics* **2020**, *26*, 1009–1030. [CrossRef]
22. European Commission. High-level expert group on artificial intelligence. In *Ethics Guidelines for Trustworthy AI*; European Commission: Brussels, Belgium, 2019. Available online: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1 (accessed on 27 February 2022).
23. Van Wynsberghe, A. Sustainable AI: AI for sustainability and the sustainability of AI. *AI Ethics* **2021**, *1*, 213–218. [CrossRef]
24. Tamburrini, G. The AI carbon footprint and responsibilities of AI scientists. *Philosophies* **2022**, *7*, 4. [CrossRef]
25. World Commission on Environment and Development. *Our Common Future*; Oxford University Press: Oxford, UK, 1990.
26. Dobson, A. *Ustice and the Environment: Conceptions of Environmental Sustainability and Theories of Distributive Justice*; Oxford University Press: Oxford, UK, 1998.
27. Wagenaar, H. Interpretation and intention in policy analysis. In *Handbook of Public Policy Analysis: Theory, Politics, and Methods*; Fischer, F., Miller, G.J., Eds.; Routledge: New York, NY, USA, 2007.
28. Yanow, D. The Communication of policy meanings: Implementation as interpretation and text. *Policy Sci.* **1993**, *26*, 41–61. [CrossRef]
29. Yanow, D. *How Does a Policy Mean? Interpreting Policy and Organizational Actions*; Georgetown University Press: Washington, DC, USA, 1997.
30. Vucetich, J.A.; Nelson, M.P. Sustainability: Virtuous or Vulgar? *BioScience* **2010**, *60*, 539–544. [CrossRef]
31. Nelson, M.P.; Vucetich, J.A. Sustainability science: Ethical foundations and emerging challenges. *Nat. Educ. Knowl.* **2012**, *3*, 12.
32. Zagonari, F. Environmental sustainability is not worth pursuing unless it is achieved for ethical reasons. *Palgrave Commun.* **2020**, *6*, 108. [CrossRef]
33. Dalsgaard, S. The commensurability of carbon: Making value and money of climate change. *HAU J. Ethnogr. Theory* **2013**, *3*, 80–98. [CrossRef]
34. Blythe, J.; Silver, J.; Evans, L.; Armitage, D.; Bennett, N.J.; Moore, M.-L.; Morrison, T.H.; Brown, K. The dark side of transformation: Latent risks in contemporary sustainability discourse. *Antipode* **2018**, *50*, 1206–1223. [CrossRef]
35. Lélé, S.M. Sustainable development: A critical review. *World Dev.* **1991**, *19*, 607–621. [CrossRef]

36. Oermann, N.O.; Weinert, A. Sustainability ethics. In *Sustainability Science*; Heinrichs, H., Martens, P., Michelsen, G., Wiek, A., Eds.; Springer: Dordrecht, The Netherlands; Berlin/Heidelberg, Germany; New York, NY, USA; London, UK, 2016; pp. 175–192.
37. Shearman, R. The meaning and ethics of sustainability. *Environ. Manag.* **1990**, *14*, 1. [CrossRef]
38. Holmes, R. Justifying Sustainable Development: A Continuing Ethical Search. *Glob. Dialogue* **2002**, *4*, 103–113.
39. Vogt, M.; Weber, C. Current challenges to the concept of sustainability. *Glob. Sustain.* **2019**, *2*, e4. [CrossRef]
40. Saldana, J. *The Coding Manual for Qualitative Researchers*; Sage Publications: Thousand Oaks, CA, USA, 2013.
41. Braun, V.; Clarke, V. *Thematic Analysis: A Practical Guide*; SAGE Publications: Thousand Oaks, CA, USA, 2021.
42. Leavy, S. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In Proceedings of the ACM/IEEE 1st International Workshop on Gender Equality in Software Engineering, New York, NY, USA, 28 May 2018.
43. Emmet, D. Regulative ideals: Kant. In *The Role of the Unrealisable*; Palgrave Macmillan: London, UK, 1994.
44. Friedman, M. Regulative and constituative. *South. J. Philos.* **1992**, *30*, 73–102. [CrossRef]
45. Liu, R.; Gailhofer, P.; Gensch, C.-O.; Köhler, A.; Wolff, F. *Impacts of the Digital Transformation on the Environment and Sustainability: Issue Paper under Task 3 from the "Service Contract on Future EU Environment Policy"*; Öko-Institut for the European Commission: Freiburg, Germany, 2019. Available online: https://ec.europa.eu/environment/enveco/resource_efficiency/pdf/studies/issue_paper_digital_transformation_20191220_final.pdf (accessed on 27 February 2022).
46. Samuel, G.; Broekstra, R.; Gille, F.; Lucassen, A. Public trust and trustworthiness in biobanking: The need for more reflexivity. *Biopreserv. Biobank.* 2022, *ahead of print*. [CrossRef] [PubMed]
47. Available online: https://sdgs.un.org/goals (accessed on 27 February 2022).
48. Dauvergne, P. *AI in the Wild*; MIT Press: London, UK, 2020.
49. Reich, R. *Saving Capitalism: For the Many, Not the Few*; Icon Books: London, UK, 2017.
50. Hunt, D.F.; Bailey, J.; Lennox, B.R.; Crofts, M.; Vincent, C. Enhancing psychological safety in mental health services. *Int. J. Ment. Health Syst.* **2021**, *15*, 33. [CrossRef] [PubMed]

sustainability

MDPI

Review

A Survey on Sustainable Surrogate-Based Optimisation

Laurens Bliek [1,2]

[1] School of Industrial Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands; l.bliek@tue.nl
[2] Eindhoven Artificial Intelligence Systems Institute, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Abstract: Surrogate-based optimisation (SBO) algorithms are a powerful technique that combine machine learning and optimisation to solve expensive optimisation problems. This type of problem appears when dealing with computationally expensive simulators or algorithms. By approximating the expensive part of the optimisation problem with a surrogate, the number of expensive function evaluations can be reduced. This paper defines sustainable SBO, which consists of three aspects: applying SBO to a sustainable application, reducing the number of expensive function evaluations, and considering the computational effort of the machine learning and optimisation parts of SBO. The paper reviews sustainable applications that have successfully applied SBO over the past years, and analyses the used framework, type of surrogate used, sustainable SBO aspects, and open questions. This leads to recommendations for researchers working on sustainability-related applications who want to apply SBO, as well as recommendations for SBO researchers. It is argued that transparency of the computation resources used in the SBO framework, as well as developing SBO techniques that can deal with a large number of variables and objectives, can lead to more sustainable SBO.

Keywords: surrogate-based optimisation; surrogate model; sequential model-based optimisation; Bayesian optimisation; Green AI; machine learning; sustainable AI

Citation: Bliek, L. A Survey on Sustainable Surrogate-Based Optimisation. *Sustainability* **2022**, *14*, 3867. https://doi.org/10.3390/su14073867

Academic Editors: Tijs Vandemeulebroucke, Aimee van Wynsberghe, Larissa Bolte and Jamila Nachid

Received: 24 February 2022
Accepted: 22 March 2022
Published: 24 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

While undergoing the climate crisis with no easy global solution in sight, some have turned their attention to Artificial Intelligence (AI) as a key technology in the pathway towards a sustainable future [1,2], causing the rise of new AI initiatives such as Climate Change AI [3] and AI for Good [4]. Though it is unlikely that any one technology will be the solution to one of humanity's greatest challenges, Machine Learning (ML) in particular is seen as a technique with potentially a large positive impact on the United Nations' Sustainable Development Goals [5] (SDGs), for example in forecasting extreme weather events, balancing supply and demand of renewable energy systems, designing zero-emission transportation systems, identifying woodlands from satellite data, and fault detection in wind turbines [1,6–9]. This does not come without cost, however, as it turns out ML is a technology with a substantial carbon footprint [2,10]. This has also been recognised in several sub-fields of ML such as natural language processing [11] and Automated ML [12] (AutoML). However, one sub-field of ML, namely, Surrogate-Based Optimisation (SBO), has not yet achieved similar positive or negative attention, even though it is particularly suitable for reducing energy consumption. In fact, SBO techniques are often especially designed to avoid having to run computationally expensive software. This is done by using an ML model as a surrogate of an expensive part of an optimisation problem.

This work investigates how SBO, as a subset of AI, can contribute to sustainability, not only by replacing computationally expensive software with more efficient ML models, but also by solving optimisation problems in sustainable applications. This is done by means of a literature review on the intersection of sustainability and SBO. The main purpose

of this literature review is to answer the question: "How is SBO applied to sustainability-related applications?" This question will be answered by identifying sustainability-related optimisation problems that are addressed using SBO, and by identifying in what way SBO is applied in these applications, e.g., which ML models and SBO techniques are used. This research is not meant to be an exhaustive list, but to give an overview of several examples of sustainable applications where SBO is applied. Such an overview can help point the way to combinations of applications and SBO techniques that add significant value. Furthermore, attention is given to the sustainability aspect of SBO itself, but only for the studies considered in this review. Finally, new avenues that can improve the use of SBO for sustainable applications are identified, both for researchers applying SBO in their own application domains, as well as for SBO researchers. The overall goal is to use SBO to improve sustainability aspects in a wide variety of applications, while also advancing research in SBO itself.

