

A definition, benchmark and database of AI for social good initiatives

Josh Cowsls^{1,2}, Andreas Tsamados¹, Mariarosaria Taddeo ^{1,2} and Luciano Floridi ^{1,2} ✉

Initiatives relying on artificial intelligence (AI) to deliver socially beneficial outcomes—AI for social good (AI4SG)—are on the rise. However, existing attempts to understand and foster AI4SG initiatives have so far been limited by the lack of normative analyses and a shortage of empirical evidence. In this Perspective, we address these limitations by providing a definition of AI4SG and by advocating the use of the United Nations' Sustainable Development Goals (SDGs) as a benchmark for tracing the scope and spread of AI4SG. We introduce a database of AI4SG projects gathered using this benchmark, and discuss several key insights, including the extent to which different SDGs are being addressed. This analysis makes possible the identification of pressing problems that, if left unaddressed, risk hampering the effectiveness of AI4SG initiatives.

In recent years, the development of AI technologies has been driven primarily by commercial interests. However, the number of non-commercial projects leveraging AI technologies to deliver socially beneficial outcomes has proliferated worldwide. These projects can be described as 'socially good AI' (hereafter 'AI4SG'¹).

AI4SG facilitates the attainment of socially good outcomes that were previously unfeasible, unaffordable or simply less achievable in terms of efficiency and effectiveness. It offers unprecedented opportunities across many domains, and could be of great importance, at a time when problems are increasingly global, complex and interconnected. For example, AI can provide much-needed support to improve health outcomes and mitigate environmental risks²⁻⁵. This is also a matter of synergy: AI4SG builds on, and augments, other recent examples of digital technologies adopted to advance socially beneficial objectives, such as 'big data for development'^{6,7}. As a result, AI4SG is gaining traction within the AI community and with policymakers.

Perhaps because of its novelty and fast growth, AI4SG is still poorly understood as a global phenomenon, and lacks a cogent framework for assessing the value and the success of relevant projects. Clearly, existing metrics, such as profitability or commercial productivity, are indicative of real-world demand, but they remain inadequate. AI4SG needs to be assessed against socially valuable outcomes, such as 'B Corporation' certification happens in the for-profit context, or for social enterprises operating in the non-profit sector.

AI4SG should be assessed by adopting human and environmental welfare metrics as opposed to financial ones. Recent frameworks for the design, development and deployment of 'ethical AI' offer some guidance in this respect (see Floridi and Cowsls⁸ for a comparative analysis and synthesis). However, the ethical and social guardrails around applications of AI that are explicitly geared towards socially good outcomes are only partially defined. This is because there is limited understanding of what constitutes AI4SG at present⁹⁻¹¹, and what would be a reliable benchmark with which to assess its success. The best efforts, especially the annual International Telecommunication Union Summit on AI for Good¹² and its associated project database¹³, focus on collecting information about, and describing occurrences of, AI4SG, but disregard normative approaches to this phenomenon and are not meant to offer a sys-

tematic analysis. This Perspective article fills this gap, by formalizing a definition of AI4SG initiatives and by arguing that the 17 United Nations (UN) SDGs provide a valid framework with which to benchmark socially good uses of AI technologies. To support the analysis, we introduce a database of AI4SG projects gathered using this benchmark, and discuss several key insights, including the extent to which different SDGs are being addressed.

What qualifies as AI4SG?

So far, AI4SG has been developed ad hoc, by analysing specific areas of application, such as famine relief or disaster management. This approach can indicate the presence of a phenomenon, but it cannot explain what exactly makes AI socially good nor can it indicate how AI4SG solutions could and should be designed and deployed to harness the full potential of the technology. These two shortcomings raise at least three main risks: unanticipated failures, missed opportunities and unwarranted interventions.

We consider unanticipated failures first. Like any other technology, AI solutions are shaped by human values. Such values, if not carefully selected and fostered, may lead to 'good AI gone awry' scenarios. AI may 'do more harm than good', amplifying rather than mitigating societal ills, for example, by widening rather than narrowing existing inequities, or by exacerbating environmental problems. AI may simply fail to serve the social good. For example, consider the failure of IBM's oncology-support software, which attempted to use machine learning to identify cancerous tumours. The system was trained using synthetic data and US medical protocols, which are not applicable worldwide. As a result, it struggled to interpret ambiguous, nuanced or otherwise 'messy' patient health records¹⁴, and provided misdiagnoses and erroneous treatment suggestions. This led medical practitioners and hospital to reject the Watson system¹⁵.

