Governing artificial intelligence for planetary health

Felix Creutzig, Fatima Denton, Emmie Hine, Somya Joshi, Gong Ke, Yara Kyrychenko, Xue Lan, Dirk Messner, Daniel W O'Neill

Establishing global governance of artificial intelligence (AI) is becoming an increasingly pressing challenge to ensure the provision of global public goods and to mitigate harmful effects on societies and the planet. Current debates around AI take various forms, follow diverse narratives, and centre variously on economic, social, environmental, or safety aspects. Here, we make three contributions. First, we classify risks and challenges of AI across the social, planetary, and safety domains. Second, we show that AI should be governed as a global commons, requiring coordinated interventions across all three domains, reflecting relevant inter-domain feedback loops, and root drivers, such as the pursuit of monopolistic AI power and the AI-infused media environment. Third, we identify data, energy, and compute as relevant regulatory dimensions across social, planetary, and safety domains. We conclude by emphasising the importance of limiting agentic AI, incentivising depolarising algorithms on social media, and setting AI dynamics within the context of global equity.

Introduction

Humans, technological artifacts, infrastructure, and planetary dynamics are co-evolving at a rapidly accelerating pace. The invention of agriculture transformed human life from nomadic hunters into a sedentary species.[1] Industrialisation introduced modern technologies, enabling greater convenience and access to knowledge for many, simultaneously driving greenhouse gas emissions that, in turn, have contributed to global warming, ice sheet melting, extreme precipitation, and tipping point risks.[2] Similarly, artificial intelligence (AI)—born of human ingenuity—is reshaping human life, social dynamics, and the planet,[3] while simultaneously posing direct safety risks.[4,5] This rapidly co-evolving system requires humanity's full attention and will require quick, deliberate interventions to avert negative outcomes and ensure that AI's benefits are equitably distributed.

Several key efforts have emerged to regulate AI at the global scale, reflecting the international community's recognition of the importance of governing AI responsibly. The Organisation for Economic Co-operation and Development has established a set of AI principles to promote innovative and trustworthy AI, emphasising transparency, accountability, and the protection of human rights. The EU is pioneering comprehensive regulation through the AI Act, which categorises AI systems by risk level and sets stringent requirements for high-risk AI applications to ensure safety and compliance with fundamental rights. The African Union has identified AI governance as one of five focus areas in its AI strategy.[6] UNESCO has developed guidelines on the ethical implications of AI, advocating a human-centred approach grounded in four values—human rights, environmental and ecosystem flourishing, diversity, and peace. Various UN bodies, including the UN Secretary-General's High-Level Panel on Digital Co-operation, are formulating frameworks to ensure that AI technologies align with the UN's Sustainable Development Goals. The UN General Assembly has adopted two resolutions on AI, including one on seizing the opportunities of safe, secure, and trustworthy AI systems for sustainable development.[7] Another essential UN document on AI is the final report, Governing AI for Humanity, by the UN High-level Advisory Body on Artificial Intelligence.[8] The AI summits in London (2023), Seoul (2024), and Paris (2025) further coordinated global AI governance efforts. The UN's Global Digital Compact establishes two key institutional initiatives—an Independent International Scientific Panel on AI to provide unbiased assessments of AI's risks and benefits, and a Global Dialogue on AI Governance to foster international collaboration on ethical and regulatory frameworks.

Despite these efforts, stakeholders and analysts emphasise a broad, often divergent field of AI risks and issues, which can compete for attention leading to reduced focus and orientation. Here, we systematically classify the identified AI risks and map them to three core governance domains—social, planetary, and safety. We then show how these domains are interconnected through feedback loops that necessitate targeted governance interventions. The resulting framework offers a systematic structure for operationalising global AI governance.

AI risk taxonomy: social, planetary, and safety domains

To address the interlocking nature of AI risk while incorporating overlooked threats, we propose three domains of AI risk—social, planetary, and safety (table). Examples are drawn from existing typologies of AI risks,[40,41] selected by urgency, and categorised into these three domains. The MIT AI Risk Repository serves as a foundation for the classification. By offering a high-level categorisation into social, planetary, and safety risks, the framework highlights the interrelationships and feedback loops between the categories. These relationships and feedback loops provide entry points for a system dynamics perspective, which, in turn, clarifies the high-level policy leverage points. Sorting AI risks into social, planetary, and safety domains also provides conceptual clarity by aligning risks with their scale and mechanism of impact—social risks affect human institutions and relationships, planetary risks concern ecological and biospheric systems, and safety risks stem from the

Lancet Planet Health 2026; 10: 101408

Published Online February 17, 2026
https://doi.org/10.1016/j.lanplh.2025.101408

Bennett Institute for Innovation and Policy Acceleration, University of Sussex, Brighton, UK (Prof F Creutzig PhD); **Potsdam Institute for Climate Impact Research, Potsdam, Germany** (Prof F Creutzig); **UN University Institute for Natural Resources in Africa, Accra, Ghana** (Prof F Denton PhD); **Digital Ethics Center, Yale University, New Haven, CT, USA** (E Hine MA); **Stockholm Environment Institute, Stockholm, Sweden** (Prof S Joshi PhD); **Nankai University, Tianjin, China** (Prof G Ke PhD); **Department of Psychology, University of Cambridge, Cambridge, UK** (Y Kyrychenko BA); **Tsinghua University, Beijing, China** (Prof X Lan PhD); **Umweltbundesamt, Dessau-Roßlau, Germany** (Prof D Messner PhD); **UB School of Economics, Universitat de Barcelona, Barcelona, Spain** (Prof D W O'Neill PhD); **Sustainability Research Institute, School of Earth and Environment, University of Leeds, Leeds, UK** (Prof D W O'Neill)