An overview of how other optimisation algorithms, namely, metaheuristic algorithms that do not make use of ML, are applied to sustainability-related applications, can be found in [13]. Such algorithms are less suited for expensive optimisation problems, as without a surrogate to guide them, they might require a prohibitively large number of evaluations of the expensive part of the problem. Besides not making use of ML, that study also does not take the sustainability aspects of the algorithms themselves into account. These are two significant differences with this work.

This paper is organised as follows: while this section gives a short introduction on the intersection of AI and sustainability, the sub-field of AI under consideration, namely, SBO, is explained and defined in Section 2. This is followed by Section 3, which investigates sustainability aspects of SBO. It defines three different ways in which SBO can contribute to sustainability. After this, Section 4 explains the method used in this literature review. This leads to a number of studies that are analysed in Section 5, where the used SBO techniques are divided into different frameworks, and several sustainable applications are identified that have benefited from SBO. Finally, Section 6 discusses the results based on the analysis, while providing recommendations for researchers that want to apply SBO to sustainable applications, and for SBO researchers themselves.

2. Surrogate-Based Optimisation

Motivated by the need for efficiently solving expensive optimisation problems, SBO algorithms such as efficient global optimisation [14] and Bayesian optimisation [15] have been developed. These algorithms make use of ML to guide the search for good solutions. The expensive optimisation problems they are designed for can involve computationally demanding simulators, or problems that depend on the outcome of other ML or optimisation algorithms. The problems are also considered 'black-box', meaning that no exact mathematical formulation is available that could be exploited. Examples are designing heat pump systems [16] and diabetes drug manufacturing [17]. The expensive part of the optimisation problem is approximated with a surrogate model in order to reduce the number of expensive computations. The surrogate model is obtained using ML on the available data of the expensive optimisation problem, and is typically updated during the optimisation process as more data becomes available; see Figure 1. The surrogate model is used inside the optimisation process, which makes SBO a powerful combination of ML and optimisation. A recent textbook introduction to SBO can be found in chapter 10 of [18].

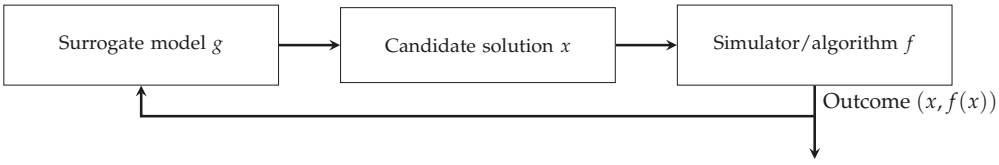


Figure 1. Simplified framework of a typical SBO algorithm. Optimisation is applied to the surrogate model instead of the expensive simulator or algorithm, giving a candidate solution. This solution is evaluated by the simulator or algorithm. The resulting outcome is given to the surrogate model to be updated using machine learning, making it more accurate over time. This gives better candidate solutions and therefore better outcomes.

As there are many synonyms of 'surrogate model' to be found in the literature, such as 'response surface model' or 'metamodel', and many related terms as well, such as 'sequential model-based optimisation', 'Bayesian optimisation', or 'AutoML', it should come as no surprise that an exact definition of SBO is lacking. This work assumes the following broad definition:

Definition 1. *Surrogate-based optimisation (SBO) is an optimisation technique that makes use of a surrogate model obtained using machine learning, usually to replace an expensive part of the optimisation problem.*

Note that this definition makes no distinction between iterative and non-iterative methods, or between surrogate-based and surrogate-assisted methods. The corresponding optimisation problem is given as

$$\operatorname{argmin}_{x \in X} \; f(x), \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}^m$ consists of m objectives that are the outcome of an expensive simulator or algorithm, $x \in \mathbb{R}^d$ consists of the decision variables, and $X \subseteq \mathbb{R}^d$ consists of the search space for the decision variables, including any constraints.

The expensive part of the optimisation Equation (1), which is usually f itself, is approximated with a machine learning model g, called the surrogate model. This is done using the outcomes obtained so far:

$$\min_g \; \sum_{i=1}^{n} L(g(x_i) - f(x_i)), \tag{2}$$

where L is a loss function such as the mean squared error or negative log-likelihood, and n is the current iteration of the SBO algorithm. Common surrogate models are Gaussian Processes [15] and random forests [19], among others. The surrogate model g is used to provide a candidate solution by finding the maximum of a so-called acquisition function α:

$$\operatorname*{argmax}_{x \in X} \; \alpha(g(x)). \tag{3}$$

This problem is much easier to solve than the original Equation (1) due to g having a closed form that is easy to evaluate; therefore, traditional optimisation methods such as derivative-based methods can be used. The acquisition function α is used to balance the trade-off between exploration and exploitation. Example acquisition functions are Expected Improvement, Upper Confidence Bound, Thompson sampling, and Entropy Search [15,20]. While the details of SBO algorithms can differ, they all contain a learning part as in (2) and an optimisation part as in (3).

3. Sustainability and Surrogate-Based Optimisation

This section proposes three definitions concerning the intersection of sustainability and SBO. Together, they are called Sustainable SBO (SSBO). Following the terminology

of [2], where the distinction is made between AI for Sustainability, i.e., using AI as a tool to achieve sustainability, and Sustainability of AI, i.e., taking carbon footprints and energy consumption into account when developing AI methods, one can define similar notions for SSBO:

Definition 2. *SBO for sustainability is concerned with applying SBO to sustainable applications, for example those that work towards the United Nations SDGs.*

Definition 3. *Sustainability of SBO is concerned with making sure the SBO algorithm itself is sustainable, e.g., does not significantly contribute to greenhouse gas emissions, has low energy consumption, is transparent about its computation costs, etc. This holds for both the ML part and the optimisation part of SBO.*

However, unlike in AI or ML in general, SBO is concerned with another aspect related to sustainability. As SBO is often used to prevent the prohibitive costs of running computationally expensive simulators multiple times, these 'savings' can be considered part of SSBO as well. The following definition is used in this work, where the name Sustainability with SBO is chosen to stay in line with the existing terminology:

Definition 4. *Sustainability with SBO is concerned with the prevention of running computationally expensive software, such as simulators or algorithms, more times than necessary.*

'More times than necessary' is ill-defined here, but a comparison can be made with any method that would be used if SBO algorithms did not exist, for example randomly searching for good outcomes of a simulator or algorithm, or applying other black-box optimisation techniques that do not make use of ML surrogates, e.g., metaheuristic algorithms.

It is this last aspect of SBO that sets it apart from other AI techniques, as the main goal of SBO is to reduce the number of expensive function evaluations for some objective function. In fact, Definition 4 is the actual purpose for which SBO has been developed in the first place, starting with algorithms such as EGO for expensive black-box optimisation [14]. At the same time, such types of 'energy savings' that are the result of using less expensive function evaluations, must be considered carefully when compared to the other SSBO definitions, so as to prevent falling into the trap of Jevons' paradox [21]. This paradox, when translated to the case of SSBO, could counter-intuitively result in using SBO with the same or an even higher number of function evaluations than any other algorithm because it is so efficient, which results in no savings of computational resources.

All in all, the three SSBO definitions must be carefully weighed against each other, similar to the weighing of sustainable AI notions according to Wynsberghe [2]: "to assess whether training or tuning of an AI model for a particular task is proportional to the carbon footprint, and general environmental impact, of that training and/or tuning". While such a 'proportionality framework' [2] is beyond the scope of this paper, Definition 2 is the main focus of this research, though the other SSBO definitions get some attention as well.

4. Survey Method

In this literature review, the search terms 'sustainable' and 'surrogate' were used. By far, this should not result in an extensive list, as both words have many synonyms (and also multiple meanings), and listing all synonyms would not only result in a number of studies too large to analyse, but would also be highly subjective. However, using these two words should be enough to serve the main goal of this survey: identifying studies that apply SBO to sustainable applications. At the same time, the terms are broad enough to cover a wide range of applications.

Only one database was used to retrieve records: SCOPUS (https://www.scopus.com, accessed on 6 February 2022). Furthermore, the time frame was limited to studies published in the 5-year period of 2017–2021. This was done to limit the number of studies to analyse in this exploratory survey, while still focusing on a time period where the topics

of sustainability and AI both gathered significant attention, and to make the search easily reproducible.