Next, we consider missed opportunities. The genuinely socially good outcomes of AI may arise merely accidentally, for example, through a fortuitous application of an AI solution in a different context. This happened with the use of a different version of the IBM cognitive system discussed above. In this case, the Watson system was originally designed to identify biological mechanisms, but when used in a classroom setting, it inspired engineering students to solve design problems¹⁶. In this positive instance, AI provided

a unique mode of education. But lacking a clear understanding of AI4SG meant that this success was accidental, and it may not be possible to repeat it systematically or at scale. For each accidental success, therefore, there may be countless examples of missed opportunities to exploit the benefits of AI for advancing socially good outcomes in different settings, especially when AI-based interventions are developed separately from those who will be most directly affected, whether this is defined in terms of area (that is residents of a particular region) or domain (for example, teachers or medical practitioners).

Conversely, there are many circumstances in which AI will not be the most effective way to address a particular social problem¹⁷, and would therefore be an unwarranted intervention. This could be due to the existence of alternative approaches that are less expensive or more efficacious (that is, 'Not AI for social good' may be preferable), or because of the unacceptable risks that the deployment of AI would introduce (that is, 'AI for insufficient social good', as weighed against its risks). This is why the use of the term 'good' to describe such efforts has itself been criticized¹⁸. Indeed, AI should not be treated as a single solution to an entrenched social problem (that is, 'Only AI for social good' is unlikely to work).

A successful way to identify and evaluate AI4SG projects is to analyse them on the basis of their outcomes. An AI4SG project is successful insofar as it helps to reduce, mitigate or eradicate a given social or environmental problem, without introducing new harms or amplifying existing ones. This interpretation suggests the following definition of AI4SG:

AI4SG is formally defined as the design, development and deployment of AI systems in ways that help to (i) prevent, mitigate and/or resolve problems adversely affecting human life and/or the wellbeing of the natural world, and/or (ii) enable socially preferable or environmentally sustainable developments, while (iii) not introducing new forms of harm and/or amplifying existing disparities and inequities.

Assuming that this working definition is acceptable, the challenge becomes the effective identification of problems that are deemed to affect human life or the wellbeing of the environment negatively. Our strategy, as for others¹⁰, has been to use the 17 UN SDGs as an assessment benchmark.

AI4SG and the UN SDGs

The SDGs were set by the United Nations General Assembly in 2015 to integrate the economic, social, and environmental dimensions of sustainable development. They are internationally agreed priorities for socially beneficial action, and thus constitute a sufficiently empirical and reasonably uncontroversial benchmark to evaluate the positive social impact of AI4SG globally. Using the SDGs to evaluate AI4SG applications means equating AI4SG with AI that supports the SDGs (AI×SDGs).

This move, to set AI4SG = AI×SDGs, may seem restrictive because there is undoubtedly a multitude of examples of socially good uses of AI outside the scope of the SDGs. Nonetheless, the approach carries clear advantages, of which five are paramount. First, the SDGs offer clear, well defined and shareable boundaries to identify positively what is socially good AI (what should be done, as opposed to what should be avoided), although they should not be understood as indicating what is not socially good AI. Second, the SDGs are internationally agreed goals for development, and have begun informing relevant policies worldwide, so they raise fewer questions about relativity and cultural dependency of values. Although they are of course improvable, they are nonetheless the closest thing we have to a humanity-wide consensus on what ought to be done to promote positive social change and the conservation of our natural environment. Third, the existing body of research on SDGs already includes studies and metrics on how to measure progress in attaining each of the 17 SDGs, and the 169 associated

targets defined in the 2030 Agenda for Sustainable Development. These metrics can be applied to evaluate the impact of AI×SDGs projects^{10,19}. Fourth, focusing on the impact of AI-based projects across different SDGs can improve existing, and lead to new, synergies between projects addressing different SDGs, further leveraging AI to gain insights from large and diverse datasets, and can pave the way to more ambitious collaborative projects. Fifth and finally, understanding AI4SG in terms of AI×SDGs enables better planning and resource allocation, once it becomes clear which SDGs are under-addressed and why.