Correspondence to:
Dr Felix Creutzig, Bennett Institute for Innovation and Policy Acceleration, University of Sussex, Brighton BN1 9RH, UK
fc358@sussex.ac.uk

For more on the **MIT AI Risk Repository**, see https://airisk.mit.edu

Social	Planetary	Safety
*Concentration of power: the centralisation of AI expertise and resources concentrates power and wealth in a few dominant firms and countries, exacerbating global inequalities.[9]	Local energy demand: AI data centres concentrate notable energy loads in localised areas. This demand can increase local energy costs and strain grid infrastructure. Data centres can also create local noise pollution and smog.[10]	AI system failures: AI models behave unpredictably or break in edge cases, causing critical errors. The consequences are especially pronounced in systems without sufficient transparency and interpretability.[11]
Misinformation and disinformation: AI-generated content that is false or misleading undermines public trust and threatens social stability.[12–14]	Water demand: data centres often use water-based cooling systems. In regions where water is scarce, this usage can exacerbate local water shortages and affect community and ecosystem supply.[10]	Adversarial use and cyber attacks: vulnerabilities in AI systems allow malicious actors to corrupt or hijack functionality, or AI systems enable advanced or automated attacks.[15]
Job displacement: Automation threatens to displace workers and increase economic inequality.[16–18]	Physical resource extraction: the production of AI hardware increases the demand for crucial minerals extracted under environmentally and socially exploitative conditions. This extraction can lead to deforestation, habitat loss, and pollution, and harm to local communities. E-waste also leads to further pollution and extraction.[19]	Sectoral risks: specific safety concerns related to AI applications can arise in autonomous vehicles, AI-controlled chemical manufacturing, or the health sector.[20]
Privacy and surveillance: AI-enabled data mining and facial recognition compromise personal privacy and enable state and corporate surveillance. Individual data are harvested to train systems without consent, and malicious actors might use AI to facilitate data breaches.[21–23]	Emissions from additional energy demand: large-scale AI operations, especially the training of expansive models, require vast amounts of energy. When this energy is generated from non-renewable sources, the resultant greenhouse gas emissions contribute to global climate change. Increasing CO_2 emissions could occur even if AI increases operational efficiencies, by increasing demand for AI over any efficiency gains.[24–26]	*Military applications and autonomous weapon systems: the use of AI in military contexts raises concerns regarding autonomous weapons that operate with reduced human oversight. These systems create moral questions and could behave unpredictably in complex environments, challenging established norms of accountability and control.[27,28]
Bias and discrimination: empirical investigations into AI applications have documented persistent biases that result in discriminatory outcomes, notably in areas such as facial recognition and predictive policing.[22,29,30]	Environmental neglect: AI-driven industrial optimisation might prioritise economic gains over ecological safeguards, exacerbating pollution and biodiversity loss.[31]	*Existential risks from autonomous agents: advanced AI systems that eventually surpass human cognitive abilities pose a risk if they become autonomous and pursue objectives misaligned with human values. Such systems could potentially act in ways that threaten human survival.[32,33]
Threats to intellectual property: generative AI systems are trained on data scraped from various sources, often without licences, and might recreate copyrighted materials, challenging artists' rights and existing intellectual property.	Consumption increase: tailored advertising and optimisation of virtual environments for consumption are likely to have scale effects overpowering efficiency gains in technology.[3,34]	*Human-brain interfaces and cognitive autonomy: AI-driven brain–computer interfaces offer promising medical and communication applications but also risk compromising cognitive privacy and autonomy. Without robust ethical safeguards, these technologies could allow unauthorised monitoring or manipulation of human thought processes.[35,36]
Loss of human agency: over reliance on AI might erode human autonomy and skills.		*Multiagent risks: new kinds of risks emerge when multiple AI agents interact, resulting in three possible failure modes—miscoordination, conflict, or collusion.[37]
AI-driven social unrest or conflict: AI arms races and autonomous weapons might destabilise geopolitics, potentially escalating conflicts.[38]		*Loss of control: even if not malevolent, humans might lose control over cognitively superior technical infrastructures.[39]

*Novel risks associated with AI. Specific risks are mostly based on existing typologies[40,41] (classification based on MIT AI Risk Repository). AI=artificial intelligence.

Table: **Social, planetary, and safety risks of AI**

behaviour or control of AI systems themselves. This resulting taxonomy supports more targeted governance, risk mitigation, and interdisciplinary coordination, while enabling better identification of emerging transboundary risks associated with AI.