The following search term was used on SCOPUS:

```
TITLE-ABS-KEY ( "surrogate" AND "sustainable" ) AND PUBYEAR > 2016
AND PUBYEAR < 2022
```

This resulted in 329 records. See Figure 2 for an overview of the methodology. Upon closer inspection, many of these records were not relevant for this review. For example, the word 'surrogate' can have different meanings in fields such as chemistry, healthcare and biology (e.g., surrogate mother). Even in optimisation, 'surrogate' could have a different meaning than intended for this survey, such as a surrogate measure that is defined by expert knowledge rather than learned from data (e.g., [22]). Such records were removed by reading the title and abstract of all 329 records and checking whether the word 'surrogate' was used in the context of surrogate models using machine learning. In case of doubt, the record was not removed. At this point, no attention was given to the sustainability aspect. Screening the records this way resulted in 89 reports, of which 2 were automatically detected as duplicates using Rayyan [23], and 5 turned out to not be accessible with the academic licenses of Eindhoven University of Technology. These reports were removed.

All 82 remaining reports were read in full, and were included in this review if they satisfied the following eligibility criteria:

- It can be seen how sustainability is a topic of the report;
- The report uses a surrogate model based on machine learning;
- The surrogate model is used inside an optimisation framework.

These criteria are specified further in the remainder of this section.

4.1. Sustainability

As this survey is performed by one person and the first criterion is especially subjective, in order to avoid bias, only in the rarest of circumstances was this criterion used to exclude a report. For two studies, no link to sustainability was found: in [24], the word 'sustainability' appeared in the affiliation, keywords, and acknowledgements, but this word or similar words appeared nowhere in the title, abstract or main text, and in [25], the same word was only found in the copyright text and once in the main text without explanation. However, it is possible that the link to sustainability has gone unnoticed in these two papers, also due to the wide range of applications covered in this survey.

4.2. Machine Learning

Some reports did not use surrogate models based on machine learning, just like when screening the abstracts, but more information from the main paper was required to notice this (e.g., [26]). This only occurred six times, as in most other cases it was straightforward to notice the lack of machine learning from screening the abstract. Note that adding the words 'machine learning' to the search query would lead to different problems, as papers were found that would satisfy such an extended query but still do not use surrogate models based on machine learning (e.g., [27]).

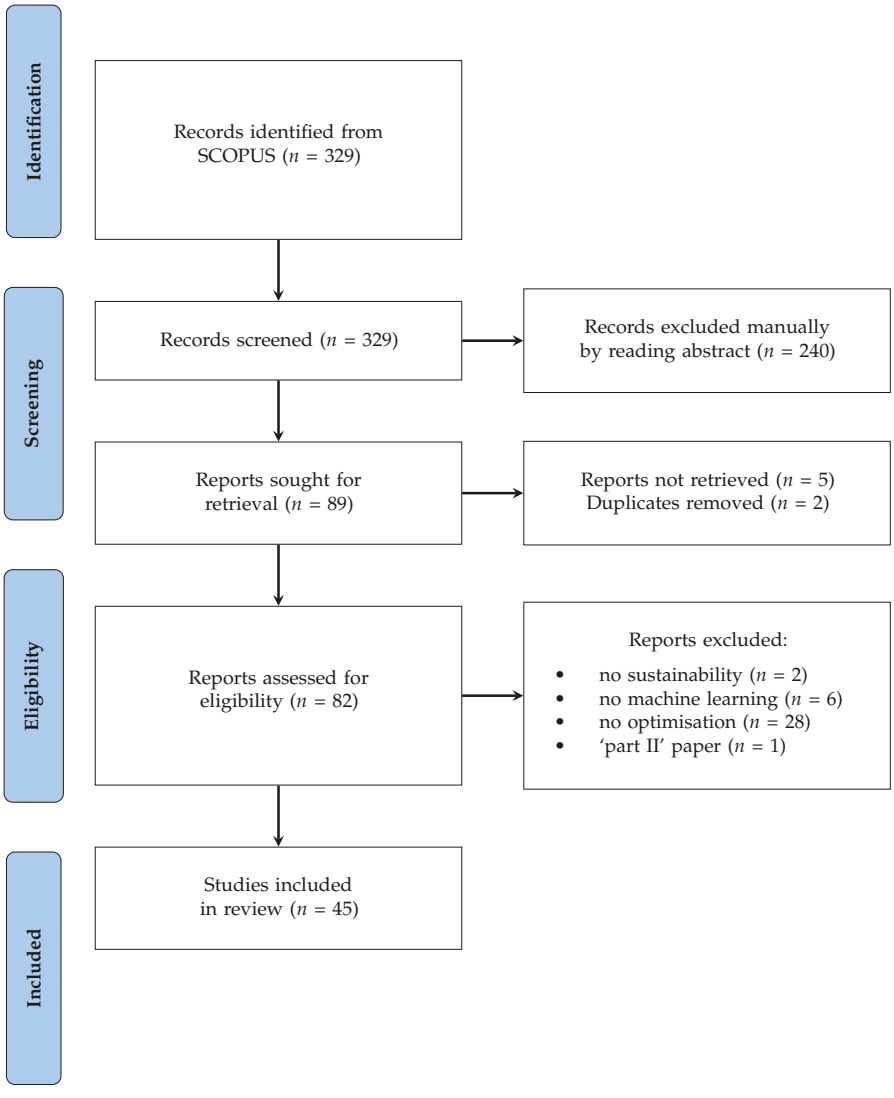


Figure 2. Literature survey flowchart.

4.3. Optimisation

Most of the studies that were excluded at this stage were not related to optimisation, even if they used surrogate models for other purposes such as prediction (e.g., [28]). In some of these studies, optimisation is only mentioned as a direction for future work (e.g., [29]). Note that adding a search query such as 'optimisation' would not solve the issue, as studies such as the earlier mentioned example [22] would be covered in the search query while not making use of SBO. Overall, 28 reports were excluded for not satisfying this criterion.

4.4. Other Criteria

Finally, two reports were a part I and part II of the same study [30,31]. These were both included but were counted together as one study. In the end, 45 studies were included in this review.

5. Surrogate-Based Optimisation for Sustainable Applications

This section analyses the studies that were included in this literature review, particularly how the studies made use of SBO for sustainable applications. Several properties of the included studies are shown in Table 1. These are: reference to the study, year, SBO framework, surrogate model used, the application related to sustainability that is addressed with SBO, domain or subject area, whether sustainability aspects of SBO are addressed, and any open questions related to SBO rather than the application. These properties are analysed in this section.

5.1. Year

Due to the search terms used in this survey, all years are covered in full: the year 2022 is not over at the time of writing this paper, but this year was not included in the survey. The number of studies included in this survey increased from 2 in 2017, to 4 in 2018, 8 in 2019, 14 in 2020, and finally 17 in 2021. Even though there is room for other search queries than the one used in this work, and therefore quantitative results could be subject to bias, this does give the indication that interest in applying SBO to sustainable applications has increased over the surveyed time period. A potential explanation for this is that both sustainability and AI were popular topics in this time period, not only in public but also in private sectors. Looking at AI investments for example: "From 2015 to 2020, the total yearly corporate global investment in AI increased by 55 billion U.S. dollars" [32]. It is likely that research on SBO, as a subset of AI, has benefited from this popularity. At the same time, the surveyed period closely follows the adoption of the Paris Agreement [33] and the United Nations SDGs [5], which have likely had a significant impact on the research focus of the time period under consideration.

5.2. Framework

While SBO methods such as Bayesian optimisation typically use an iterative approach, where surrogate models are constantly updated and used to search for better candidate solutions, this was not the only framework in which surrogate models found in this review were used. Non-iterative or direct approaches were more common, probably due to not including terms like 'Bayesian optimisation' or 'sequential model-based optimisation' in the search query. This work divides the included studies in five frameworks: Sequential Model-Based Optimisation, Predict-then-Optimise, Optimise-then-Predict, Predict-then-Interact, Bi-level Optimisation, Automated Machine Learning, and finally 'review' for review papers. Note that these frameworks are by no means extensive, and that they may be subject to bias and sometimes overlap, as is common when trying to divide optimisation algorithms into categories.

5.2.1. Sequential Model-Based Optimisation (SMBO)

This iterative procedure is the one used by Bayesian optimisation and similar algorithms [15,34]. Starting from an initial set of function evaluations, the surrogate model is learned. Then, an iterative procedure starts where (1) the surrogate model is used to suggest a new candidate point, (2) the expensive objective is evaluated at the new candidate point, and (3) the new evaluation is used to update the surrogate model. These three steps repeat until a stopping criterion such as maximum number of objective evaluations is satisfied. This iterative procedure is shown in Figure 1. This framework ensures that only promising parts of the search space are approximated by the surrogate model, which can reduce the required number of expensive function evaluations. An example of an included study that uses this framework in a water management problem is [35], where the surrogate model is repeatedly used in an optimisation problem solved by a genetic algorithm, the optimal points are given to a high-fidelity hydrodynamics simulator, and the surrogate is then updated with the outcome of the simulator.