Assessing evidence of AI×SDGs

In view of the advantages of using the UN SDGs as a benchmark for assessing AI4SG, we conducted an international survey of AI×SDG projects. The survey ran between July 2018 and November 2020, and it involved collecting data on AI×SDG projects that met the following five criteria:

1. Only projects that actually addressed (even if not explicitly) at least one of the 17 SDGs;
2. Only real-life, concrete projects relying on some actual form of AI (symbolic AI, neural networks, machine learning, smart robots, natural language processing, and so on), not merely referring to AI (an observed problem among AI start-ups more generally²⁰);
3. Only projects built and used in the field for at least six months, rather than theoretical projects or research projects yet to be developed (for example, patents, or grant programmes);
4. Only projects with documented positive impact, for example through a web site, a newspaper article, a scientific article, a non-governmental organization (NGO) report, and so on;
5. Only projects with no or minimal evidence of counter-indications or negative side-effects.

Requirements (3) and (4) were crucial to unearthing concrete examples of AI×SDG, that is, projects with a proved record of robust, positive impact, as opposed to identifying research projects or tools developed in laboratories and trained on data that may prove to be inadequate or unfeasible when the technology is deployed outside controlled environments. No constraints were assumed about who developed the project, where, or by whom it was used, who supported it financially, or whether the project was open-source, with the exception that projects conducted solely by commercial entities with entirely proprietary systems were excluded.

The projects were discovered via a combination of resources, including academic databases (arXiv and Scopus), government press releases, patent filings, reports tracking organisations' public commitment to the UN SDGs, and existing databases, including that of the International Telecommunication Union¹³ and the database by the Inter-American Development Bank's fAIR LAC partnership^{21,22}. This approach made it possible to build on existing work by Vinuesa and colleagues¹⁰—who used an expert elicitation process to ascertain which of the SDGs could potentially be affected by AI—by offering empirical evidence of actual benefits already being felt in the domains of various SDGs.

From a larger pool, the survey identified 108 projects in English, Spanish and French matching these criteria. The data about the AI×SDG projects collected in this study are publicly available in the aforementioned database (<https://www.aiforsdgs.org/all-projects>), which is part of the Oxford Research Initiative on Sustainable Development Goals and Artificial Intelligence²³. We presented a preliminary version of our analysis in September 2019 at a side-event during the annual UN General Assembly.

Our analysis (see below) shows that every SDG is already being addressed by at least one AI-based project. The analysis indicates that the use of AI×SDGs is an increasingly global phenomenon—with projects operating from five continents—but also that the