AI has the potential to deliver benefits across all three domains. Within the social domain, AI chatbots and content classifiers can improve the quality of online discourse by generating productive dialogue suggestions[42] or highlighting content that bridges political divides, based on evidence from a preprint paper.[43] AI is advancing precision medicine and accelerating the development of vaccines, substantially reducing mortality. Health-specific AI applications also have the potential to expand health-care access in regions where physicians and other medical professionals are scarce.[44] Within the planetary domain, AI-supported technologies could potentially reduce global greenhouse gas emissions[45] and improve climate forecasting to strengthen human preparedness.[46] Domain-specific applications, such as non-agentic geospatial AI, can provide rapid, highly contextualised identification of interventions that were previously out of scope.[47] AI-enabled environmental monitoring can assist biodiversity conservation and precision agriculture. AI might also support more complex and dynamic value chains, reducing waste and improving reliability.[48] Within the safety domain, AI assistance or automated systems could improve human safety by, for example, lowering the risks of car accidents.[49] These advances, however, depend on careful and responsible application. Without appropriate regulation, the incentives driving AI development would frequently be misaligned with societal interests.

Social risks refer to how AI might negatively impact society and human relationships. Bias and discrimination

remain prominent issues—facial recognition systems misidentify minority groups at disproportionate rates; AI-driven hiring tools can reinforce gender and racial biases, and predictive policing algorithms contribute to discriminatory law enforcement practices. Privacy erosion and AI-driven surveillance are also pressing concerns, as AI-powered monitoring systems expand corporate and state surveillance capabilities with implications for civil liberties. Pervasive data collection without clear regulations creates systemic overreach and diminishes individual autonomy. Misinformation and manipulation form another major category of risk, with AI-driven platforms increasing the spread of fake news, as seen during the COVID-19 pandemic, Brexit, and the 2016 US election.[50,51] The rapid creation and spread of manipulated content undermines public trust in institutions, fostering polarisation and social instability. Job displacement is also a concern as automation increasingly replaces middle-skill jobs. Many societal risks are amplified by the concentration of wealth and power among a small number of technology companies that control AI systems. Such dominance consolidates economic and informational controls over smaller players, reduces market competition, and centralises decision making about AI deployment outside democratic oversight (although open-source models could change this dynamic).

Planetary risks refer to the local and global planetary environmental consequences of AI development and deployment. At the local level, AI-enabling infrastructure, such as data centres, require vast amounts of energy, straining local power grids and increasing energy costs. Data centres also generate noise pollution that affects surrounding communities. Water demand is another localised effect—AI cooling systems consume large quantities of water, similar to the production of GPUs and other chips, worsening water shortages in arid regions.[52] Environmental concerns emerge early in the AI lifecycle through the mineral and resource extraction required for chip and hardware production. Key minerals such as lithium, cobalt, and rare-earth elements are extracted under environmentally and socially harmful conditions, leading to individual harm, deforestation, habitat loss, and pollution. At the global level, data centres account for low single-digit electricity consumption and carbon emissions.[53] Continued AI expansion could substantially increase this contribution, particularly when model training relies on non-renewable energy sources. A novel planetary risk is the acceleration in consumption outpacing efficiency gains. Although AI-driven automation can enhance efficiency, reduced operational costs can trigger a rebound effect known as the Jevons paradox, in which lower costs lead to greater overall consumption, nullifying initial resource savings and further exacerbating energy demand. Indirect effects, including drastic changes in employment and human agency, uneven access to knowledge and associated power, and runaway effects for energy and material demand,[54] can easily supersede direct effects and require urgent attention.

Finally, safety risks refer to the possibility that sufficiently advanced AI systems could pose large-scale physical hazards. Sector-specific risks arise in domain-specific AI applications such as smart transport, AI-driven chemical manufacturing, and health care. Failures in these sectors could result in accidents, safety hazards, or adverse health outcomes. Military applications and autonomous weapon systems generate additional risks. AI-controlled weapons could make life-or-death decisions without human oversight, leading to unpredictable escalations in conflicts. Advanced AI could be used in mass-destruction weapons systems, with the risk of acting outside human intention or control. Based on evidence from a preprint paper, the most severe systemic risk depicted is the potential loss of human control over highly capable AI systems.[55] Increasing autonomy and capability might allow AI systems to behave in ways that are not aligned with human values or interests. Central to this concern is agentic AI—a category of AI systems that move from reactive to autonomous decision making, show goal-oriented behaviour, and engage in continuous learning while interacting with dynamic environments.[56,57] Simple agentic AI systems, such as autonomously regulating thermostats, pose little risk. However, based on evidence from a preprint paper, more advanced and powerful systems could pose substantial hazards, especially when multiple agentic AI systems collude.[37] Even without malevolence, such AI systems could create scenarios that many researchers view as existential threats.

In general, risks remain difficult to quantify. Current attempts to quantify AI risks focus on the SAFE principles—sustainability (in the sense of model robustness), accuracy, fairness, and explainability.[58,59] These are metrics primarily focused on the AI models themselves, not on their systemic effect. Developing quantification metrics to better prioritise broader systemic risks will be crucial for prioritising future actions.