5.2.2. Predict-then-Optimise (PtO)

This procedure is also called a direct procedure [36]. The surrogate model is learned from sampled data of the expensive objective, using e.g., Latin hypercube sampling, and then the optimisation problem is solved once with the new surrogate model. See Figure 3. It is possible that an iterative procedure is used for the learning process, however, no optimisation problem is solved within the iterative procedure: accuracy of the surrogate is the only goal. In some fields like chemical engineering, SMBO is known to outperform PtO [37], but in other fields like building design, it is still an open question which approach is better [36]. An example of an included study that uses the PtO framework in sustainable food production is given in [38], where the crop water demand is predicted with a neural network surrogate model, and the resulting optimisation problem is solved with nonlinear optimisation.

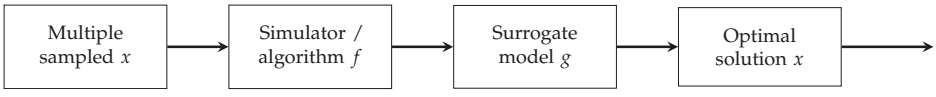


Figure 3. Predict-then-Optimise (PtO) framework.

5.2.3. Optimise-then-Predict (OtP)

Though it is arguable whether this framework is considered SBO, it was included as it satisfies the definition used in this work. Using a similar terminology as for PtO, in OtP, first, an optimisation problem is solved using standard optimisation algorithms, for multiple situations or contexts (e.g., wind conditions). Then, a surrogate model is trained on the resulting data. The model is then used to generalise or visualise the outcomes of the optimisation problem for new situations, i.e., prediction of optimal outputs is the final goal. See Figure 4. An example of an included study that uses this framework when planning charging stations for electric vehicles is found in [39], where a mathematical program is solved multiple times for different situations such as number of electric vehicles or energy storage capacity, and the surrogate model then predicts the optimal annual profit of the system for all possible situations.

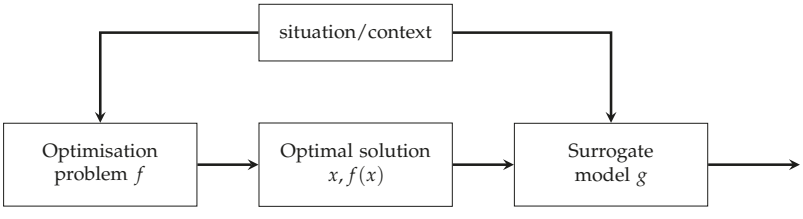


Figure 4. Optimise-then-Predict (OtP) framework.

5.2.4. Predict-then-Interact (PtI)

Keeping the terminology consistent, PtI is similar to PtO, but instead of using an algorithm to solve the optimisation problem with the trained surrogate, a human interacts with the surrogate. See Figure 5. If used for design, the designer can use the surrogate to manually solve an optimisation problem defined by their own constraints and objectives, which are often related to creativity and cannot always be defined mathematically. Two included studies use this framework, both for building design [40,41].

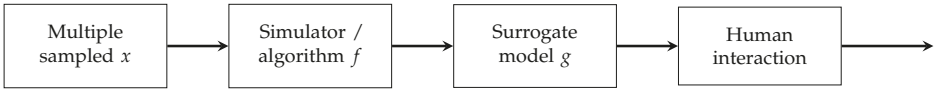


Figure 5. Predict-then-Interact (PtI) framework.

5.2.5. Bi-level Optimisation (BlO)

In this framework, there are two nested optimisation problems, where the outer optimisation problem depends on the results of the inner optimisation problem. The surrogate model is used to approximate the inner optimisation problem. The framework is the same as in Figure 1, but with f consisting of two nested problems. An example of an included study that uses this framework for sustainable land development is [42], where the problem is formulated as a Stackelberg game between government (who wants to balance land use between food, energy and water) and land developers (who want to maximise profit), and a surrogate model approximates the decisions of the land developers.

5.2.6. Automated Machine Learning (AutoML)

In this framework, the final task is similar as in machine learning (ML), e.g., prediction or classification, but SBO is used (typically in the SMBO framework) to automatically solve part of the ML procedure that is usually done by hand, such as choosing which ML model to use or tuning the hyperparameters. The framework is again the same as in Figure 1, but with f consisting of the ML procedure. The only included study that uses this framework is [43], which uses SBO to select the models and hyperparameters of a ML model that predicts groundwater levels. Note that AutoML is also considered a sub-field of AI that does not necessarily always make use of SBO, but here it is considered an SBO framework to separate it from the other frameworks.

5.2.7. Framework Discussion

In the included studies, not all frameworks had the same frequency of appearance. As mentioned, the PtO framework was quite common. Different optimisers were used in the optimisation step, such as gradient descent (e.g., [44]), NSGA-II (e.g., [45]), tabu search (e.g., [46]), CMA-ES (e.g., [31]), and more. The SMBO framework was most common in studies from the chemical engineering and computer science domains, among others. It is possible that this framework is not yet well known in other domains, despite its earlier mentioned advantages. Examples of optimisers used in the SMBO framework are NSGA-II [47,48], simulated annealing [35], multi-objective particle swarm optimisation [49], and more. Bi-level optimisation problems were often the result of having multiple stakeholders—which is common in sustainable applications, such as road users and government [50], or land owners and government [42]. OtP problems often appeared when outside factors such as traffic load or weather came into play [39,51]. Two studies used the PtI framework, both for building design [40,41]. There were also some comprehensive reviews for SBO in building design [36,52] and other specific applications, but none for sustainable applications in general. Finally, the AutoML framework appeared only once [43], though replacing 'surrogate' with 'AutoML' or similar terms in the search query would likely yield more results.

5.3. Surrogate Model

The type of surrogate model used in the included studies varied a lot, mostly depending on the framework. Artificial Neural Networks (ANNs) are very popular in PtO, while Gaussian Processes (GPs) or Radial Basis Functions (RBFs) are more common in other frameworks such as SMBO or BlO. As noted in one of the included studies [52], the popularity of ANNs can likely be explained by the success of deep learning in the last decade and by researchers having access to more powerful hardware that allows more complex models. Other surrogate models were Multivariate Adaptive Regression Spline (MARS), Support Vector Machine (SVM) or Support Vector Regression (SVR), linear and polynomial regression, piece-wise linear models, Random Forest (RF), Recurrent Neural Network (RNN), and ensembles of multiple models. A general explanation of how to use ML models such as ANNs or GPs in SBO can be found in, e.g., [18].

5.4. Application and Domain

The goal of this study is to identify sustainable applications where SBO is applied, both to point SBO researchers to new interesting and relevant problems, and to make researchers in these sustainable application domains aware of the power of SBO. The range of applications is quite broad, from groundwater management to electric vehicles. The terms in the 'application' column of Table 1 were chosen manually after reading the studies, and might be subject to bias, especially considering the broad range of topics. Therefore, the domains of the application are also mentioned: these are retrieved from SCOPUS directly. A broad list of domains is covered, but engineering, environmental science, energy and computer science were the most common domains, covering over half of all included studies as seen in Figure 6. While the prominence of energy and environmental science is to be expected when searching for sustainability-related studies, the engineering domain is likely well-represented due to SBO being considered a subset of engineering optimisation [18]. Similarly, SBO is considered part of AI, which can be considered part of the computer science domain. Overall, due to the diversity of the sustainable applications that were found, this study has already achieved its main goal.

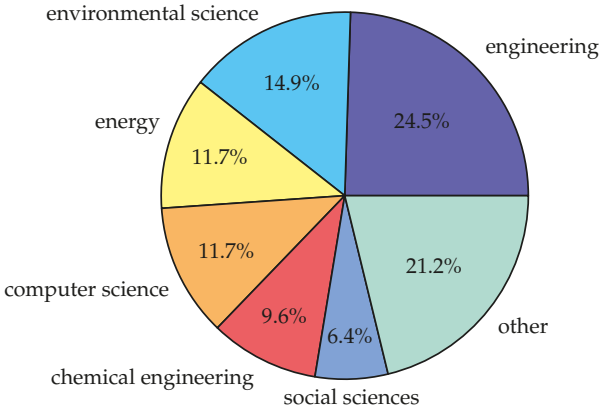


Figure 6. Domains of the studies encountered in this literature survey.

5.5. Sustainable SBO

As explained in Section 3, this work considers three types of Sustainable SBO (SSBO). SBO for Sustainability is about the sustainability aspects of the application itself. This work assumes that all the included studies cover this aspect in some way or another: e.g., [53] uses SBO to increase soil health by 7.6%, while [54] finds several new stable chemical compounds for sustainable energy applications using SBO. This assumption is made in order to reduce bias, and because of the wide range of applications which requires expertise in many different science domains to fully understand the sustainability aspects of the applications. The corresponding column in Table 1 ignores this aspect of SSBO and only reveals whether other SSBO aspects were covered. This was done by manually inspecting the studies and is therefore still subject to bias, so only examples and general insights are given here.