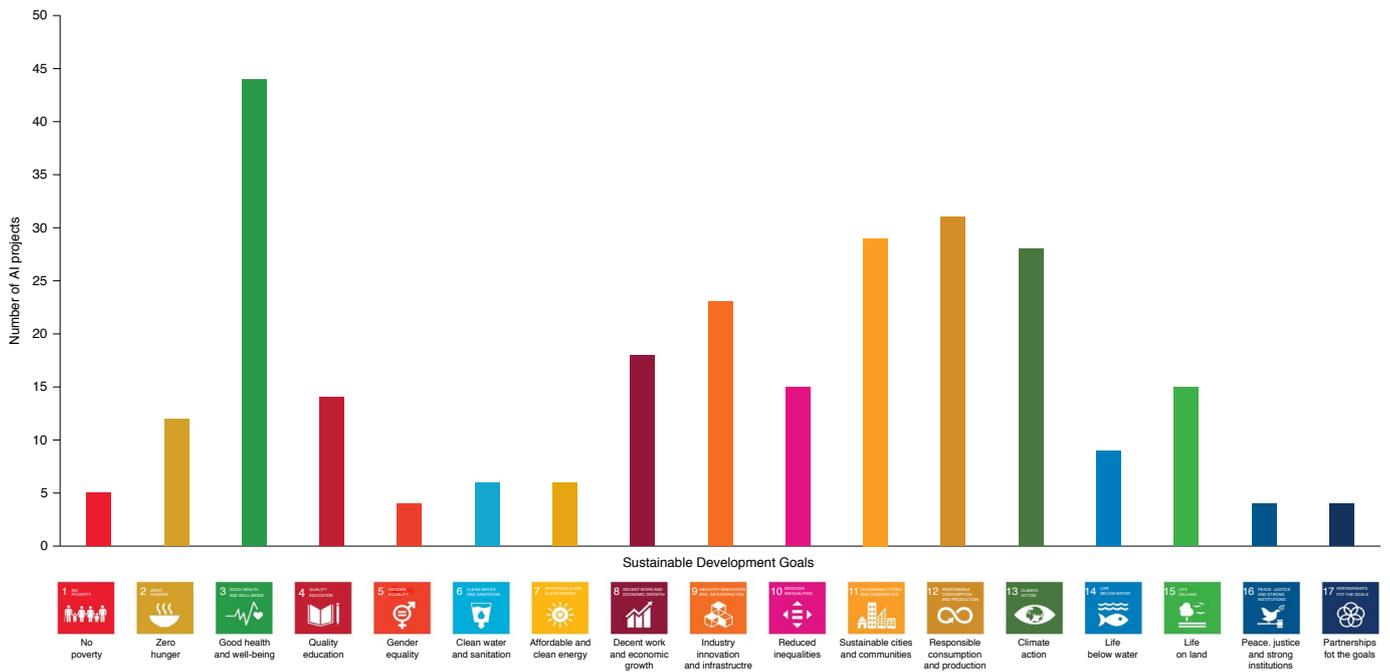


Fig. 1 | Projects addressing the SDGs. A survey sample of 108 AI projects found to be addressing the SDGs globally.

phenomenon may not be equally distributed across the SDGs (Fig. 1). SDG 3 (‘Good Health and Well-Being’) leads the way, while SDGs 5 (‘Gender Equality’), 16 (‘Peace, Justice and Strong Institutions’), and 17 (‘Partnerships for the Goals’) appear to be addressed by fewer than five projects (see Fig. 1).

It is important to note that the use of AI to tackle at least one of the SDGs does not necessarily result in success. We note also that a project could address multiple SDGs simultaneously or at different timescales and in different ways. Moreover, even complete success for a given project would be exceedingly unlikely to result in the eradication of all of the challenges associated with an SDG. This is chiefly because each SDG concerns entrenched challenges that are widespread and structural in nature. This is well reflected by the way SDGs are organized: all 17 SDGs have several targets, and some targets in turn have more than one metric of success. The survey shows which SDGs are being addressed by AI at a high level, but more finely grained analysis is required to assess the extent of the positive impact of AI-based interventions with respect to specific SDG indicators, as well as possible cascade effects and unintended consequences.

The unequal allocation of efforts detected by our survey may be due to the constraints of our survey criteria but, given the degree of systematic analysis and search conducted, it more probably signals underlying divergence in how suitable it is to use AI technology to address each SDG. For instance, the suitability of AI to a given problem also rests on the ability to formalize that problem at a suitable level of abstraction. It may be that SDGs such as ‘Gender Equality’ or ‘Peace and Justice and Strong Institutions’ are harder to formalize than problems which pertain more directly to the allocation of resources, such as ‘Affordable and Clean Energy’ or ‘Clean Water and Sanitation’. We also observe a different allocation of efforts along geographical lines. For example, ‘Reduced Inequalities’, ‘Quality Education’ and ‘Good Health and Well-Being’ were the main goals pursued by projects based in South America (25 out of the 108 projects). The more detailed questions prompted by the survey, such as what explains the observed divergence and how it may be overcome or what may explain the different geographical distribution of projects will require further work and are being addressed by our current research.