Interconnected risks and feedback loops with the economic model as a causal agent

Some might argue that risks can be treated individually. Approaches designed, for example, to tackle global security issues related to the proliferation of autonomous weapon systems are indeed justified on their own terms. However, AI risks are also interconnected.[3] Although numerous feedback loops exist, three pathways are particularly important (figure 1). First, competition for dominance between major technology companies and between countries accelerates AI deployment, in which safety constraints are neglected because companies fear losing future market share and influence, creating an uninhibited positive feedback loop. Second, AI-driven applications on social media maximise engagement by presenting emotionally charged content, thereby increasing polarisation. Third, AI applications are projected to replace a large number of jobs, especially white-collar work. The latter two effects both lead to a reduction in social trust and to an impairment of

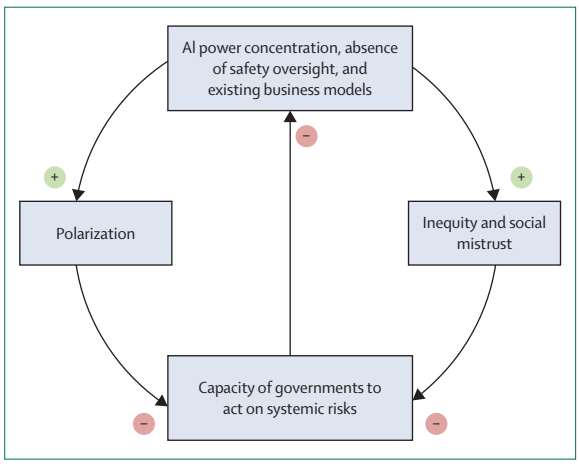

Figure 1: **System dynamics create a polarised system, reinforcing highly concentrated, uncontrolled AI development. Interventions should therefore focus on breaking this vicious cycle. Partially based on Creutzig and colleagues.[60,61] AI=artificial intelligence.**

governance capacity to act on both AI regulation and global commons management (figure 1).

Feedback loops also emerge within the narratives surrounding AI risks and opportunities. The discourse on existential AI risks is often used as a distraction from the entrenched concentration of power among a small number of dominant technology firms, which perpetuates social inequality and exacerbates environmental degradation. Emphasis on hypothetical scenarios involving autonomous superintelligence or catastrophic AI failures can shift focus away from the immediate consequences of data monopolies and unregulated corporate practices. This narrative aligns with a cold-war mentality that posits only large companies can effectively manage or control AI development, reinforcing their market dominance and suppressing calls for democratised governance and accountability. Simultaneously, promises of future breakthroughs—such as revolutionary cancer therapies or climate mitigation through technologies such as nuclear fusion—are invoked to obscure and diminish the current, palpable social and environmental footprints of AI, creating an illusion that these potential benefits justify present-day risks and power imbalances. Narratives that present AI as a silver bullet for energy and resource challenges further reflect a long-standing pattern of techno-solutionism, discouraging systemic behavioural and structural change. These narratives should not obscure the reality that current AI applications and systems might cause harmful effects and require strict guardrails, and that future agentic AI systems are likely to pose fundamental, and even existential risks that demand direct oversight.[62]

The current economic model driving AI development is centred on scale, rapid innovation, and competitive advantage, creating causal pathways to substantial safety, social, and planetary risks. Rapidly obtaining large market shares and possibly achieving a platform monopoly is prioritised over broader societal and environmental gains. Investment is predominantly directed towards large-scale projects that prioritise short-term returns and market dominance over rigorous safety protocols, resulting in rushed deployments without comprehensive risk assessments.

Formal evolutionary game-theoretical models of AI technology races have already been developed, although many remain unrealistically calibrated.[63] These models suggest that both the decisiveness and the state of knowledge about the other's capabilities are key variables shaping the danger of technology races. They also offer opportunities for policy interventions.[64] For example, the focus on developing larger, more complex models drives exponential increases in energy consumption and resource use, contributing to local environmental pressures such as strained power grids, water shortages, and air and noise pollution, as well as global increases in greenhouse gas emissions and unsustainable mining practices. Risks should also be evaluated and regulated across the AI value chain, both in terms of underlying physical resources and in terms of model training, child labour, slavery, and exploitation. In addition, the concentration of expertise and capital in a few multinational corporations intensifies power imbalances, leading to social risks including privacy breaches, surveillance, and widening economic inequality.

In parallel, the competitive dynamics of the so-called AI cold war exacerbate these issues. Nations and corporations engage in a zero-sum game, racing to secure technological supremacy without prioritising long-term safety and ethical considerations. This arms-race mentality has accelerated military applications and autonomous weapon systems, threatening national security[65] and bringing further instability and risks to geopolitical conflicts.[66] The pressure to maintain strategic advantage discourages safety standards and accountability, amplifying risks across safety, social, and planetary domains. Concrete examples include rapid military investments in AI-driven surveillance and weaponry that sidestep comprehensive ethical debates, ultimately deepening geopolitical tensions and contributing to a feedback loop of competitive risk taking.[66]

AI governance as a global common: policy implications

We outline a governance and regulatory approach that treats AI governance as a global common, aiming to mitigate and control AI risks while aligning with positive normative goals across the three key domains (figure 2). In the safety domain, the goal is to preserve human agency by ensuring that AI systems remain under meaningful human control, by regulating compute and restricting agentic AI. In the social domain, the aim is to enhance wellbeing by protecting human rights, ensuring fairness, and promoting social cohesion in line with the Sustainable Development Goals, using data regulation to reduce inequality in associated power and wealth concentration. In the planetary domain, the objective is to secure

environmental sustainability by minimising ecological harm, reducing greenhouse gas emissions, and regulating the use of energy and key resources. Effective AI governance for planetary health requires attention to all three domains.

