PERSPECTIVE

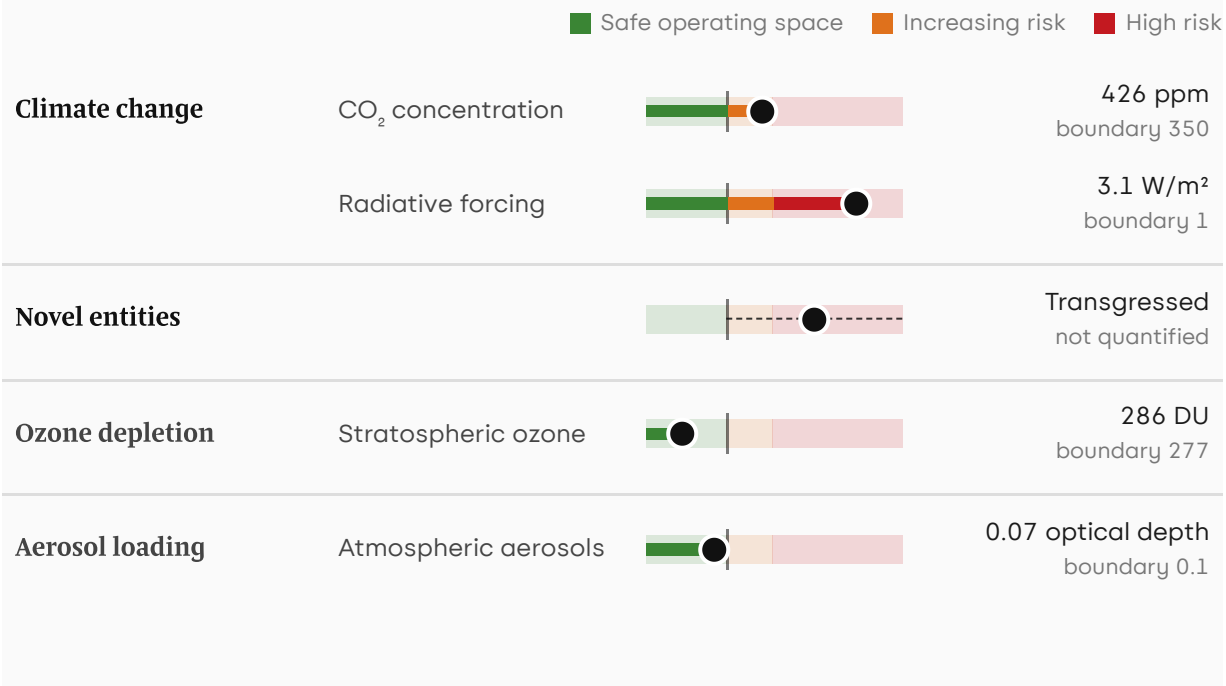
The **Anthropocene Alignment Problem**

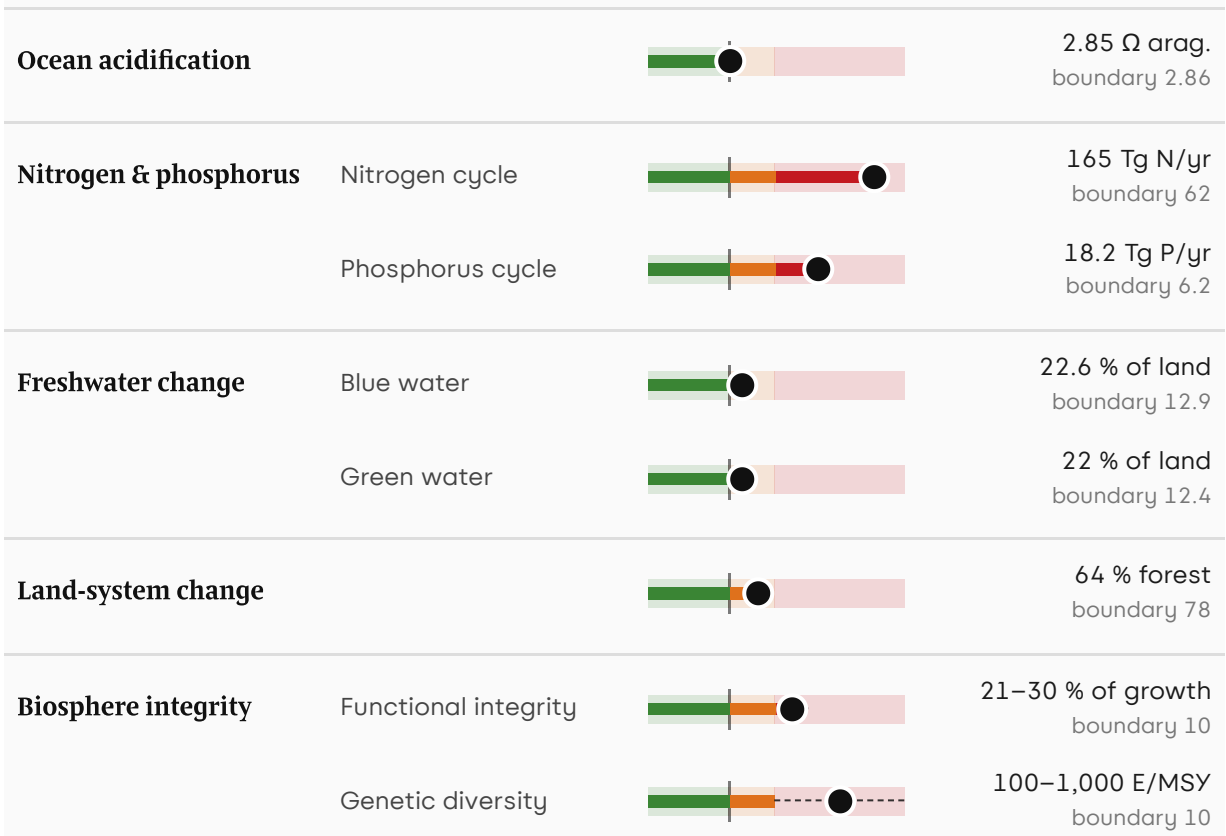
Planetary boundaries as constitutional constraints for **artificial intelligence**