Sustainability with SBO is about the prevention of running computationally intensive simulators or algorithms. Since this is the main reason for using SBO, it is assumed that all included studies take this aspect into account when choosing their methods. However, most studies do not quantify this aspect, which makes it difficult to determine whether the benefits (SBO for Sustainability and Sustainability with SBO) are worth the computational resources of SBO itself (Sustainability of SBO). An example of an included study that does quantify this is [55], where the total time of the SBO approach was estimated at 15 h for evaluating the expensive simulator and 1 000 s for the ML and optimi-

sation parts of the SBO approach, while directly optimising the expensive simulator using the same optimisation procedure without a surrogate model was estimated to take 330 h. In other words, SBO has lead to approximately a 95% reduction in computational resources for this application, when compared to other optimisation techniques that do not make use of ML. Similar savings in computation time were reported for sustainable building design in one of the included reviews [36], with 97% as the largest reported number. Studies that quantify this aspect of SSBO get a checkmark in the corresponding column in Table 1.

Finally, studies that discuss Sustainability of SBO itself, for example by mentioning the trade-off in computational resources between expensive optimisation problem and SBO framework, or by quantifying the computation time or energy usage of their SBO framework, get a checkmark in the corresponding column. An example is [44], where computation times for training the ANN surrogate, using the surrogate for optimisation, and evaluating the expensive simulator are all reported. The study estimates that the surrogate model is 4000 times faster than the expensive simulator after performing the 'predict' step of the PtO framework. If the entire PtO framework is taken into account, the study estimates that the SBO framework is more efficient than a regular optimisation framework (in this case the gradient descent or Nelder–Mead simplex algorithm) in the situation that the expensive simulator is called more than 168,000 times. The authors conclude that "the use of ANN in multidisciplinary optimization frameworks transfers the computational cost of the aircraft optimization task to the ANN training process".

It should be noted that the above study was an exception: almost none of the studies included in this work mentioned this last aspect of SSBO, i.e., the computational resources of the SBO framework. What was usually mentioned is the number of samples used to train the surrogate model, but not the time used to train or optimise the surrogate. In some studies, computation times of the SBO framework were reported in the supplementary material, e.g., [30]. In others, such as [56], the ML and optimisation parts of the SBO framework were considered to be negligible, so only the total computation time of using the SBO framework and evaluating the expensive simulator together were mentioned. While it is indeed the case for many applications that the SBO framework has negligible computation time compared to the expensive simulator, this is certainly not the case for all applications, as seen with the study [44] earlier. Therefore, it is important to keep track of both the computation time of the SBO framework itself and the computation time spent on calling the expensive simulator, and to separate these.

5.6. Open Questions

Most studies posed several open questions related to their application; however, many studies also contained more general open questions related to SBO. The latter are denoted in this column, in order to stimulate SBO researchers to tackle these open questions. These questions were found manually, mainly by inspecting the conclusions and future work sections of the papers, and are by no means an exhaustive list of open problems in SBO, even for the applications in this review. Still, many commonalities were found.

One of the most common open questions is related to the high dimensionality present in many applications. Many SBO algorithms struggle when a large number of variables is present, e.g., over 200 variables, which can occur in applications such as building design [31]. Fortunately, high dimensionality is an active area of research in SBO [57,58], though it remains the question which approach works best in practice and how to solve the problem efficiently.

Another common open question was that of dealing with multiple objectives, or specifically, many objectives, as most problems in this review already had to deal with two or three objectives and some of them mentioned adding more in the future work section. For example, Ref. [47] uses two objectives but mentions at least five more in the future work section. Multi-objective problems were common in this review due to the nature of sustainable applications: often, ecological, economical, and social aspects of the same problem had to be considered. This was also noticed in a related review on metaheuristic

optimisation algorithms for sustainability-related applications, where multi-objective problems outnumbered single-objective problems at least four to one [13]. Furthermore, this could be another reason for the popularity of the PtO framework compared to SMBO, as not all SMBO techniques are equipped for dealing with multi-objective problems. How to deal with a large number of objectives (e.g., more than 1–3) in SBO remains an open question.

The question of robustness and generalisation was also often encountered. Just because a surrogate model is accurate in one situation, does not mean it is accurate in similar situations, such as different weather conditions when designing heat pump systems [16]. In the ML community this question has gained a lot of attention in the past, but in SBO, generalisation aspects are less well understood. Especially in the SMBO framework, where the surrogate only approximates a small part of the search space, generalisation typically plays a smaller role than in traditional ML or in the PtO framework. Research on contextual bandits [59] and similar ideas could be of use here.

Other open questions are concerned with: making efficient use of parallelisation, having multiple users in the PtI framework, including historic data in the SMBO framework, hyperparameter optimisation for surrogate learning, using smooth and/or interpretable surrogates, and general challenges in optimisation such as mixed variables, multimodality, nonlinearity and nonconvexity. Though many of these questions are gaining attention in recent SBO research, such as mixed variables [60,61], these open questions can serve as an incentive for SBO researchers to tackle such problems.

Table 1. Properties of the surveyed literature. SSBO = Sustainable Surrogate-Based Optimisation, SMBO = Sequential Model-Based Optimisation, PtO = Predict then Optimise, OtP = Optimise then Predict, PtI = Predict then Interact, BlO = Bi-level Optimisation, AutoML = Automated Machine Learning.

Study	Year	Framework	Surrogate	Application	Domain	SSBO	Open Questions
[62]	2017	PtO	MARS	groundwater extraction	engineering	✓	parallelisation
[44]	2017	PtO	ANN	aviation	engineering	✓	dimension reduction
[38]	2018	PtO	ANN	food production	chemical engineering; computer science	-	assumptions
[42]	2018	BlO	unknown	land development	chemical engineering; computer science	-	-
[55]	2018	SMBO	ANN	production systems	chemical engineering; chemistry	✓	sustainability in objective
[63]	2018	PtO	SVR	groundwater extraction	environmental science; earth and planetary sciences	-	hyperparameter optimization
[64]	2019	PtO	SVR	groundwater extraction	environmental science; social sciences	-	-
[39]	2019	OtP	polynomial, RBF, GP	electric vehicles	energy; engineering	-	high dimensionality
[47]	2019	PtO	polynomial	outdoor thermal comfort	social sciences; engineering	-	many objectives
[65]	2019	PtO	GP	thermal comfort	environmental science	-	-
[36]	2019	review	multiple	building design	engineering	✓	high dimensionality; smoothness; efficiency; interpretability

Table 1. *Cont.*

Study	Year	Framework	Surrogate	Application	Domain	SSBO	Open Questions
[66]	2019	PtO	SVR	groundwater management	environmental science	-	multiple objectives
[17]	2019	SMBO	RBF	drug manufacturing	chemical engineering; chemistry; energy; environmental science	-	multiple objectives
[51]	2019	OtP	ANN	building renovation	engineering	✓	generalisation; efficient sampling
[67]	2020	BlO	ANN	water management	environmental science	-	multiple and fuzzy objectives
[48]	2020	PtO	ANN	solar heat system	energy; engineering; environmental science; business, management and accounting	-	-
[52]	2020	review	multiple	building design	engineering	-	incorporate behavioural data; reproducibility
[40]	2020	PtI	RF; ensemble ANN	building design	engineering	-	multiple objectives; multiple users
[41]	2020	PtI; PtO	RF	building design	social sciences; computer science; arts and humanities	✓	multiple users
[68]	2020	SMBO	GP	sea transport	engineering; computer science	-	parallelisation
[35]	2020	SMBO	ANN	water management	environmental science; engineering	-	many objectives
[69]	2020	OtP	linear	cooling tower	engineering; environmental science	-	-
[70]	2020	PtO	ANN	building energy management	engineering	-	efficiency
[45]	2020	PtO	ANN	air conditioning	energy	-	transfer learning
[54]	2020	SMBO	GP	material discovery	chemical engineering; materials science	-	include historic data
[71]	2020	SMBO	polynomial	public transport	mathematics; computer science	-	complex variable interactions; visualisation
[72]	2020	review	multiple	hydro-cracking	energy; environmental science; social sciences	✓	multidisciplinarity
[73]	2020	BlO	RBF; GP; polynomial	transportation networks	engineering; social sciences; decision sciences	-	hyperparameter optimisation; model selection
[74]	2021	PtO	linear; SVM	product design	engineering; chemical engineering	-	robustness
[46]	2021	PtO	multiple	urban logistics	mathematics; computer science	✓	robustness