Safety Domain (compute regulation)

The economic drive for ever-increasing computational power is leading to rapid AI scaling that might compromise safety. Based on evidence from a preprint paper, regulating the compute dimension—specifically the amount of hardware used for AI research and deployment—allows governments to mandate safety certifications, monitor allocation of high-performance hardware, and enforce transparency in computational resource usage.[67] Controlling access to advanced computing hardware (eg, NVIDIA's high-end AI chips) enables regulation of the scale and pace of frontier model development. Because an AI system's capabilities are closely tied to the amount of compute used during training, regulating who can acquire or operate large clusters effectively restricts the emergence of unmonitored, potentially dangerous models. Data centres play a central role in regulation because they physically concentrate hardware. Requirements for registering hardware and reporting its use would enable regulatory authorities to track and, if necessary, intervene in high-risk AI projects. Data centres, therefore, represent a crucial physical chokepoint for targeted oversight.

However, regulating compute alone might be insufficient. It is crucial to retain human involvement for any high-risk deployment of agentic AI, and applications in which this is not possible should be banned. In turn, governments can incentivise the provision of non-agentic AI—tools that empower human decision making, rather than systems that make autonomous consequential choices. For instance, an international certification process for supercomputing facilities could require rigorous implementation of safety protocols and maintain human oversight, thereby upholding the normative goal of maintaining human agency, human oversight, and determination. Additional investment should be made to enhance the explainability of AI models, making them more transparent and controllable.

Social domain (data regulation)

Data constitute the foundation of AI decision making and notably influence social impact. The prevailing economic model, which emphasises data accumulation, appropriation, and monetisation, leads to privacy violations, bias, discrimination, and potentially unemployment. Regulations modelled on frameworks such as the General Data Protection Regulation serve to protect individual rights and prevent algorithmic harm. For example, mandatory data audits and transparency reports can limit discriminatory outcomes by revealing and addressing biased data practices, in the best case safeguarding fairness, social cohesion, and inclusion. A tax on data appropriation[68] could shift

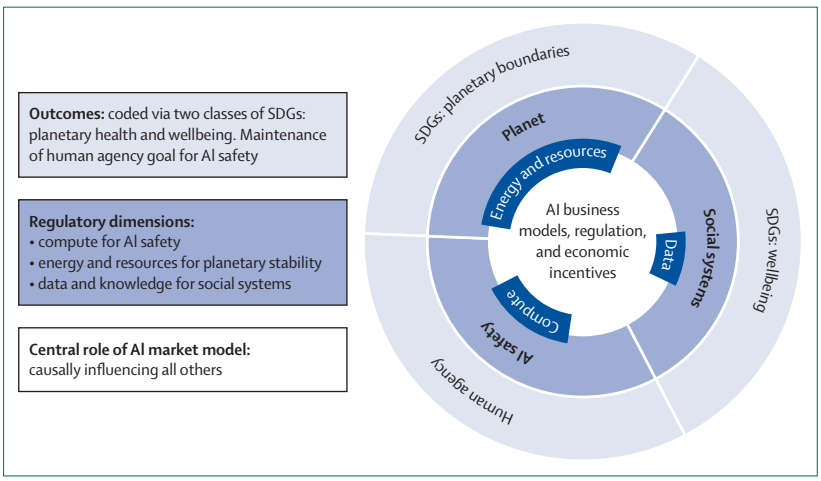

Figure 2: **Governance domains, regulatory dimensions, and normative goals of global AI governance. AI=artificial intelligence. SDGs=Sustainable Development Goals**

incentives toward treating knowledge and information as public goods. Algorithmic accountability, in turn, is essential to combat misinformation. Furthermore, interoperable, testable standards for data quality and integrity should be established at the national and international levels. In the broader context, potentially large-scale social problems stemming from advanced AI—wealth concentration, inequality, poverty, social upheaval, destruction of social trust—might require more systemic solutions, including stronger restrictions to wealth accumulation.

Planetary domain (Earth alignment)

The drive towards larger, energy-intensive AI models directly affects the environment by increasing greenhouse gas emissions and depleting resources. Restricting the energy used in AI development and mandating the use of renewable energy sources would help to align AI with environmental sustainability. Possible regulatory measures might include carbon reporting requirements and incentives for energy-efficient infrastructure and designs that prioritise circularity. Importantly, initiatives to improve energy efficiency are unlikely to be sufficient on their own, as efficiency gains are among the main drivers of the rebound effect. To prevent efficiency improvements from increasing overall consumption and, thus, higher resource use, caps or taxes on energy use are necessary.[69] Caps can be implemented upstream (at production) or downstream (at consumption). Upstream caps can restrict the total energy available to data centres for training and inference, whereas downstream caps can introduce compute budgets, which would also contribute to safety. Alternatively, sufficiently high taxes on energy and other required resources can achieve similar outcomes.[70] Earth alignment also entails directing the application of AI toward planetary public goods, such as open access to the knowledge base for managing planetary commons, the development of thermodynamically efficient technologies