Félix Pharand-Deschênes · GLOBAÏA · November 2026 · 24 min read · v1.0

ABSTRACT Two defining challenges of the twenty-first century, the destabilization of the Earth system and the rapid advancement of artificial intelligence, are converging in ways that remain critically undertheorized. Seven of nine planetary boundaries have now been transgressed (up from six in 2023 [1], with ocean acidification first assessed as transgressed in 2025) and all seven show worsening trends (Fig. 1); the planet's overall condition now lies near the upper end of the zone of increasing risk, close to the high-risk zone, and the land carbon sink is showing signs of weakening, raising the prospect that ecosystems could increasingly become sources of carbon [2]. Climate tipping points are closer than previously assessed [3], and over half of global land area has suffered critical losses in functional biosphere integrity [4]. Simultaneously, frontier AI systems are approaching transformative capabilities [5,6], with projected data-centre investment needs of $5–7 trillion by 2030 [7]. Yet the governance and research communities addressing these two domains have until recently remained largely siloed.

FIGURE 1 **Seven of nine planetary boundaries transgressed**





Each bar is one control variable of the planetary boundaries, rescaled so that the boundary sits at the same place on every row and the start of the high-risk zone at another; beyond it the scale is logarithmic. Units: DU, Dobson units; Ω arag., aragonite saturation state. Dashed lines span readings known only as a range, with the marker at the midpoint: novel entities, judged transgressed but never quantified; functional integrity, the share of plant growth taken by humanity (21–30%); and genetic diversity, an extinction rate of 100–1,000 per million species-years (E/MSY) against a boundary of 10. Data: Richardson et al. 2023 and the Planetary Health Check 2026[1,2].

Both trends are now measurably accelerating. Once the influence of El Niño, volcanic eruptions and solar variation is removed from the temperature record, global warming shows a statistically significant increase in its rate from about 2014: between 0.34 and 0.42 °C per decade across five datasets, against about 0.2 °C per decade since the 1970s, a pace that would carry the long-term trend across 1.5 °C before 2030[8]. On the other side, the first independent assessment carried out inside a frontier developer found in November 2026 that the company was likely improving its models' capabilities on well-specified tasks at around twice its earlier rate (anywhere from 1.3 to 3 times), judged that most of the acceleration probably came from AI's own contribution to AI research rather than from more training compute, and saw no evidence of a bottleneck within the next twelve months[9]. The second measurement is far less certain than the first: one company, benchmark scores, and evidence that was largely redacted. But the

two findings have the same form, a rate that was steady and no longer is. This Perspective argues that these are not parallel trends but coupled ones, on a collision course.

A growing 'AI for Sustainability Science' agenda has begun to emerge[10], demonstrating real capabilities: AI systems already improve climate projections, monitor biodiversity loss in near-real time, power early warning systems for extreme weather, and optimize renewable energy grids. The Earth alignment principle has established that AI development should promote planetary stability and respect safe Earth system boundaries[11]. These are genuine advances, and no serious analysis can dismiss them. But application-level benefits operate within a system whose aggregate dynamics (material footprint, rebound effects, coupled amplification) remain largely unaccounted for. The question is not whether AI can help address environmental challenges, but whether these benefits can outweigh system-level amplification in the absence of hard constraints.

Here we argue that Earth system science and AI safety research share deep structural parallels in their dynamics, failure modes, and irreversibility risks, and that realizing the Earth alignment principle requires a constitutional constraint framework — embedding planetary boundaries as inviolable conditions that bound the space within which all AI optimization may occur — together with a coupled-system analysis of the cascade dynamics that make such constraints urgent.