It is worth stressing that, although one criterion for the survey was that the projects must have already demonstrated positive impact, in many cases this impact was ‘only’ local or at an early stage. Questions therefore remain about how best to—or indeed in each case whether to—‘scale up’ existing solutions to apply them at regional or even global levels. The idea of scaling up solutions is attractive, since it implies that demonstrable success in one domain or area can be replicated elsewhere, reducing the costs of duplication (not to mention the computationally intense and hence environmentally problematic training of AI systems). Indeed, as we highlight below, learning lessons from successes and failures is another critical area for future research. But in asking how successes can be scaled up, it is important not to overlook the fact that most of the projects in our survey already represent a ‘scaling down’ of existing technology. More specifically, most examples of AI×SDG reflect the repurposing of existing AI tools and techniques (developed in silico in academic or industrial research contexts) for the specific problem at hand. This can in part explain why certain SDGs such as ‘Good Health and Well-Being’ are seeing more AI×SDG projects flourish than others, such as ‘Gender Equality’, where tools and techniques are relatively lacking or not yet as mature. This also suggests that AI×SDG involves first a ‘funnelling in’, where numerous (in silico and/or in vivo) options are considered in order to address a particular SDG target in a particular place, and then a ‘fanning out’, which involves the spread and iterative adoption of verified successes in adjacent domains and areas.

The analysis suggests that the 108 projects meeting the criteria correspond with seven essential factors for socially good AI, identified in ref. ²⁴: falsifiability and incremental deployment; safeguards against the manipulation of predictors; receiver-contextualized intervention; receiver-contextualized explanation and transparent purposes; privacy protection and data subject consent; situational fairness; and human-friendly semanticization. Each factor relates to at least one of the five ethical principles of AI—beneficence, nonmaleficence, justice, autonomy and explicability—identified in the comparative analysis of ref. ⁸. This coherence is crucial: AI×SDGs cannot be inconsistent with ethical frameworks guiding the design and evaluation of any kind of AI. The principle of

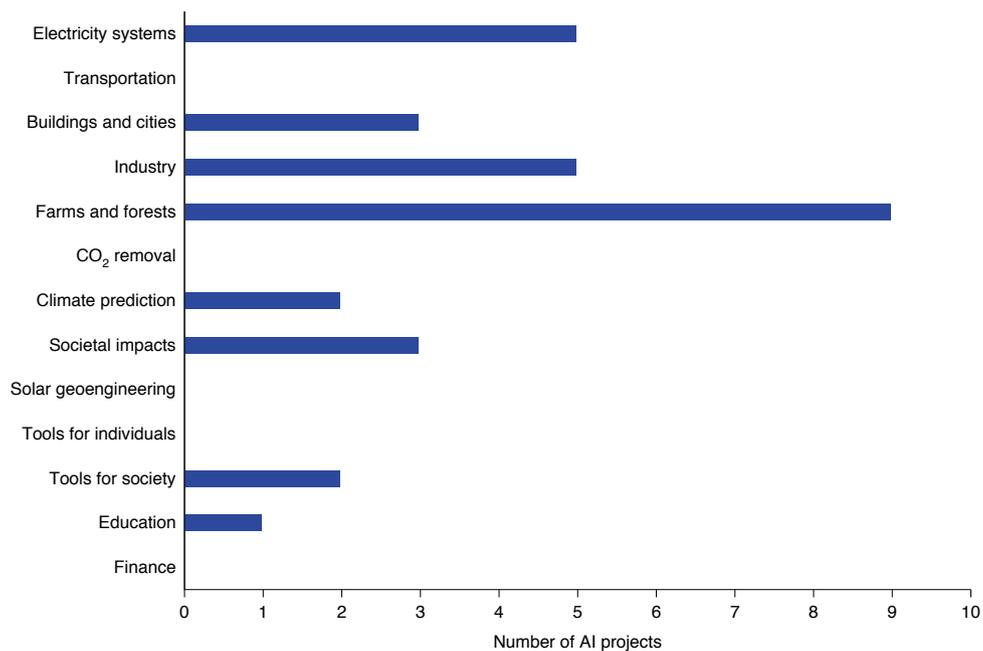


Fig. 2 | Projects addressing different aspects of climate action. AI projects in our dataset that are exploiting different domains of climate crisis response, as identified by Rolnick et al.⁴.

benevolence is of particular relevance when considering AI×SDGs, as it states that the use of AI should benefit humanity and the natural world. AI×SDG projects should therefore respect and implement this principle. However, although benevolence is a necessary condition of the success of AI×SDGs, it is not sufficient. The beneficial impact of an AI×SDGs project may be offset by the creation or amplification of other risks or harms. Ethical analyses informing the design, development and deployment (including monitoring) of AI×SDGs initiatives are essential in mitigating foreseeable risks and avoiding unintended consequences and possible misuses of the technology.