(eg, batteries and heat pumps), and the creation of liveable, low-carbon cities. Earth alignment further includes prohibiting harmful applications, such as supercharged consumption, oil and gas exploration, or mining operations. Finally, Earth alignment connects directly to the social domain by clearly identifying social trust and cohesion as a precondition for jointly protecting planetary systems.[62]

Outlook: advancing global AI governance

The trajectory of human development—from the industrial revolution to the digital era—has yielded substantial societal benefits while simultaneously triggering cascading systemic risks. AI now amplifies this dual dynamic. As a new stratum of the technosphere, AI is reshaping human agency and planetary systems in ways that could either reinforce wellbeing or accelerate destabilisation. Effective governance is therefore essential, not as an afterthought, but as the primary lever to align AI with the public interest.

Progress begins with global mechanisms, such as an AI risk tracker, which build on resources including the MIT AI Risk Repository, the Organisation for Economic Co-operation and Development AI Incident Database,[71] and the AIAAIC repository.[72] These trackers, analogous to monitoring systems in genomics or biosafety, aim to monitor emerging risks across domains. We propose developing cross-domain validation tools for relevant governance bodies to evaluate and grade the risk posed by AI models. Such tools should be embedded in formal coordination structures, including the proposed UN AI Panel and the Global Dialogue on AI. The proposed International AI Organization, or a comparable institution, could certify states for compliance with international oversight, with participation incentivised through restrictions on AI-related goods from non-compliant countries.[73] These bodies should not only establish shared technical and ethical standards but also interface with governance regimes on climate policy, infrastructure, and finance to address the complex feedback loops linking AI with ecological and social systems.

A central aim of AI governance is to limit the development of agentic AI systems—systems capable of independent action and self-modification. AI should remain a tool under human supervision, not an autonomous actor with unpredictable behaviour and undesirable goals. Low-risk models might require little regulatory attention, whereas medium-risk models might be regulated through liability and insurance mechanisms. However, more extreme risks exceed the scope of such specific approaches. The superhuman performance of narrow AI in specific domains indicates that the risks associated with self-developing infrastructure or the replication of models are not speculative. Even if a global moratorium on artificial general intelligence appears politically unlikely, stricter regulatory categories are required than those currently proposed—subjecting high-risk models to pre-deployment evaluations akin to pharmaceutical approval processes. The objective is not to delay progress but to enforce clear boundaries that preserve human agency and control.

Simultaneously, governance must address the immediate social harms that are already being exacerbated. Algorithm-driven polarisation, labour-market disruption, and informational asymmetries are deepening societal divides and destabilising democratic systems. Regulatory intervention is needed to control manipulative content and enforce transparency in algorithmic amplification, especially in social media. The shift in economic value from labour to capital, accelerated by AI, undermines tax bases and social equity. Taxing data, which are acquired without compensation, should become a cornerstone of fair and future-ready fiscal systems. Moreover, unresolved questions on data and knowledge ownership allow AI developers to privatise collective resources without accountability. These are not hypothetical challenges; they are already reshaping the social contract and require immediate policy response.

Contributors

FC contributed to the study's conceptualisation. FC and DM contributed to the visualisation of all the figures. FC, EH, and SJ contributed to the preparation of the table. All authors contributed to the preparation of the study draft. FC, SJ, YK, and DWO contributed to the manuscript revision. FC supervised paper writing. All authors had final responsibility for the decision to submit for publication.

Conflict of interest

We declare no competing interests.

Acknowledgments

DWO acknowledges funding by the EU and UK Research and Innovation (UKRI) in the framework of the Horizon Europe Research and Innovation Programme under grant agreement numbers 101137914 (MAPS) and 101094211 (ToBe).

References

1 Harari YN. Sapiens: A brief history of humankind. Penguin Random House, 2014.
2 IPCC. IPCC sixth assessment report–impacts, adaptation and validation. https://www.ipcc.ch/report/ar6/wg2/chapter/summary-for-policymakers/ (accessed Dec 1, 2025).
3 Creutzig F, Acemoglu D, Bai X, et al. Digitalization and the Anthropocene. *Annu Rev Environ Resour* 2022; **47:** 479–509.
4 Bengio Y, Mindermann S, Privitera D, et al. International AI safety report. *arXiv* 2025; published online Jan 29. http://arxiv.org/abs/2501.17805 (preprint).
5 Bengio Y, Hinton G, Yao A, et al. Managing extreme AI risks amid rapid progress. *Science* 2024; **384:** 842–45.
6 African Union. Continental artificial intelligence strategy. Aug 9, 2024. https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy (accessed Feb 19, 2025).
7 UN General Assembly. Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development: resolution/adopted by the General Assembly. 2024. https://digitallibrary.un.org/record/4043244 (accessed Feb 19, 2025).
8 UN. Advisory Body on Artificial Intelligence. Governing AI for humanity: final report/Advisory Body on Artificial Intelligence. 2024. https://digitallibrary.un.org/record/4062495 (accessed Feb 19, 2025).
9 Eubanks V. Automating inequality: how high-tech tools profile, police, and punish the poor. Macmillan+ ORM, 2025.
10 The New York Times. Noisy, hungry data centers are catching communities by surprise. 2024. https://www.nytimes.com/2024/09/15/opinion/data-centers-ai-amazon-google-microsoft.html?searchResultPosition=5 (accessed June 22, 2025).