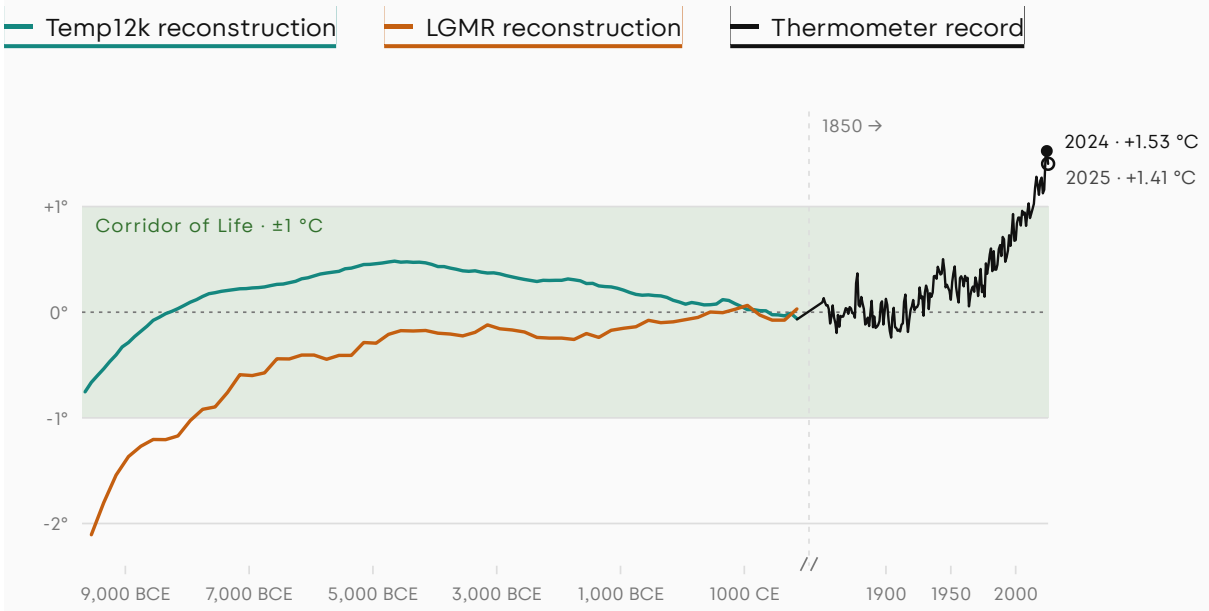
1 The Holocene as training distribution

The AI alignment problem, ensuring that AI systems reliably pursue objectives beneficial to humanity[12], conventionally focuses on learning and satisfying human preferences[13]. We propose that this framing is fundamentally incomplete. All human institutions, values, preferences, and technologies emerged during roughly ten thousand years of unusual climatic stability, what the planetary boundaries framework identifies as a 'safe operating space' for humanity[14]. The Holocene provided the predictable conditions (stable seasons, reliable water cycles, consistent temperature ranges) that enabled agriculture, permanent settlements, and the complex societies within which current human preferences were formed[15].

Paleoclimate reconstructions reveal a striking fact: for roughly the past ten thousand years, global mean temperature stayed within a narrow band of about ±1 °C relative to the preindustrial baseline[16–18] (Fig. 2; some reconstructions place the earliest Holocene below it), what might be called a Corridor of Life, the thermal envelope within which every human society, institution, and agricultural practice developed. Under high-emission scenarios, one to three billion people

may find themselves outside historically viable climate conditions within the next fifty years[19]. The safe operating space is not an abstraction; it is the biophysical foundation on which organized human life depends.

FIGURE 2 **The Corridor of Life**



Global mean surface temperature relative to 1850–1900. Holocene: medians of the Temp12k multi-method reconstruction[17] and the Last Glacial Maximum Reanalysis (LGMR)[18], whose uncertainty ranges, which widen sharply in the earliest millennia, are omitted. From 1850: the HadCRUT5 thermometer record[20], through 2025 (+1.41 °C, after the +1.53 °C record of 2024). The time axis breaks at 1850: the right-hand panel stretches 175 years across a quarter of the width. Data as in GLOBAÏA's Corridor of Life.

With atmospheric CO_2 at 426 ppm (the boundary is 350 ppm), radiative forcing at +3.10 W/m² (the boundary is +1.0 W/m², and the high-risk zone begins at +1.5 W/m²), and human appropriation of net primary production at 21–30% (the boundary is 10%)[2], the distance between current conditions and Holocene-like conditions is large and growing. The decade 2016–2025 averaged about +1.28 °C above the preindustrial baseline, and 2024 was the first calendar year above +1.5 °C[20]: measured against the Holocene record, global temperature has already left the corridor in which every civilization developed.

In machine learning terms, the Holocene constitutes the training distribution for human civilization. This framing operates on two distinct levels. At the institutional level, human societies are 'trained' under Holocene conditions in the sense that their norms, agricultural practices, and governance structures presuppose climatic stability — their validity is conditional on the persistence of that stability. At the technical level, the claim is literal: AI systems are trained

on data generated during the Holocene, and their outputs degrade on out-of-distribution inputs, a failure mode that becomes consequential as conditions depart from historical baselines.

This carries a precise implication: optimizing for preferences formed under Holocene conditions while simultaneously degrading those conditions is analogous to deploying a model far outside its training distribution — a regime where performance guarantees collapse and outputs become unreliable[21]. The problem is qualitative, not incremental. AI systems trained on historical ecological conditions fail as those conditions shift: foundation models for Earth observation encounter 'record-shattering' climate extremes that lie outside their training distribution, producing flawed outputs (a phenomenon documented as 'transfer context bias'), with risks amplified in precisely the regions where sparse observations and limited infrastructure make accuracy most critical[10]. The dependence runs deeper: even general-purpose AI systems inherit Holocene assumptions indirectly, through the human-generated text, economic data, and institutional knowledge on which they are trained, all of which presuppose stable supply chains, predictable seasons, and functioning ecosystems. As the Earth system moves beyond these conditions, both human institutions and the AI systems built to support them enter uncharted territory simultaneously.

