https://doi.org/10.1038/s41893-025-01536-6

The Earth alignment principle for artificial intelligence

Owen Gaffney, Amy Luers, Franklin Carrero-Martinez, Berna Oztekin-Gunaydin, Felix Creutzig, Virginia Dignum, Victor Galaz, Naoko Ishii, Francesca Larosa, Maria Leptin & Ken Takahashi Guevara

At a time when the world must cut greenhouse gas emissions precipitously, artificial intelligence (AI) brings large opportunities and large risks. To address its uncertain environmental impact, we propose the 'Earth alignment' principle to guide AI development and deployment towards planetary stability.

According to the most recent Earth system assessment, six of nine planetary boundaries have been transgressed and climate tipping points are much closer than previously thought[1,2]. This requires urgent, disruptive and cooperative action to address the underlying drivers of planetary instability. It has been routinely stated that artificial intelligence (AI), a disruptive technology, could be a powerful tool to accelerate innovation and scale climate action, protect biodiversity and reduce environmental degradation[3,4]. However, there is a paradox at the heart of AI's potential contribution to stewardship of Earth: AI is being deeply embedded in the existing global socio-economic system responsible for driving the destabilization of the Earth system.

Conversations regarding AI and the Earth system often focus narrowly on AI's own rapidly growing environmental footprint[5]. Given the severely constrained carbon budget to meet the United Nations (UN) Paris Agreement, any emissions growth is deeply concerning. However, the systemic societal changes AI is likely to drive, and resultant emissions trajectories, could be a larger concern[6].

A key feature of technological advance is the Jevons paradox, a dynamic where efficiencies drive down production costs, leading to cheaper products driving increased consumer demand[7]. This ultimately results in greater overall consumption, waste and emissions. The paradox is relevant to AI development, but with potentially unprecedented implications. Unlike many previous technologies, AI's impact has the potential to be exceptionally broad, affecting virtually every sector of society[3,8]. The unique and increasing capacity of AI for generalized problem solving, rapid learning and autonomous adaptation suggests it could impact economic sectors at a scale and acceleration far beyond previous technological developments. Moreover, AI's potential to learn and improve itself sets it apart from previous technological shifts, possibly disrupting the process of scientific and technological inquiry. This could create a rapid feedback loop of innovation, production and consumption with profound implications for the Earth system.

Beyond economic impacts, AI poses risks of large-scale social harms, including exacerbating social injustice, eroding social stability and weakening our shared understanding of reality[8]. Given these potentially far-reaching effects on the economy, society and Earth, the development and deployment of AI warrant even greater concern than previous technologies affected by the Jevons paradox, requiring new governance approaches to manage this risk.

Aligning AI policy with stewardship of Earth

As governments mobilize to address the complex challenges posed by AI, some are beginning to grapple with two interconnected yet distinct concepts: the alignment problem — from a socio-technical standpoint, ensuring AI systems reliably pursue objectives beneficial to humanity — and the ethical implications dominated by safety, security and economic concerns (for example, risk of job losses).

In August 2024, the European Union's AI Act came into force, categorizing AI system risk from 'minimal' and 'limited' to 'high' and 'unacceptable'. The Act aims to protect European citizens from, for example, biometric categorization, social scoring, predictive policing and subliminal influencing. Other countries are exploring policy responses and the UN is taking steps to advance global policy (Governing AI for Humanity). The United States is a notable exception. A 2023 Executive Order, 'Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence', which included environmental considerations, was revoked by the new administration on 20 January 2025. In addition, several large technology companies have developed internal governance frameworks for ethical and human-aligned AI. While these efforts are crucial, a critical foundational principle is missing from current policy discussions: the urgent need to align AI systems with the stability of the Earth system.

Good AI governance for planetary stewardship must focus both on the opportunities to mitigate Earth system risks and on reining in forces with the potential to accelerate destabilization. To address this, here we introduce the 'Earth alignment' principle for AI. Principles are used to guide behaviour in complex social systems. Examples include the 'precautionary principle', the principle of 'do no harm' or, in climate negotiations, 'common but differentiated responsibility'. We define Earth alignment as the principle of aligning the development, deployment and use of AI to promote planetary stability and stewardship for the benefit of humankind. This principle covers both the technical focus of AI alignment and the broader ethical considerations of AI's impact on our planet, emphasizing that truly beneficial AI must not only serve immediate human goals but also safeguard the long-term health and resilience of the global environment on which all life depends.