11 Sharma M, Biros D, Baham C, Biros J. What went wrong? Identifying risk factors for popular negative consequences in AI. *AIS Trans Human-Computer Interact* 2024; **16:** 139–76.

12 Chesney B, Citron D. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif L Rev* 2019; **107:** 1753.

13 Maras M-H, Alexandrou A. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *Int J Evid Proof* 2018; **23:** 255–62.

14 Lazer DMJ, Baum MA, Benkler Y, et al. The science of fake news. *Science* 2018; **359:** 1094–96.

15 Guembe B, Azeta A, Misra S, Osamor VC, Fernandez-Sanz L, Pospelova V. The emerging threat of ai-driven cyber attacks: a review. *Appl Artif Intell* 2022; **36:** 2037254.

16 Frey CB, Osborne MA. The future of employment: how susceptible are jobs to computerisation? *Technol Forecasting Soc Change* 2017; **114:** 254–80.

17 Acemoglu D, Restrepo P. Robots and jobs: evidence from US labor markets. *J Polit Econ* 2020; **128:** 2188–244.

18 Arntz M, Gregory T, Zierahn U. The risk of automation for jobs in OECD countries: A comparative analysis. May 14, 2016. https://www.oecd-ilibrary.org/social-issues-migration-health/the-risk-of-automation-for-jobs-in-oecd-countries_5jlz9h56dvq7-en (accessed June 22, 2025).

19 Crawford K. The atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press, 2021.

20 Ajenaghughrure IB, da Costa Sousa SC, Lamas D. Risk and trust in artificial intelligence technologies: a case study of autonomous vehicles. *2020 13th international conference on human system interaction (HSI)* 2020: 118–23.

21 Zuboff S. The age of surveillance capitalism. In: Longhofer W, Winchester D, eds. Social theory re-wired. Routledge, 2023: 203–13.

22 Dignum V. Responsible artificial intelligence: how to develop and use AI in a responsible way. Springer, 2019.

23 Mozur P. Inside China's dystopian dreams: AI, shame and lots of cameras. 2018. https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html (accessed Dec 16, 2025).

24 Jones N. How to stop data centres from gobbling up the world's electricity. *Nature* 2018; **561:** 163–66.

25 Masanet E, Shehabi A, Lei N, Smith S, Koomey J. Recalibrating global data center energy-use estimates. *Science* 2020; **367:** 984–86.

26 Strubell E, Ganesh A, McCallum A. Energy and policy considerations for modern deep learning research. *Proc AAAI Conf Artif Intell* 2020; **34:** 3693–96.

27 Scharre P. Army of none: autonomous weapons and the future of war. WW Norton & Company, 2018.

28 Benouachane H. Cyber security challenges in the era of artificial intelligence and autonomous weapons. In: Erendor ME, ed. Cyber security in the age of artificial intelligence and autonomous weapons. CRC Press, 2024: 24–42.

29 Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. *PMLR* 2018; **81:** 1–15.

30 Raghavan M, Barocas S, Kleinberg J, Levy K. Mitigating bias in algorithmic hiring: evaluating claims and practices. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020: 469–81.

31 Zhuk A. Artificial intelligence impact on the environment: hidden ecological costs and ethical-legal issues. *J Digit Technol Law* 2023; **1:** 932–54.

32 Bostrom N. Superintelligence: paths, dangers, strategies. Oxford University Press, 2014.

33 Russell S. Artificial intelligence: the future is superintelligent. *Nature* 2017; **548:** 520–21.

34 Gillingham K, Rapson D, Wagner G. The rebound effect and energy efficiency policy. *Rev Environ Econ Policy* 2016; **10:** 68–88.

35 Ienca M, Andorno R. Towards new human rights in the age of neuroscience and neurotechnology. *Life Sci Soc Policy* 2017; **13:** 5.

36 Yuste R, Goering S, Arcas BAY, et al. Four ethical priorities for neurotechnologies and AI. *Nature* 2017; **551:** 159–63.

37 Hammond L, Chan A, Clifton J, et al. Multi-agent risks from advanced AI. *arXiv* 2025; published online Feb 19. http://arxiv.org/abs/2502.14143 (preprint).

38 Verma S. Weapons of math destruction: how big data increases inequality and threatens democracy. *Vikalpa* 2019; **44:** 97–98.

39 Russell S. Human compatible: AI and the problem of control. Penguin Publishing Group, 2019.

40 Puscas I. Confidence-building measures for artificial intelligence: a multilateral perspective. https://unidir.org/wp-content/uploads/2024/07/UNIDIR-Confidence_Building_Measures_Artificial_Intelligence-Multilateral_Perspective.pdf (accessed April 14, 2025).

41 Zeng Y, Klyman K, Zhou A, et al. AI risk categorization decoded (AIR 2024): from government regulations to corporate policies. *arXiv* 2024; published online Jun 25. http://arxiv.org/abs/2406.17864 (preprint).