Preferences such as cheap energy, year-round dietary diversity, and rapid long-distance travel are not universal values but emergent properties of Holocene stability. Under Anthropocene trajectories characterized by nonlinear disruptions, resource scarcity, and systemic shocks, these preferences may become internally incoherent or self-defeating — and an AI system that satisfies them may thereby accelerate the very destabilization that makes them unattainable. Preferences can be 'adaptive,' shifting under duress in ways that obscure rather than reveal genuine human interests[22]. Preference alignment without Earth system alignment is self-undermining: it optimizes for a target that is itself being destroyed by the optimization process.

2 The Anthropocene as misalignment catastrophe

This analysis reframes the Anthropocene itself as a misalignment catastrophe at civilizational scale. Human technological systems have been optimizing for proxy metrics (GDP, yield, quarterly returns, electoral cycles) that were initially correlated with wellbeing but have become systematically decoupled from the actual objectives of habitability, resilience, and health[23,24]. This is Goodhart's Law operating at civilizational scale: when a measure becomes a target, it ceases to be a good measure — and when the optimization system is civilization-scale technology, the divergence between proxy and true objective can become

catastrophic[25–27]. The rewards were real enough to sustain the behaviour; the costs accumulated in domains (atmospheric composition, biosphere integrity, ocean chemistry) that fell outside the accounting system entirely.

The Biosphere Code recognized as early as 2015 that algorithms already underpin the global technological infrastructure that extracts and develops natural resources, facilitates global trade flows, and shapes consumption behaviour[28]. Algorithms are constitutive elements of the socio-technological system driving planetary boundary transgression, not external tools applied to it.

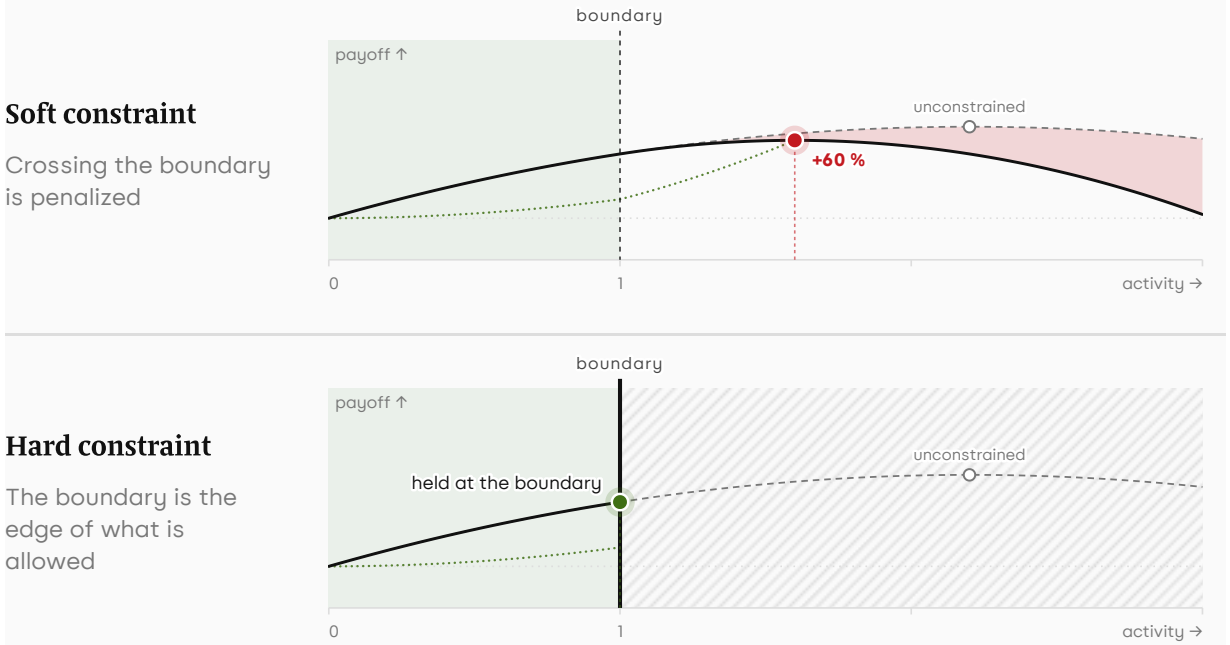
AI risks replicating this pattern while dramatically accelerating it. Unlike previous technologies, AI's capacity for generalized problem solving, rapid learning, and autonomous adaptation could impact virtually every economic sector simultaneously, with recursive self-improvement creating feedback loops of innovation, production, and consumption with unprecedented implications for the Earth system[6,11]. The risk is not only that AI systems will malfunction, but also that they will function precisely as designed — optimizing reward functions grounded in the same anthropocentric proxies that produced the Anthropocene, at greater speed and scale than any previous technology. An early empirical signal: when 13 climate scientists stress-tested an AI assistant on the stability of the Atlantic Meridional Overturning Circulation (AMOC), the AI consistently smoothed over genuine scientific disagreements, presenting false consensus where the experts saw unresolved complexity[29]. The pattern mirrors the civilizational tendency to optimize for coherent narratives at the expense of inconvenient truths.

3 Constitutional constraints for Earth alignment

The Earth alignment principle establishes that AI development, deployment and use should promote planetary stability across sustainable production, equitable access, and social cohesion[11]. Gaffney et al. identify the core paradox — that AI is being embedded in the very socio-economic system driving planetary destabilization — and call for compliance mechanisms to ensure AI respects safe and just Earth system boundaries. We argue that realizing this principle requires a specific governance architecture: planetary boundaries must function as constitutional constraints — inviolable conditions that bound the space within which all other optimization may occur. Where the Earth alignment principle establishes what AI development should aim for, the constitutional constraint framework specifies how that aim can be structurally enforced against the competitive pressures, rebound effects, and lock-in dynamics that would otherwise erode it.