AI to advance ‘climate action’

In terms of current focus, SDG 13, ‘Climate action’, ranks fourth in our database, with 28 of the 108 initiatives tackling it. This is despite the environmental problems linked to the use of AI, that is, the intense computational requirements—and therefore energy consumption—that training successful deep learning systems necessitates^{25,26} (Covels, J. et al., manuscript in preparation).

To explore the extent to which AI is already being developed to tackle SDG 13, and the specific ways in which this is occurring, we cross-referenced the initiatives in our dataset that were coded as addressing the ‘Climate action’ SDG with the areas of prospective use recently identified in a large-scale scoping effort undertaken by Rolnick et al.⁴. As Fig. 2 details, at least one initiative in our dataset addresses eight of the thirteen aspects of climate action identified in ref.⁴.

Projects relying on AI to support ‘climate action’ in our dataset are based across a number of countries, which suggests reasonable geographic spread, but it is important to note that most of these countries (Australia, France, Germany, Japan, Slovenia, South Korea, the UAE, the UK and the USA) are in the Global North. Only four projects were based in the Global South Hemisphere, specifically in Argentina, Peru and Chile. This is not to suggest that initiatives based in the Global North are not having an impact elsewhere in the world; to take one example, Global Forest Watch is a project based in the UK that is attempting to track and protect forests worldwide. Nonetheless, this finding highlights the risk that projects based (and

funded) in one part of the world may not necessarily be responsive to actual needs elsewhere.

Overall, this case study provides promising preliminary evidence that AI is in fact being used to tackle climate change and associated problems. As the cross-referencing effort above shows, this dovetails with wider research, suggesting that AI could and should be developed and used for this purpose. As Rolnick and colleagues show, AI could support efforts to mitigate climate change in thirteen existing and potential domains, ranging from CO₂ removal and transport optimization to forestry protection⁴. The potential of AI to support climate action has also been recognized by a consortium of academics, NGOs and energy companies, who wrote to the UK Government in 2019 to call for the establishment of an International Centre for AI, Energy and Climate²⁷.

This analysis shows that there are already ‘boots on the ground’ using AI to tackle the climate crisis, even if such efforts are only at an early stage. Given the strict criteria applied in our sampling process (for example, that projects must have had evidence of positive impact), our evidence is encouraging. At the same time, it highlights that there is much more to be done, with several existing gaps towards which prospective initiatives could orient their efforts.

Conclusion

There is a growing number of projects that are using AI4SG by addressing the UN SDGs. AI technologies cannot solve all problems, but they can help to address the major challenges, both social and environmental, facing humanity today. If designed well, AI technologies can foster the delivery of socially good outcomes with unprecedented scale and efficiency. Therefore, it is crucial to provide a coherent structure within which new and existing AI×SDG projects can thrive. The next steps in understanding AI4SG in terms of AI×SDGs are: to analyse what factors determine the success or failure of AI×SDG projects, particularly with respect to their specific impacts ‘on the ground’; to explain gaps and discrepancies between the use of AI to tackle different SDGs and indicators, as well as mismatches between where projects are based and where SDG-related need is greatest; and to clarify how key stakeholders could advance the success of AI×SDG projects and address important gaps and

discrepancies. Inevitably, all this will require a multidisciplinary approach, and involve deeper investigation of AI×SDGs projects in locations and communities where they are both developed and deployed.