The Earth alignment principle is differentiated from the principle of 'do no harm' or the 'precautionary principle' in that it actively seeks to leverage AI's capabilities to drive sustainability transformations across human activities while constraining developments that impede this goal. This leverage includes, for example, AI deployment and use that drives down greenhouse gas emissions or contributes to reversing loss of biodiversity. Earth alignment, therefore, provides a unique

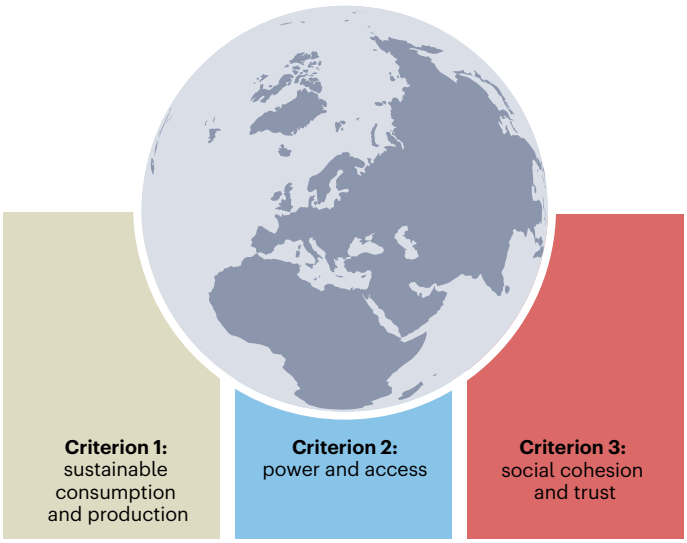

Fig. 1 | Achieving strong Earth alignment requires taking account of three criteria simultaneously. AI's transformative impact demands a systemic stewardship approach that goes beyond direct environmental impacts to encompass trajectories of economies and societies.

safeguard alongside other AI principles in that it specifically values the state of the biosphere.

Earth alignment is not binary. AI systems will exist on a spectrum ranging from strong alignment to weak alignment or even strong misalignment. Here we outline three criteria for strong Earth alignment (Fig. 1) that must be met over the entire life cycle of AI systems, their development, deployment and use.

- AI systems should help to accelerate the transition to sustainable production and consumption in ways that respect planetary boundaries, or at least do not obstruct these objectives[4,9] (criterion 1).
- AI systems should be developed, deployed and used in ways that ensure equitable access to AI tools for global sustainability and avoid concentrations of power (criterion 2).
- AI systems should be developed, deployed and used to support greater societal cohesion, build trust and provide access to reliable information for planetary stewardship (criterion 3).

Criterion 1. Strongly aligned AI development and deployment prioritizes innovation and deployment for rapid decarbonization and biosphere preservation and restoration. In the energy sector, AI is already being used to support grid decarbonization and reliability through better forecasting[10]. AI has the potential to bring efficiencies to manufacturing and supply chains, and support development of new materials and processes[11]. In the food and health sectors it can improve fertilizer application, support personalized sustainable diets, and deliver more accurate weather and disease forecasting[12–14]. In transport and urban planning, AI has the potential to deliver systemic efficiency gains, for example, by enabling a shift from private cars to shared mobility. But innovations must be viewed systemically and consider the Jevons paradox. Autonomous vehicles, for example, may increase congestion due to empty trips and preferences for longer commutes and cars over public transport[15].

Without good governance, AI has the potential to supercharge unsustainable production and consumption through, for example, being used for enhanced fossil fuel extraction or AI-based microtargeting in advertising. If business models for AI systems are based on advertising revenues, for example, the profit incentive may be in direct conflict with emissions reductions goals. If uses of AI cause cheap products to flood the market, even if emissions per unit are lower, the sheer volumes may drive up emissions and degrade nature. Implementing the Earth alignment principle means managing these risks, designing for circularity and ensuring that all AI-assisted production and consumption respects safe and just Earth system boundaries[16].

The environmental impacts of the development and operations of AI models also need to be addressed. Strong Earth alignment must apply to energy and material used for AI infrastructure. AI development should aim for full decarbonization. Data centres must avoid excessive water consumption in water-stressed regions and be built for material circularity and efficiency.