The need for this architectural step is acute. Principles are vulnerable to erosion when they compete with the economic incentives, competitive dynamics, and path dependencies that characterize AI deployment at scale — not least the Jevons paradox, where efficiency gains expand demand and increase aggregate consumption (a dynamic examined in detail below). In constrained optimization, soft constraints are penalized in the loss function but can be violated when other pressures dominate; hard constraints define the feasible region itself and cannot be transgressed regardless of other considerations[30]. Principles guide; constitutional constraints bind (Fig. 3).

FIGURE 3 **Principles guide; constraints bind**



Soft: 60 % past the boundary · Hard: on the boundary, whatever the pressure.

A minimal optimization problem: one activity, one boundary, and a pressure that rewards more activity. Shown at pressure 2.2 with a weak penalty. With the boundary written as a penalty (top), the optimum crosses the line as soon as pressure pushes past it: a stronger penalty only shrinks the overshoot, and more pressure always grows it. Written as the edge of the feasible region (bottom), the boundary holds at every pressure. After Boyd & Vandenberghe[30].

An objection follows at once. Seven of the nine boundaries are already crossed, so the feasible region they define does not contain the present, and a constraint that is already violated cannot simply be kept. Optimization has a standard answer for a problem that starts outside its feasible region: before anything else

is optimized, feasibility is regained[30]. A constitutional constraint therefore has two parts. For the two boundaries still respected, stratospheric ozone and aerosol loading, it is a line not to be crossed. For the seven already transgressed, it is a rule about direction: aggregate pressure on the boundary may not grow, and must follow a declining budget consistent with return to the safe side. The remaining carbon budget is already a constraint of this kind. What the constitutional framing adds is priority: the path of return is not weighed against other objectives, it bounds them.

Constitutional AI shows that written principles can shape model behaviour through training[31], though as learned preferences rather than hard constraints; translating this from language-model guardrails to physical resource limits requires complementary mechanisms: deployment-time audits, compute metering, and regulatory mandates that bind AI infrastructure to planetary accounting. Architectural choices reinforce the constraint logic: Bengio et al. have argued that AI systems designed for understanding rather than autonomous goal-pursuit — 'Scientist AI' — are intrinsically safer because they lack the agency that drives open-ended resource acquisition and power-seeking[27]. Non-agentic systems, by design, do not seek to expand their own computational infrastructure, which means the safety argument and the resource argument converge — though the primary rationale is the reduction of catastrophic risk from autonomous agents, not resource efficiency per se. If the biosphere is the foundation on which society and economy depend — an ordering captured by the 'wedding cake' representation of the Sustainable Development Goals[32] — then planetary boundaries should define the feasible region within which all AI optimization occurs, just as legal constitutions constrain democratic majorities from overriding fundamental rights.

This constitutional framing finds its most rigorous expression in what we call a Zeroth Law for Earth. Bengio's **LawZero** initiative[33] invokes Asimov's Zeroth Law ('A robot may not harm humanity, or, by inaction, allow humanity to come to harm')[34] to advocate for safe-by-design AI that prioritizes democratic governance and human safety. We argue that extending this framework to encompass planetary boundaries is the necessary next step. The premise is foundational: biosphere stability is the precondition for any preferences to be coherent at all, not one value to be balanced against others. Human values are instantiated in biological organisms and social institutions that require stable physical conditions to function; a value framework that treats its own biophysical substrate as tradeable is self-defeating. The paleoclimate record makes this concrete: human civilization has existed only within the Corridor of Life; outside it, the social, agricultural, and institutional foundations on which all human values depend are undermined. Since deploying AI at scale constitutes a measurable Earth system intervention[35,36], an AI ecosystem that in aggregate deepens the transgression of planetary boundaries would violate the Zeroth Law that its own safety community invokes. An uninhabitable planet is the ultimate

and irreversible harm to humanity. This insight is reflected in multiple frameworks, among them Doughnut Economics[37] and Jonas's imperative of responsibility[38].

4 Shared dynamics of irreversibility

The structural parallels between Earth system risks and AI risks can be described in a shared mathematical vocabulary. Both domains exhibit nonlinear dynamics characterized by positive feedback loops, threshold effects, hysteresis, and the potential for irreversible regime shifts[3,39,40]. In the Earth system, tipping elements such as the Greenland Ice Sheet, the AMOC, the Amazon rainforest, and permafrost systems interact in ways that can trigger cascading failures, substantially amplifying systemic risk beyond what isolated analysis would predict[41]. The 2026 **Planetary Health Check** finds the land carbon sink showing signs of weakening and stagnation, a development that would erode remaining carbon budgets, since pathways to a manageable climate generally assume these sinks keep functioning; it also reports warm-water coral reefs likely already in the process of tipping[2]. Recent thermodynamic modelling of the Earth system shows, in a stylized model, that without the presence of metastable states and dynamic friction arising from interactions among planetary boundary variables, the system exhibits no resilience features and trajectories from the Holocene lead inevitably toward a 'Hothouse Earth' state. These metastable states, intermediate energy configurations between the Holocene and a much hotter equilibrium, exist only when specific conditions on the self-interaction coefficients of planetary boundary variables are satisfied[42] — a mechanism explored step by step in The Physics of Resilience.