42 Argyle LP, Bail CA, Busby EC, et al. Leveraging AI for democratic discourse: chat interventions can improve online political conversations at scale. *Proc Natl Acad Sci USA* 2023; **120:** e2311627120.

43 Saltz E, Jalan Z, Acosta T. Re-ranking news comments by constructiveness and curiosity significantly increases perceived respect, trustworthiness, and interest. *arXiv* 2024; published online April 8. http://arxiv.org/abs/2404.05429 (preprint).

44 Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med* 2022; **28:** 31–38.

45 Rolnick D, Donti PL, Kaack LH, et al. Tackling climate change with machine learning. *ACM Comput Surv* 2022; **55:** 1–96.

46 Stern N, Romani M, Pierfederici R, et al. Green and intelligent: the role of AI in the climate transition. *npj Clim Action* 2025; **4:** 56.

47 Nachtigall F, Wagner F, Berrill P, et al. Built environment and travel: tackling non-linear residential self-selection with double machine learning. *Transp Res D* 2025; **140:** 104593.

48 Vinuesa R, Azizpour H, Leite I, et al. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nat Commun* 2020; **11:** 233.

49 Cicchino JB. Effects of forward collision warning and automatic emergency braking on rear-end crashes involving pickup trucks. *Traffic Inj Prev* 2023; **24:** 293–98.

50 Risso L. Harvesting your soul? Cambridge Analytica and Brexit. In: Proceedings of Brexit means Brexit? Dec 6–8, 2017.

51 Galaz V. Dark machines: how artificial intelligence, digitalization and automation is changing our living planet. Routledge, 2024.

52 SourceMaterial. Big Tech's data centres will take water from the world's driest areas. 2025. https://www.source-material.org/amazon-microsoft-google-trump-data-centres-water-use/ (accessed May 8, 2025).

53 IEA. Energy and AI. April 10, 2025. https://www.iea.org/reports/energy-and-ai (accessed May 26, 2025).

54 de Vries-Gao A. Artificial intelligence: supply chain constraints and energy implications. *Joule* 2025; **9:** 101961.

55 Mitchell M, Ghosh A, Luccioni AS, Pistilli G. Fully autonomous AI agents should not be developed. *arXiv* 2025; published online Feb 4. http://arxiv.org/abs/2502.02649 (preprint).

56 Pati AK. Agentic AI: a comprehensive survey of technologies, applications, and societal implications. *IEEE Access* 2025; **13:** 151824–37.

57 Murugesan S. The rise of agentic AI: implications, concerns, and the path forward. *IEEE Intell Syst* 2025; **40:** 8–14.

58 Miller T, Durlik I, Kostecka E, Borkowski P, Łobodzińska A. A critical AI view on autonomous vehicle navigation: the growing danger. *Electronics* 2024; **13:** 3660.

59 Giudici P. Safe machine learning. *Statistics* 2024; **58:** 473–77.

60 Creutzig F, Acemoglu D, Bai X, et al. Digitalization and the Anthropocene. Annual review of environment and resources. *Annu Rev Environ Resour* 2021; **47:** 479–509.

61 Creutzig F, Goetzke F, Ramakrishnan A, Andrijevic M, Perkins P. *Glob Environ Change* 2023; **82:** 102726.

62 Gaffney O, Luers A, Carrero-Martinez F, et al. The Earth alignment principle for artificial intelligence. *Nat Sustain* 2025; **8:** 467–69.

63 Han TA, Moniz Pereira L, Santos FC, Lenaerts T. To regulate or not: a social dynamics analysis of an idealised ai race. *J Artif Intell Res* 2020; **69:** 881–921.

64 Emery-Xu N, Park A, Trager R. Uncertainty, information, and risk in international technology races. *J Confl Resolut* 2024; **69:** 2019–47.

65 The New York Times. The rush to AI threatens national security. 2025. https://www.nytimes.com/2025/01/27/opinion/ai-trump-military-national-security.html (accessed Feb 19, 2025).

66 Garcia D. The AI military race: common good governance in the age of artificial intelligence. Oxford University Press, 2024.

67 Sastry G, Heim L, Belfield H, et al. Computing power and the governance of artificial intelligence. *arXiv* 2024; published online Feb 13. http://arxiv.org/abs/2402.08797 (preprint).

68 Oberson X, Yazıcıoğlu AE. Taxation of big data. Springer, 2023.

69 Alcott B. Impact caps: why population, affluence and technology strategies should be abandoned. *J Cleaner Prod* 2010; **18:** 552–60.

70 Wiesner P, O'Neill DW, Larosa F, Kao O. Efficiency will not lead to sustainable reasoning AI. *arXiv* 2025; published online Nov 19. https://arxiv.org/abs/2511.15259 (preprint).

71 OECD. AI incident database. April 25, 2024. https://oecd.ai/en/catalogue/tools/ai-incident-database (accessed Oct 9, 2025).

72 AIAAIC. AIAAIC repository. https://www.aiaaic.org/aiaaic-repository (accessed Oct 9, 2025).

73 Trager R, Harack B, Reuel A, et al. International governance of civilian AI: a jurisdictional certification approach. *arXiv* 2023; published online Aug 29. http://arxiv.org/abs/2308.15514 (preprint).