Criterion 2. Planetary stewardship is a collective action problem[6]. As AI tools become increasingly powerful, economic and technological power will probably concentrate within wealthy nations with profound consequences for collective action challenges. Research has shown that AI applications for achieving the Sustainable Development Goals are biased mainly towards issues relevant to nations where most AI researchers live[3]. If AI systems for sustainability are primarily designed to meet the needs of wealthy nations, transformation elsewhere could slow and social inequalities could widen. Widening inequality reduces trust in governance institutions, which can hamper sustainability goals[17]. Hence, more equitable access to AI tools, and checks on concentration of power, are important criteria for Earth alignment.

Criterion 3. Planetary stewardship requires social cohesion to allow societies to take long-term decisions. Societies with higher social and political trust are better at reaching agreement on public good provisions and protection of global commons[17]. Without strong governance, AI technologies can be weaponized to destabilize democracies and slow or reverse progress towards stewardship goals. Misaligned technologies can supercharge the creation and dissemination of disinformation and misinformation, and thereby undermine trust in democratic institutions and the science needed to tackle the climate crisis. Conversely, strongly aligned systems will support collaboration, social cohesion and informed collective decision-making for stewardship of Earth.

The three criteria overlap with other ethical AI principles relating to issues around political power or social cohesion. However, there are important distinctions: Earth alignment prioritizes deep integration of environmental stewardship with socio-economic stability, and it recognizes that ecological sustainability is more challenging without social cohesion and trust in governance institutions. By adhering to these criteria and this integrated approach, we can address systemic risk and the paradox of AI's larger socio-economic impact. The recent report from the UN Secretary-General's High-Level Advisory Body on AI (Governing AI for Humanity) provides strong support for Earth alignment, advocating for inclusive global frameworks that prevent power concentration and equitable access to AI technology, and emphasizing the role of AI in achieving sustainability goals.

Applying the Earth alignment principle

The goal of the principle is to guide innovation, use and deployment of AI. As such, it is relevant to three groups of actors: governments and international organizations, companies, and investors.

Nations and international bodies such as the UN should consider explicit categorization of AI systems or applications as 'high risk' or 'unacceptable risk' if technologies could foreseeably be a threat to the stability of the Earth system based on the criteria above. A second priority is to strategically direct investments in AI-driven research and innovation towards planetary stewardship emphasizing open-source initiatives and investments beyond the wealthiest nations. This needs to be coupled with incentives for global AI tech transfer, including towards actors with modest resources such as municipalities to support initiatives like low-carbon urban planning or AI-related agricultural innovations. It should also include policies to ensure fair access to critical AI infrastructure, such as cloud computing and data, and prevent power concentrations. With the most powerful AI tools concentrated in wealthy nations, this will be challenging to achieve.

National and international institutions must move to mandatory environmental reporting for companies. This should include broadening the scope of AI's environmental impact beyond operations to include business models and societal impact, for example, risks relating to misinformation and disinformation. Compliance mechanisms will be needed urgently to ensure the principle is adopted.

Companies, organizations and other entities developing and using AI tools should include Earth alignment within their governance frameworks and risk assessments. This starts with the business models for AI tools: can models based on advertising revenues be compatible with systemic societal emissions reductions and societal cohesion? Are other business models more likely to achieve these goals?

Scale out of AI must not compromise absolute reduction in carbon and ecological footprints from operations, breaking with the current trend of rising carbon dioxide emissions from AI operations. Recent advances, for example, the DeepSeek model, indicate that high performance may be possible with more efficient algorithms and smaller infrastructure, although in the longer run these efficiencies may boost demand as per the Jevons paradox[18].

Beyond metrics like greenhouse gas emissions and ecosystem degradation, risk assessments should evaluate whether AI systems foster social cohesion and trust. This could involve measuring access to reliable, non-biased data, and evaluating misinformation or manipulation through AI-driven platforms. It may require restricting or prohibiting algorithmic design in social media that actively harms social trust and feeds polarization.

Investors and philanthropy should include Earth alignment as a criterion for AI investment. Finally, all actors will need incentives to share information and know-how across the north–south divide.

The growth in power and ubiquity of AI systems creates an immense opportunity to accelerate action to stabilize the Earth system, but also carries unprecedented risks of destabilization. We believe that Earth alignment should be a foundation for an international framework to guide AI system development, deployment and use in order to manage systemic risks in relationship to Earth system stability and stewardship. The definition of Earth alignment may evolve as AI technologies progress and their influence becomes more pervasive.