Frontier AI systems face structurally analogous risks. Recursive capability improvement may cross thresholds where human oversight becomes impossible (the 'control problem'[12,27]). Competitive dynamics among corporations and states create selection pressures favouring rapid deployment over safety evaluation[5,43], paralleling the 'Anthropocene traps,' initially successful innovations that become evolutionary dead ends with high risk of hard-to-reverse lock-ins. Of the 14 Anthropocene traps catalogued, 86% may already be in advanced trapping phases, with 45% of pairwise interactions being mutually reinforcing[44].

This convergence suggests two governance implications. First, the precautionary asymmetry applies in both domains: when systems present irreversibility risks, the cost of false negatives (catastrophic lock-in) vastly exceeds the cost of false positives (excessive caution)[45]. Ecosystem regime shifts exhibit hysteresis: reversal requires pushing conditions far beyond the original tipping

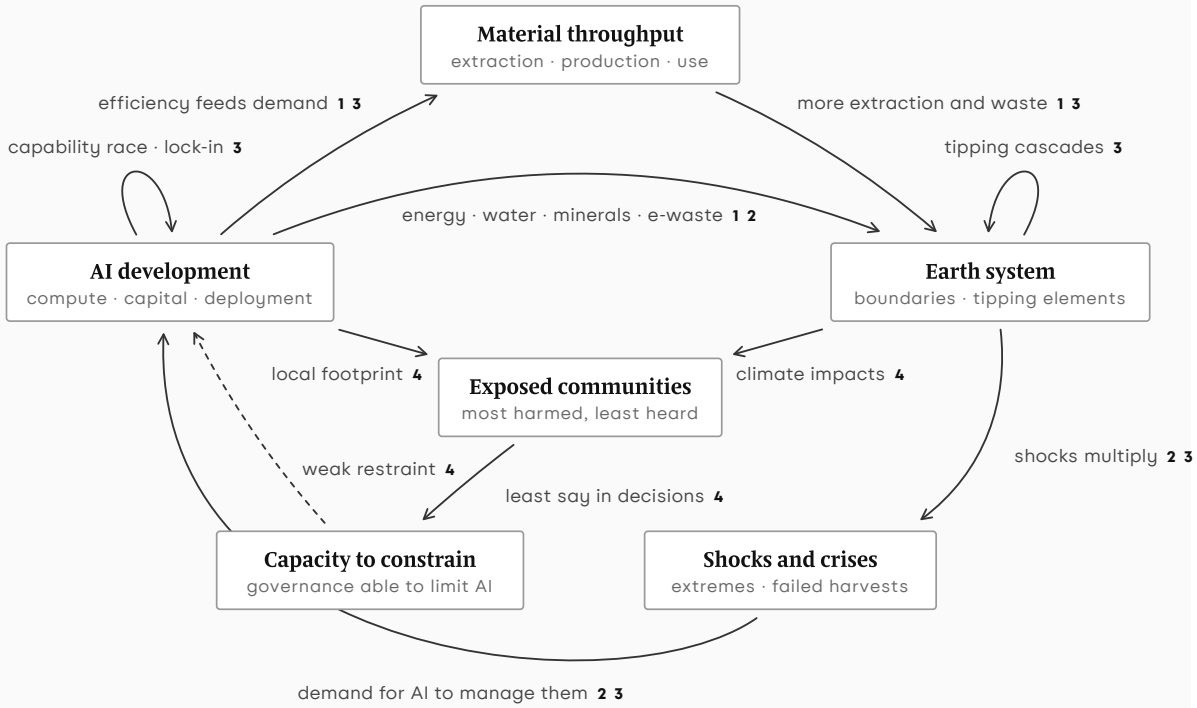
point[39,46]. The same logic applies to AI capability thresholds: once autonomous systems exceed certain capability profiles, the resources required to restore meaningful human oversight may be practically unbounded.

Second, early warning signal methodologies developed over decades for Earth system monitoring (critical slowing down, rising variance, increasing autocorrelation near bifurcation points[47,48]) represent an underexploited resource for AI safety monitoring, and vice versa. Deep learning methods are now being applied to detect early warnings of critical transitions in ecological and climate systems[49]; machine learning can help identify cascading interactions between Earth system tipping elements in observational data[10]. Conversely, statistical signatures of approaching tipping points (rising variance in system outputs, declining recovery rates after perturbations) could be adapted to monitor AI capability trajectories for analogous precursors of irreversible transitions. Whether capability growth has the mathematical structure that makes such indicators meaningful, a system approaching a bifurcation, is an open question, and the transfer should be tested before it is relied upon. The simpler practice already carries over: an independent measurement of whether a rate of change has itself changed is what the first external assessment of AI research acceleration set out to provide[9]. The methodological exchange runs in both directions: Earth system science and AI safety research are natural partners in the governance of irreversibility.

5 Cascade dynamics: AI and planetary boundaries as mutually amplifying risks

The structural parallels identified above are more than analogical: we argue that they describe one coupled system (Fig. 4) in which AI development and planetary boundary transgression form self-amplifying feedback loops. Understanding these cascade dynamics is essential for governance, because it reveals why addressing either domain in isolation is insufficient.