Table 1. *Cont.*

Study	Year	Framework	Surrogate	Application	Domain	SSBO	Open Questions
[75]	2021	review; PtO	multiple	process design; material design	chemical engineering; computer science; energy; engineering; environmental science; materials science	✓	high dimensionality; generalisation
[30,31]	2021	PtO	ANN	building design	energy; materials science	✓	high dimensionality and constraints
[53]	2021	SMBO	RBF	soil health	agriculture and biological sciences; computer science	✓	variable reduction
[49]	2021	SMBO	ANN	hydropower reservoir	engineering; computer science	-	high dimensionality; parallelisation
[50]	2021	BlO	RBF	electric vehicles	business, management and accounting; engineering; social sciences	✓	mixed variables and constraints
[56]	2021	SMBO	GP ensemble	chemical process	business, management and accounting; energy; engineering; environmental science	✓	high dimensionality; multimodality
[43]	2021	AutoML	RBF; GP	groundwater management	computer science; decision sciences; mathematics	✓	generalisation
[76]	2021	PtO	ANN	urban drainage systems	environmental science	✓	divide problem into subproblems
[77]	2021	PtO	RNN	bridge maintenance	business, management and accounting; engineering	-	-
[78]	2021	PtO	ANN	chemical process	chemical engineering; chemistry; engineering	-	robustness
[79]	2021	PtO	RBF ensemble	concrete barriers	computer science; engineering; mathematics	-	-
[80]	2021	PtO	ANN	water management	energy	-	multiple objectives; accuracy
[16]	2021	SMBO	RBF	heat pump system	energy	-	multiple objectives; constraints; robustness; discrete variables
[81]	2021	PtO	ANN	thermal comfort	engineering	✓	generalisation
[82]	2021	PtO	piece-wise linear	agricultural system	energy; chemistry; chemical engineering; environmental science	✓	high dimensionality; nonlinearity; nonconvexity

6. Discussion and Conclusions

Answering the main research question, "How is SBO applied to sustainability-related applications?", several sustainability-related applications from a wide variety of research domains were identified, as well as several frameworks in which SBO were used. Some

of these frameworks (such as SMBO and BlO) used an iterative procedure where the surrogate is continuously updated, while others (such as PtO and OtP) trained the surrogate only once. The AutoML framework, where SBO is applied to a more general ML problem, opens up the path of applying SBO indirectly to many more sustainable applications that make use of ML. Overall, besides the sustainability aspect of the application (SBO for Sustainability), many researchers applied SBO to prevent having to run expensive simulators or algorithms multiple times, which is another sustainability aspect related to SBO denoted Sustainability with SBO in this work. However, researchers were not always transparent about the computational resources spent on the ML and optimisation parts of SBO itself: Sustainability of SBO was often not considered. In some cases, these computational resources were negligible compared to those spent on the expensive simulators and algorithms, but in other cases they were not. This makes it difficult to analyse the trade-off between different sustainability aspects of SBO, for example whether to use a more complex but time-consuming ML model in order to save a few more calls to the expensive simulator or algorithm, or to spend more computational resources to increase the sustainability aspects of the application itself (e.g., reduce CO_2 emissions). Therefore, the following recommendations are made to application researchers to improve Sustainable SBO:

- Report the hardware used for the SBO framework and the hardware used for the expensive simulator or algorithm (these are often the same).
- Report the number of calls to the expensive simulator or algorithm and the computation time used for this.
- Report the computation time used for training the surrogate model, including model selection and hyperparameter optimisation.
- Report the computation time used for the optimisation part of SBO. If this cannot be separated, this can be merged with the point above.
- Estimate the time it would take to solve the optimisation problem without a surrogate model, if the expensive simulator or algorithm was optimised directly.
- Consider using an iterative framework such as SMBO instead of PtO to potentially reduce the number of calls to the expensive simulator or algorithm.

Besides these SBO-specific recommendations, the following recommendation is related to sustainability of AI in general:

- If possible, report not just the computation times, but also the energy consumption (and energy mix used) or even CO_2 emissions used for the computations.

While this last point is not necessary for making the trade-off between Sustainability of SBO and Sustainability with SBO, i.e., deciding whether using SBO is more efficient than using other optimisation algorithms, it is necessary for also including SBO for Sustainability in the trade-off. To give a concrete example: if 100 tonnes of CO_2 can be mitigated by designing a sustainable solar heat system in Mexico, as is done in one of the included studies [48], but training the ML model would emit over 200 tonnes of CO_2, as estimated can happen for complex ML tasks such as natural language processing [11], the decision whether to apply SBO to this application is not that easy to make. Fortunately, the carbon footprint of the ML models used in SBO are likely not that high, but without any transparency on this issue it will be difficult to prevent SBO from heading the same direction as natural language processing, with a large carbon footprint as a result. This same call for transparency on the environmental impact of AI techniques is found in other related subfields of AI, such as autoML [12]. Examples of tools to measure the CO_2 emissions of ML are Carbontracker [83] and Machine Learning Emissions Calculator [84], and a similar example for algorithms in general is Green Algorithms [85].

For SBO researchers, the same recommendations above are made, but the reviewed studies themselves also contained recommendations in the form of open questions. These can be used to determine which challenges in SBO to tackle next. The most common open questions found in the reviewed studies were related to high dimensionality, and multi-objective optimisation. This is in line with the results of a questionnaire on more

general real-world optimisation problems [86], i.e., problems not necessarily related to SBO or sustainability, where having two or more objectives and tens or hundreds of variables was quite common. Especially researchers using the SMBO framework should take this into account when designing their algorithms. Note that in that same questionnaire, objectives that took up to an hour to evaluate were quite common, which indicates the importance of SBO for real-world optimisation problems, and the potential of Sustainability with SBO.

All in all, it can be concluded that there is great potential in responsibly using SBO to make the world more sustainable.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The author would like to thank Shane Ó Seasnáin for his valuable feedback on a first draft of this paper.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ML	Machine Learning
SDG	Sustainable Development Goal
SBO	Surrogate-Based Optimisation
SSBO	Sustainable Surrogate-Based Optimisation
SMBO	Sequential Model-Based Optimisation
PtO	Predict-then-Optimise
OtP	Optimise-then-Predict
PtI	Predict-then-Interact
BlO	Bi-level Optimisation
AutoML	Automated Machine Learning
ANN	Artificial Neural Network
GP	Gaussian Process
RBF	Radial Basis Function
MARS	Multivariate Adaptive Regression Spline
SVM	Support Vector Machine
SVR	Support Vector Regression
RF	Random Forest
RNN	Recurrent Neural Network

References

1. Vinuesa, R.; Azizpour, H.; Leite, I.; Balaam, M.; Dignum, V.; Domisch, S.; Felländer, A.; Langhans, S.D.; Tegmark, M.; Nerini, F.F. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nat. Commun.* **2020**, *11*, 1–10. [CrossRef] [PubMed]
2. Wynsberghe, A.R.V. Sustainable AI: AI for sustainability and the sustainability of AI. *AI Ethics* **2021**, *1*, 213–218. [CrossRef]
3. Climate Change AI. Available online: https://www.climatechange.ai/ (accessed on 17 February 2021).
4. AI for Good. Available online: https://ai4good.org/ (accessed on 17 February 2021).
5. Sustainable Development Goals. Available online: https://sdgs.un.org/goals (accessed on 17 February 2021).
6. Rolnick, D.; Donti, P.L.; Kaack, L.H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A.S.; Milojevic-Dupont, N.; Jaques, N.; Waldman-Brown, A.; et al. Tackling Climate Change with Machine Learning. *ACM Comput. Surv.* **2022**, *55*, 1–96. [CrossRef]
7. Wang, J.; Yang, Y.; Wang, T.; Sherratt, R.S.; Zhang, J. Big data service architecture: A survey. *J. Internet Technol.* **2020**, *21*, 393–405.
8. Wang, W.; Yang, Y.; Li, J.; Hu, Y.; Luo, Y.; Wang, X. Woodland labeling in Chenzhou, China, via deep learning approach. *Int. J. Comput. Intell. Syst.* **2020**, *13*, 1393. [CrossRef]