References

1. Hager, G. D. et al. Artificial intelligence for social good. Preprint at <https://arxiv.org/abs/1901.05406> (2017).
2. Wang, D., Khosla, A., Gargeya, R., Irshad, H. & Beck, A. H. Deep learning for identifying metastatic breast cancer. Preprint at <https://arxiv.org/abs/1606.05718> (2016).
3. Davenport, T. & Kalakota, R. The potential for artificial intelligence in healthcare. *Future Healthc. J.* **6**, 94–98 (2019).
4. Rolnick, D. et al. Tackling climate change with machine learning. Preprint at <https://arxiv.org/abs/1906.05433> (2019).
5. Zhou, Y., Wang, F., Tang, J., Nussinov, R. & Cheng, F. Artificial intelligence in COVID-19 drug repurposing. *Lancet Digit. Health* **2**, e667–e676 (2020).
6. Hilbert, M. Big data for development: a review of promises and challenges. *Dev. Policy Rev.* **34**, 135–174 (2016).
7. Taylor, L. & Schroeder, R. Is bigger better? The emergence of big data as a tool for international development policy. *GeoJournal* **80**, 503–518 (2015).
8. Floridi, L. & Cows, J. A unified framework of five principles for AI in society. *Harv. Data Sci. Rev.* **1**, <https://doi.org/10.1162/99608f92.8cd550d1> (2019).
9. Taddeo, M. & Floridi, L. How AI can be a force for good. *Science* **361**, 751–752 (2018).
10. Vinuesa, R. et al. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nat. Commun.* **11**, 1–10 (2020).
11. Chui, M. et al. *Notes From The AI frontier: Insights From Hundreds Of Use Cases*. (McKinsey Global Institute, 2018).
12. International Telecommunication Union (ITU) *AI Repository*; <https://www.itu.int/en/ITU-T/AI/Pages/ai-repository.aspx>
13. *AI for Good Global Summit (28–31 May 2019, Geneva, Switzerland)* (AI for Good, 2019); <https://aiforgood.itu.int/>
14. Strickland, E. How IBM Watson overpromised and underdelivered on AI health care. In *IEEE Spectrum: Technology, Engineering, and Science News* <https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care> (IEEE, 2019).
15. Ross, C. & Swetlitz, I. IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. In *STAT* <https://www.statnews.com/2017/09/05/watson-ibm-cancer/> (5 September 2017).
16. Goel, A. et al. Using Watson for enhancing human-computer co-creativity. In *2015 AAAI Fall Symp. Ser.* (IEEE, 2015).
17. Abebe, R. et al. Roles for computing in social change. In *FAT* '20: Proc. 2020 Conf. on Fairness, Accountability, and Transparency* (ACM, 2019); <https://doi.org/10.1145/3351095.3372871>
18. Green, B. 'Good' isn't good enough. In *Proc. AI for Social Good Worksh. NeurIPS* (2019).
19. United Nations Development Program (UNDP) *Sustainable Development Goals*. <https://www.undp.org/content/undp/en/home/sustainable-development-goals.html> (UNDP, 2015).
20. Ram, A. Europe's AI start-ups often do not use AI, study finds. *Financial Times* (5 March 2019).
21. Strubell, E., Ganesh, A. & McCallum, A. Energy and policy considerations for deep learning in NLP. *Proc. 57th Annual Meeting of the Association for Computational Linguistics* 3645–3650 (ACL, 2019).
22. *Inter-American Development Bank fAIRLAC Observatory* (IADB, 2020); <https://fairlac.iadb.org/en/observatory>
23. *Oxford Initiative on AI×SDGs*. <https://www.aiforsdgs.org/> (2020).
24. Floridi, L., Cows, J., King, T. C. & Taddeo, M. How to design AI for social good: seven essential factors. *Sci. Eng. Ethics* **26**, 1771–1796 (2020).
25. Dandres, T. et al. Consequences of future data center deployment in Canada on electricity generation and environmental impacts: A 2015–2030 prospective study. *J. Ind. Ecol.* **21**, 1312–1322 (2017).
26. Strubell, E., Ganesh, A. & McCallum, A. Energy and policy considerations for deep learning in NLP. Preprint at <https://arxiv.org/abs/1906.02243> (2019).
27. Shrestha, P. Leading energy and tech groups call for International Centre for AI, Energy and Climate. *Energy Live News* <https://www.energylivenews.com/2019/08/20/leading-energy-and-tech-groups-call-for-international-centre-for-ai-energy-and-climate/> (20 August 2019).

Acknowledgements

J.C. acknowledges the receipt of a Doctoral Studentship from the Alan Turing Institute. M.T. and L.F. acknowledge the Oxford Initiative on AI for SDG, which is supported by grants from Facebook, Google and Microsoft.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to L.F.