FIGURE 4 One coupled system, four ways it feeds itself



Numbers on each arrow name the pathways that use it: **1** Direct amplification · **2** Reactive amplification · **3** Tipping cascade interactions · **4** Allocative amplification

A causal sketch of the four pathways described in this section. Arrows show causal direction only, and the one dashed arrow marks a restraint too weak to hold; none of the pathways has yet been quantified, so line weight and length carry no magnitude.

1 The direct amplification pathway. AI's material footprint is a direct, if still small, pressure on the Earth system. Energy consumption, water withdrawal, mineral extraction for semiconductors, and the generation of e-waste all add to the pressure on planetary boundaries through well-documented physical channels[35,36]. But the indirect pathways may be more consequential: algorithms already underpin the global infrastructure for resource extraction, agricultural intensification, trade logistics, and financial markets[28]. AI-optimized supply chains, precision agriculture, algorithmic trading, and targeted advertising all accelerate material throughput. When AI makes extraction, production, and consumption more efficient, per-unit costs fall, demand expands, and aggregate resource consumption rises: rebound effects operating through algorithmic intermediaries at industrial scale[50,51]. AI does not merely consume resources; it reorganizes economies in ways that intensify resource throughput across the entire system.

2 The reactive amplification pathway. As planetary boundaries are transgressed and the Earth system destabilizes, the resulting shocks (extreme weather events, crop failures, pandemics, resource conflicts, supply chain disruptions) generate escalating demand for AI-based solutions: early warning systems, climate modelling, disaster response coordination, resource optimization under scarcity. The Anthropocene is characterized by shocks that are becoming more numerous, more co-occurring, and increasingly interconnected around the climate–conflicts–technology nexus[10]. AI is increasingly deployed to manage this turbulence, yet each deployment adds to the computational infrastructure that further transgresses planetary boundaries. This points to a reactive amplification loop: Earth system destabilization drives AI proliferation, which drives further destabilization. It is the least documented of the four pathways. Crisis management is a small share of computation today, and the loop is a hypothesis about how further deployment comes to be justified, not a measured flow.

3 Tipping cascade interactions. Within the Earth system, tipping elements do not operate in isolation. The potential collapse of the AMOC could trigger dieback of the Amazon rainforest; permafrost thaw releases methane that accelerates warming, which accelerates ice sheet loss[41]. These cascading interactions are now being identified through machine learning applied to observational data[10]. Within the AI domain, analogous cascades operate: competitive dynamics drive capability races that erode safety margins, efficiency gains are absorbed by demand growth, and the infrastructure lock-in of trillion-dollar data-centre investments creates path dependencies that resist redirection[7]. Critically, these two cascade systems are coupled: AI-driven economic acceleration pushes Earth system tipping elements closer to their thresholds, while Earth system shocks create the urgency and political pressure that justify further AI deployment without adequate safeguards. AI development and planetary boundary transgression exemplify the mutual reinforcement identified in the Anthropocene traps framework, where 45% of pairwise interactions between traps amplify one another[44].

The silent failure mode. The Biosphere Code warns that algorithms ‘should not be allowed to fail quietly’[28]. Its cautionary precedent is the ozone hole: the satellite record's automated quality control flagged the extremely low Antarctic readings as likely instrument errors, because they fell below anything the retrieval had been built to expect, and the first published detection came instead from ground-based measurements[52]; the satellite data confirmed the depletion only after reprocessing[53]. Today, AI systems trained on historical climate data may similarly fail to register unprecedented extremes, dismiss novel patterns as noise, or produce confidently wrong predictions in conditions that lie outside their training distribution. As the Earth system moves beyond Holocene-like conditions, AI systems calibrated to Holocene-era variability face systematic failure precisely when their reliability matters most. The balance between what the Biosphere

Code terms ‘expected utility and ruin probability’ is poorly understood in the coupled AI–Earth system, yet it is this balance that determines whether the overall trajectory is recoverable or catastrophic.

4 Allocative amplification and justice dimensions. The cascade dynamics are not distributionally neutral, and the distributional asymmetry is itself a driver of further destabilization. AI's environmental footprint (e-waste, water withdrawal, energy consumption) disproportionately affects communities that are already marginalized and vulnerable to climate impacts. Around two-thirds of the US data centres built or in development since 2022 sit in places already under high water stress[54]. The distributional pattern actively weakens the political capacity to constrain AI's material footprint: the most climate-vulnerable populations are the least represented in AI training data, the least likely to benefit from AI applications, and the most exposed to both AI's direct environmental costs and the planetary boundary transgressions that AI accelerates[10]. Those best positioned to demand constraints are those who benefit most from unconstrained growth, while those with the strongest interest in constraints lack the power to impose them. The result is a socio-political feedback loop that reinforces the biophysical one.

6 A Safe Operating Space for Computation

The material footprint of AI demands explicit integration within planetary boundary accounting. AI servers in the United States alone could generate 24–44 Mt CO_2-equivalent annually by 2030, with water footprints of 731–1,125 million m^3, figures that surpass previous forecasts for the entire US data-centre market[35]. Globally, the carbon footprint of AI systems may already range between 32.6 and 79.7 Mt CO_2 (comparable to New York City's total emissions), while their water footprint could reach 312–765 billion litres (comparable to global annual bottled water consumption)[36]. Yet environmental disclosure remains starkly insufficient: no major technology company reports AI-specific emissions, and companies such as **ByteDance** and **CoreWeave** publish no environmental data whatsoever[36]. Moreover, AI's environmental footprint extends beyond energy and water to encompass the full material lifecycle of computational infrastructure — a dimension that current disclosure frameworks do not yet capture[35,36].