9. Maheswari, R.U.; Umamaheswari, R. Wind Turbine Drivetrain Expert Fault Detection System: Multivariate Empirical Mode Decomposition based Multi-sensor Fusion with Bayesian Learning Classification. *Intell. Autom. Soft Comput.* **2020**, *26*, 479–488. [CrossRef]
10. Schwartz, R.; Dodge, J.; Smith, N.; Etzioni, O. Green AI. *Commun. ACM* **2020**, *63*, 54–63. [CrossRef]
11. Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Florence, Italy, 2019; pp. 3645–3650. [CrossRef]
12. Tornede, T.; Tornede, A.; Hanselle, J.; Wever, M.; Mohr, F.; Hullermeier, E. Towards Green Automated Machine Learning: Status Quo and Future Directions. *arXiv* **2021**, arXiv:abs/2111.05850.
13. Sadollah, A.; Nasir, M.; Geem, Z.W. Sustainability and Optimization: From Conceptual Fundamentals to Applications. *Sustainability* **2020**, *12*, 2027. [CrossRef]
14. Jones, D.R.; Schonlau, M.; Welch, W.J. Efficient Global Optimization of Expensive Black-Box Functions. *J. Glob. Optim.* **1998**, *13*, 455–492. [CrossRef]
15. Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R.P.; de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, *104*, 148–175. [CrossRef]
16. Vering, C.; Wüllhorst, F.; Mehrfeld, P.; Müller, D. Towards an integrated design of heat pump systems: Application of process intensification using two-stage optimization. *Energy Convers. Manag.* **2021**, *250*, 114888. [CrossRef]
17. Ho, C.H.; Yi, J.; Wang, X. Biocatalytic Continuous Manufacturing of Diabetes Drug: Plantwide Process Modeling, Optimization, and Environmental and Economic Analysis. *ACS Sustain. Chem. Eng.* **2019**, *7*, 1038–1051. [CrossRef]
18. Martins, J.R.; Ning, A. *Engineering Design Optimization*; Cambridge University Press: Cambridge, UK, 2021.
19. Lindauer, M.; Eggensperger, K.; Feurer, M.; Biedenkapp, A.; Deng, D.; Benjamins, C.; Ruhkopf, T.; Sass, R.; Hutter, F. SMAC3: A versatile Bayesian optimization package for hyperparameter optimization. *J. Mach. Learn. Res.* **2022**, *23*, 1–9.
20. Wang, Z.; Jegelka, S. *Max-Value Entropy Search for Efficient Bayesian Optimization*; ICML: Sydney, Australia, 2017; pp. 3627–3635.
21. Alcott, B. Jevons' paradox. *Ecol. Econ.* **2005**, *54*, 9–21. [CrossRef]
22. Sarkar, A.; Bardhan, R. A simulation based framework to optimize the interior design parameters for effective Indoor Environmental Quality (IEQ) experience in affordable residential units: Cases from Mumbai, India. *IOP Conf. Ser. Earth Environ. Sci.* **2019**, *294*, 012060. [CrossRef]
23. Ouzzani, M.; Hammady, H.; Fedorowicz, Z.; Elmagarmid, A. Rayyan—A web and mobile app for systematic reviews. *Syst. Rev.* **2016**, *5*, 210. [CrossRef]
24. Nahvi, A.; Sadoughi, M.K.; Arabzadeh, A.; Sassani, A.; Hu, C.; Ceylan, H.; Kim, S. Multi-objective Bayesian optimization of super hydrophobic coatings on asphalt concrete surfaces. *J. Comput. Des. Eng.* **2019**, *6*, 693–704. [CrossRef]
25. Nguyen, D.D.; Nguyen, L. An Adaptive Control for Surrogate Assisted Multi-objective Evolutionary Algorithms. *Adv. Intell. Syst. Comput.* **2021**, *1270*, 123–132.
26. Kazi, S.R.; Short, M.; Biegler, L.T. Synthesis of Combined Heat and Mass Exchange Networks Via a Trust Region Filter Optimisation Algorithm Including Detailed Unit Designs. *Comput. Aided Chem. Eng.* **2021**, *50*, 13–18.
27. Genedy, R.A.; Ogejo, J.A. Using machine learning techniques to predict liquid dairy manure temperature during storage. *Comput. Electron. Agric.* **2021**, *187*, 106234. [CrossRef]
28. Hoque, M.M.; Rahman, M.T.U. Landfill area estimation based on solid waste collection prediction using ANN model and final waste disposal options. *J. Clean. Prod.* **2020**, *256*, 120387. [CrossRef]
29. Giselle Fernandez-Godino, M.; Grosskopf, M.J.; Nakhleh, J.B.; Wilson, B.M.; Kline, J.L.; Srinivasan, G. Identifying Entangled Physics Relationships through Sparse Matrix Decomposition to Inform Plasma Fusion Design. *IEEE Trans. Plasma Sci.* **2021**, *49*, 2410–2419. [CrossRef]
30. Ekici, B.; Kazanasmaz, Z.T.; Turrin, M.; Taşgetiren, M.F.; Sariyildiz, I.S. Multi-zone optimisation of high-rise buildings using artificial intelligence for sustainable metropolises. Part 1: Background, methodology, setup, and machine learning results. *Sol. Energy* **2021**, *224*, 373–389. [CrossRef]
31. Ekici, B.; Kazanasmaz, Z.T.; Turrin, M.; Taşgetiren, M.F.; Sariyildiz, I.S. Multi-zone optimisation of high-rise buildings using artificial intelligence for sustainable metropolises. Part 2: Optimisation problems, algorithms, results, and method validation. *Sol. Energy* **2021**, *224*, 309–326. [CrossRef]
32. Artificial Intelligence (AI) Worldwide—Statistics & Facts. Available online: https://www.statista.com/topics/3104/artificial-intelligence-ai-worldwide (accessed on 18 February 2021).
33. The Paris Agreement. Available online: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement (accessed on 18 February 2021).
34. Hutter, F.; Hoos, H.H.; Leyton-Brown, K. Sequential Model-Based Optimization for General Algorithm Configuration. In *Learning and Intelligent Optimization*; Coello, C.A.C., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 507–523.
35. Saadatpour, M. An Adaptive Surrogate Assisted CE-QUAL-W2 Model Embedded in Hybrid NSGA-II AMOSA Algorithm for Reservoir Water Quality and Quantity Management. *Water Resour. Manag.* **2020**, *34*, 1437–1451. [CrossRef]
36. Westermann, P.; Evins, R. Surrogate modelling for sustainable building design—A review. *Energy Build.* **2019**, *198*, 170–186. [CrossRef]

37. Garud, S.S.; Karimi, I.A.; Kraft, M. Smart sampling algorithm for surrogate model development. *Comput. Chem. Eng.* **2017**, *96*, 103–114. [CrossRef]
38. Woldesellasse, H.; Govindan, R.; Al-Ansari, T. Role of analytics within the energy, water and food nexus—An Alfalfa case study. *Comput. Aided Chem. Eng.* **2018**, *44*, 997–1002.
39. Chen, Y.; Dababneh, F.; Zhang, B.; Kassaee, S.; Smith, B.T.; Liu, X.; Momen, A.M. Surrogate Modeling for Capacity Planning of Charging Station Equipped With Photovoltaic Panel and Hydropneumatic Energy Storage. *J. Energy Resour. Technol.* **2019**, *142*, 50907. [CrossRef]
40. Brown, N.C.; Jusiega, V.; Mueller, C.T. Implementing data-driven parametric building design with a flexible toolbox. *Autom. Constr.* **2020**, *118*, 103252. [CrossRef]
41. Brown, N.C. Design performance and designer preference in an interactive, data-driven conceptual building design scenario. *Des. Stud.* **2020**, *68*, 1–33. [CrossRef]
42. Avraamidou, S.; Beykal, B.; Pistikopoulos, I.P.; Pistikopoulos, E.N. A hierarchical Food-Energy-Water Nexus (FEW-N) decision-making approach for Land Use Optimization. *Comput. Aided Chem. Eng.* **2018**, *44*, 1885–1890.
43. Müller, J.; Park, J.; Sahu, R.; Varadharajan, C.; Arora, B.; Faybishenko, B.; Agarwal, D. Surrogate optimization of deep neural networks for groundwater predictions. *J. Glob. Optim.* **2021**, *81*, 203–231. [CrossRef]
44. Secco, N.R.; De Mattos, B.S. Artificial neural networks to predict aerodynamic coefficients of transport airplanes. *Aircr. Eng. Aerosp. Technol.* **2017**, *89*, 211–230. [CrossRef]
45. Saikia, P.; Gaurav.; Rakshit, D. Designing a clean and efficient air conditioner with AI intervention to optimize energy-exergy interplay. *Energy AI* **2020**, *2*, 100029. [CrossRef]
46. Yang, J.; Lau, H.C. A Learning and Optimization Framework for Collaborative Urban Delivery Problems with Alliances. *Lect. Notes Comput. Sci.* **2021**, *13004*, 316–331.
47. Du, Y.; Mak, C.M.; Li, Y. A multi-stage optimization of pedestrian level wind environment and thermal comfort with lift-up design in ideal urban canyons. *Sustain. Cities Soc.* **2019**, *46*, 101424. [CrossRef]
48. May Tzuc, O.; Bassam, A.; Ricalde, L.J.; Jaramillo, O.A.; Flota-Bañuelos, M.; Escalante Soberanis, M.A. Environmental-economic optimization for implementation of parabolic collectors in the industrial process heat generation: Case study of Mexico. *J. Clean. Prod.* **2020**, *242*, 118538. [CrossRef]
49. Saadatpour, M.; Javaheri, S.; Afshar, A.; Sandoval Solis, S. Optimization of selective withdrawal systems in hydropower reservoir considering water quality and quantity aspects. *Expert Syst. Appl.* **2021**, *184*, 115474. [CrossRef]
50. Liu, H.; Zou, Y.; Chen, Y.; Long, J. Optimal locations and electricity prices for dynamic wireless charging links of electric vehicles for sustainable transportation. *Transp. Res. Part Logist. Transp. Rev.* **2021**, *152*, 102187. [CrossRef]
51. Sharif, S.A.; Hammad, A. Developing surrogate ANN for selecting near-optimal building energy renovation methods considering energy consumption, LCC and LCA. *J. Build. Eng.* **2019**, *25*, 100790. [CrossRef]
52. Roman, N.D.; Bre, F.; Fachinotti, V.D.; Lamberts, R. Application and characterization of metamodels based on artificial neural networks for building performance simulation: A systematic review. *Energy Build.* **2020**, *217*, 109972. [CrossRef]
53. Ramos-Castillo, M.; Orvain, M.; Naves-Maschietto, G.; de Faria, A.B.B.; Chenu, D.; Albuquerque, M. Optimal agricultural spreading scheduling through surrogate-based optimization and MINLP models. *Inf. Process. Agric.* **2021**, *8*, 159–172. [CrossRef]
54. Flores, R.A.; Paolucci, C.; Winther, K.T.; Jain, A.; Torres, J.A.G.; Aykol, M.; Montoya, J.; Nørskov, J.K.; Bajdich, M.; Bligaard, T. Active Learning Accelerated Discovery of Stable Iridium Oxide Polymorphs for the Oxygen Evolution Reaction. *Chem. Mater.* **2020**, *32*, 5854–5863. [CrossRef]
55. González-Garay, A.; Guillén-Gosálbez, G. SUSCAPE: A framework for the optimal design of SUStainable ChemicAl ProcEsses incorporating data envelopment analysis. *Chem. Eng. Res. Des.* **2018**, *137*, 246–264. [CrossRef]
56. Dai, M.; Yang, F.; Zhang, Z.; Liu, G.; Feng, X. Energetic, economic and environmental (3E) multi-objective optimization of the back-end separation of ethylene plant based on adaptive surrogate model. *J. Clean. Prod.* **2021**, *310*, 127426. [CrossRef]
57. Oh, C.; Gavves, E.; Welling, M. BOCK: Bayesian optimization with cylindrical kernels. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 3868–3877.
58. Kirschner, J.; Mutný, M.; Hiller, N.; Ischebeck, R.; Krause, A. *Adaptive and Safe Bayesian Optimization in High Dimensions via One-Dimensional Subspaces*; PMLR: Long Beach, CA, USA, 2019; pp. 3429–3438.
59. Krause, A.; Ong, C. Contextual Gaussian process bandit optimization. *NIPS* **2011**, *24*, 2447–2455.
60. Bliek, L.; Verwer, S.; Weerdt, M.D. Black-box mixed-variable optimisation using a surrogate model that satisfies integer constraints. In Proceedings of the Genetic and Evolutionary Computation Conference Companion, Lille, France, 10–14 July 2021; pp. 1851–1859.
61. Daxberger, E.A.; Makarova, A.; Turchetta, M.; Krause, A. Mixed-Variable Bayesian Optimization. *arXiv* **2019**, arXiv:1907.01329.
62. Roy, D.K.; Datta, B. A surrogate based multi-objective management model to control saltwater intrusion in multi-layered coastal aquifer systems. *Civ. Eng. Environ. Syst.* **2017**, *34*, 238–263. [CrossRef]
63. Lal, A.; Datta, B. Modelling saltwater intrusion processes and development of a multi-objective strategy for management of coastal aquifers utilizing planned artificial freshwater recharge. *Model. Earth Syst. Environ.* **2018**, *4*, 111–126. [CrossRef]
64. Lal, A.; Datta, B. Optimal Groundwater-Use Strategy for Saltwater Intrusion Management in a Pacific Island Country. *J. Water Resour. Plan. Manag.* **2019**, *145*, 4019032. [CrossRef]