Owen Gaffney[1,2,3] ✉, Amy Luers[4], Franklin Carrero-Martinez[5], Berna Oztekin-Gunaydin[5], Felix Creutzig[6,7,8], Virginia Dignum[9], Victor Galaz[10,11], Naoko Ishii[12], Francesca Larosa[13], Maria Leptin[14] & Ken Takahashi Guevara[15]

[1]Nobel Prize Outreach, Stockholm, Sweden. [2]Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, The Netherlands. [3]Exponential Roadmap Initiative, Stockholm, Sweden. [4]Microsoft Corporation, Redmond, WA, USA. [5]National Academy of Sciences, Washington DC, USA. [6]Potsdam Institute for Climate Impact Research (PIK), Potsdam, Germany. [7]Bennett Institute for Innovation and Policy Acceleration, Business School, University of Sussex, Brighton, UK. [8]Technical University Berlin, Berlin, Germany. [9]AI Policy Lab, Umeå University, Umeå, Sweden. [10]Stockholm Resilience Centre, Stockholm University, Stockholm, Sweden. [11]Beijer Institute of Ecological Economics, Royal Swedish Academy of Sciences, Stockholm, Sweden. [12]Center for Global Commons, University of Tokyo, Tokyo, Japan. [13]FLOW, Engineering Mechanics, KTH Royal Institute of Technology, Stockholm, Sweden. [14]European Research Council, Brussels, Belgium. [15]Instituto Geofísico del Perú, Lima, Peru.

✉e-mail: owen.gaffney@nobelprize.org

Published online: 28 March 2025

References

1. Richardson, K. et al. *Sci. Adv.* **9**, eadh2458 (2023).
2. Armstrong McKay, D. I. et al. *Science* **377**, eabn7950 (2022).
3. Vinuesa, R. et al. *Nat. Commun.* **11**, 233 (2020).
4. Larosa, F. et al. *Nat. Clim. Change* **13**, 497–499 (2023).
5. Luers, A. et al. *Nature* **628**, 718–720 (2024).
6. Creutzig, F. et al. *Annu. Rev. Environ. Resour.* **47**, 479–509 (2022).
7. Alcott, B. *Ecol. Econ.* **54**, 9–21 (2005).
8. Bengio, Y. et al. *Science* **384**, 842–845 (2024).
9. Rolnick, D. et al. *ACM Comput. Surv.* **55**, 42 (2022).
10. Rozite, V., Miller, J. & Oh, S. Why AI and energy are the new power couple. *IEA* (2 November 2023); https://go.nature.com/4bFupwT
11. Papadimitriou, I., Gialampoukidis, I., Vrochidis, S. & Kompatsiaris, I. *Comput. Mater. Sci.* **235**, 112793 (2024).
12. Price, I. et al. *Nature* **637**, 84–90 (2025).
13. Thadani, N. N. et al. *Nature* **622**, 818–825 (2023).
14. Papastratis, I., Konstantinidis, D., Daras, P. & Dimitropoulos, K. *Sci. Rep.* **14**, 14620 (2024).
15. *Self-Driving Vehicles: Seventh Report of Session 2022–23* (UK House of Commons Transport Committee, 2023); https://go.nature.com/4bF9xWF
16. Rockström, J. et al. *Nature* **619**, 102–111 (2023).
17. Wilkinson, R. G. & Pickett, K. E. *Nature* **627**, 268–270 (2024).
18. Gibney, E. *Nature* https://doi.org/10.1038/d41586-025-00229-6 (2025).

Acknowledgements

This paper is an outcome of the 'Global Sustainability and Science Integrity in the Age of Generative AI' workshop convened by the US National Academy of Sciences, Nobel Prize Outreach and Microsoft following the Nobel Prize Summit, Truth, Trust and Hope held in Washington, DC in May 2023. The views expressed in this article are the opinions of the authors and not necessarily the views of their affiliated organizations.

Author contributions

O.G., A.L., F.C.-M. and B.O.-G. conceived and wrote the article. F.C. contributed to the writing, revision and review. V.D., V.G., N.I., F.L., M.L. and K.T.G. contributed to the revision and review.

Competing interests

A.L. is employed by Microsoft. V.G. receives part-time funding from Google.org and has received funding from Google DeepMind. All other authors declare no competing interests.