65. Xu, G.; Xing, X.; Nguyen, V.T.; Poh, H.J.; Lou, J. CFD-driven optimization of air supplies deployment in an air-conditioned office. *IOP Conf. Ser. Earth Environ. Sci.* **2019**, *238*, 12054. [CrossRef]
66. Lal, A.; Datta, B. Multi-objective groundwater management strategy under uncertainties for sustainable control of saltwater intrusion: Solution for an island country in the South Pacific. *J. Environ. Manag.* **2019**, *234*, 115–130. [CrossRef] [PubMed]
67. Hasanzadeh, S.K.; Saadatpour, M.; Afshar, A. A fuzzy equilibrium strategy for sustainable water quality management in river-reservoir system. *J. Hydrol.* **2020**, *586*, 124892. [CrossRef]
68. Pedrielli, G.; Xing, Y.; Peh, J.H.; Koh, K.W.; Ng, S.H. A real time simulation optimization framework for vessel collision avoidance and the case of singapore strait. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 1204–1215. [CrossRef]
69. Guerras, L.S.; Martín, M. On the water footprint in power production: Sustainable design of wet cooling towers. *Appl. Energy* **2020**, *263*, 114620. [CrossRef]
70. Gonçalves, D.; Sheikhnejad, Y.; Oliveira, M.; Martins, N. One step forward toward smart city Utopia: Smart building energy management based on adaptive surrogate modelling. *Energy Build.* **2020**, *223*, 110146. [CrossRef]
71. Leprêtre, F.; Fonlupt, C.; Verel, S.; Marion, V. Combinatorial Surrogate-Assisted Optimization for Bus Stops Spacing Problem. *Lect. Notes Comput. Sci.* **2020**, *12052*, 42–52.
72. Iplik, E.; Aslanidou, I.; Kyprianidis, K. Hydrocracking: A perspective towards digitalization. *Sustainability* **2020**, *12*, 7058. [CrossRef]
73. Rodriguez-Roman, D.; Ritchie, S.G. Surrogate-based optimization for multi-objective toll design problems. *Transp. Res. Part Policy Pract.* **2020**, *137*, 485–503. [CrossRef]
74. Mohajeri, M.J.; van den Bergh, A.J.; Jovanova, J.; Schott, D.L. Systematic design optimization of grabs considering bulk cargo variability. *Adv. Powder Technol.* **2021**, *32*, 1723–1734. [CrossRef]
75. Zhou, T.; Gani, R.; Sundmacher, K. Hybrid Data-Driven and Mechanistic Modeling Approaches for Multiscale Material and Process Design. *Engineering* **2021**, *7*, 1231–1238. [CrossRef]
76. Seyedashraf, O.; Bottacin-Busolin, A.; Harou, J.J. A Disaggregation-Emulation Approach for Optimization of Large Urban Drainage Systems. *Water Resour. Res.* **2021**, *57*, e2020WR029098. [CrossRef]
77. Abdelkader, E.M.; Moselhi, O.; Marzouk, M.; Zayed, T. Integrative Evolutionary-Based Method for Modeling and Optimizing Budget Assignment of Bridge Maintenance Priorities. *J. Constr. Eng. Manag.* **2021**, *147*, 4021100. [CrossRef]
78. Vázquez, D.; Guillén-Gosálbez, G. Process design within planetary boundaries: Application to CO_2 based methanol production. *Chem. Eng. Sci.* **2021**, *246*, 116891. [CrossRef]
79. Ozcanan, S.; Atahan, A.O. Minimization of Accident Severity Index in concrete barrier designs using an ensemble of radial basis function metamodel-based optimization. *Optim. Eng.* **2021**, *22*, 485–519. [CrossRef]
80. Tariq, R.; Cetina-Quiñones, A.J.; Cardoso-Fernández, V.; Daniela-Abigail, H.L.; Soberanis, M.A.; Bassam, A.; De Lille, M.V. Artificial intelligence assisted technoeconomic optimization scenarios of hybrid energy systems for water management of an isolated community. *Sustain. Energy Technol. Assess.* **2021**, *48*, 101561. [CrossRef]
81. Azevedo, L.; Gomes, R.; Silva, C. Influence of model calibration and optimization techniques on the evaluation of thermal comfort and retrofit measures of a Lisbon household using building energy simulation. *Adv. Build. Energy Res.* **2021**, *15*, 630–661. [CrossRef]
82. Wang, H.; Palys, M.; Daoutidis, P.; Zhang, Q. Optimal Design of Sustainable Ammonia-Based Food-Energy-Water Systems with Nitrogen Management. *ACS Sustain. Chem. Eng.* **2021**, *9*, 2816–2834. [CrossRef]
83. Anthony, L.F.W.; Kanding, B.; Selvan, R. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. *arXiv* **2020**, arXiv:2007.03051.
84. Lacoste, A.; Luccioni, A.; Schmidt, V.; Dandres, T. Quantifying the Carbon Emissions of Machine Learning. *arXiv* **2019**, arXiv:1910.09700.
85. Lannelongue, L.; Grealey, J.; Inouye, M. Green Algorithms: Quantifying the Carbon Footprint of Computation. *Adv. Sci.* **2021**, *8*, 2100707. [CrossRef] [PubMed]
86. Van der Blom, K.; Deist, T.M.; Volz, V.; Marchi, M.; Nojima, Y.; Naujoks, B.; Oyama, A.; Tušar, T. Identifying Properties of Real-World Optimisation Problems through a Questionnaire. *arXiv* **2021**, arXiv:cs.NE/2011.05547.

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Sustainability Editorial Office
E-mail: sustainability@mdpi.com
www.mdpi.com/journal/sustainability

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34

www.mdpi.com

ISBN 978-3-0365-6601-6