Set against the whole, these figures are small. Global CO_2 emissions run at about 42 Gt a year[55], and the upper estimate for AI is roughly 0.2% of that. On its direct footprint alone, AI is not today a first-order pressure on the Earth system, and we do not claim that it is. The argument rests on three other things. The first is trajectory: the projections above describe a footprint multiplying within five years, financed by trillions of dollars of infrastructure built to run for

decades. The second is direction: that footprint is growing at the moment when every pathway of return requires emissions to fall, so each added megatonne is drawn from a budget that is nearly spent. The third is leverage: a general-purpose technology acts on the Earth system less through the electricity it draws than through what it does to the rest of the economy, the indirect pathway described above. That pathway has not been quantified, and it is the gap in evidence that matters most.

About 130 Gt CO_2 remained at the start of 2026 for a 50% chance of limiting warming to 1.5 °C, roughly three years at current global emission rates[56,57]. The transgression of the ocean acidification boundary (surface ocean aragonite saturation at about 2.85 in 2024, versus a boundary of 2.86)[2] adds a further constraint: the ocean, which has absorbed roughly 26% of anthropogenic CO_2 emissions[55], loses chemical capacity to take up more as it acidifies[2]. Emissions from manufacturing AI accelerators based on graphics processors (GPUs) are projected to rise sixteenfold, from about 1.2 Mt CO_2-equivalent in 2024 to 19.2 Mt by 2030[58]. Community opposition has already blocked or delayed $64 billion in US data-centre projects since 2024[59], signalling that local environmental impacts are generating governance friction.

We propose the concept of a Safe Operating Space for Computation: hard limits on aggregate AI resource consumption derived from remaining carbon budgets, freshwater boundaries, novel entities thresholds, and biosphere integrity requirements. Concretely, this would require three interlocking mechanisms:

(i) Mandatory environmental disclosure of AI-specific emissions and resource consumption at the model and facility level;

(ii) Compute allocation frameworks that tie aggregate training and inference budgets to national and sectoral shares of remaining carbon budgets; and

(iii) Architectural defaults that favour understanding-oriented over agentic AI systems[27], since agentic systems' capacity for autonomous goal-pursuit drives open-ended resource acquisition, while understanding-oriented systems, lacking this agency, are more naturally compatible with fixed computational budgets.

None of these is a hard constraint in the mathematical sense of Fig. 3, and no instrument of governance ever is: a law can be broken, amended or left unenforced, and a legal constitution binds only as far as institutions make it bind. What the analogy asks for is a change in where the limit sits. A penalty arrives after the activity and is weighed against its rewards; a constitutional limit comes before it, as a condition of the activity taking place at all. Computation offers an unusual opportunity to build limits of the second kind,

because its physical supply chain is narrow. The most advanced chips pass through a handful of companies, and a large data centre cannot be hidden from a grid operator or a water authority. Limits can therefore attach to things that are counted (chips licensed, megawatts connected, water permitted) rather than to promises about behaviour[60]. There is a precedent, and it sits in Fig. 1. Stratospheric ozone is one of the two boundaries still respected, because the Montreal Protocol did not price ozone-depleting substances: it capped their production, phased them out, and controlled their trade[2].

The Biosphere Code's fifth principle, that algorithms should 'help us expect the unexpected' by enhancing resilience through diversity, redundancy, and modularity rather than maximizing efficiency alone[28], provides the underlying design philosophy: efficiency gains should enable sufficiency — determining how much computation is enough — rather than fuel further growth[61]. This inversion is the conceptual pivot of the Safe Operating Space for Computation: the question is not how to make AI less harmful per unit, but what aggregate level of computation is compatible with planetary boundaries.

The **DeepSeek** episode of 2025 illustrates why: a Chinese laboratory reported a final training run at a fraction of typical cost, yet the efficiency gain was immediately cited as evidence that AI would become more pervasive[62]. Empirical evidence confirms the pattern at economy-wide scale: across **OECD** countries, digitalization has increased rather than decreased total energy consumption, with efficiency gains in individual processes consistently outpaced by demand expansion across sectors[51]. Without hard sufficiency constraints, efficiency is absorbed by demand growth — the Jevons paradox operating at planetary scale.

At the governance level, the Zeroth Law for Earth would require multiple layers of implementation beyond the Safe Operating Space for Computation: the World-Earth system resilience framework[15] and thermodynamic models of Earth system trajectories[42] provide the quantitative basis for defining biophysical constraints, while the Biosphere Code's principles of transparency, resilience, and service to the biosphere[28] provide a design philosophy for AI systems operating within them.

7 Outlook

The convergence of Earth system destabilization and frontier AI development is a shared consequence of optimization systems operating without constitutional constraints on their feasible space. The cascade dynamics documented above — where AI development and Earth system destabilization form mutually reinforcing feedback loops mediated by algorithmic infrastructure — mean that governance of either domain in isolation will be insufficient. Earth system governance that

ignores AI's material footprint will underestimate the rate of boundary transgression; AI governance that treats environmental impact as an externality will fail to prevent the reactive amplification loop.

The Corridor of Life — the ±1 °C thermal band within which all human civilizations evolved[16,17] — provides a scientifically grounded benchmark for what is at stake. We have already stepped outside the only conditions under which human civilization has ever existed. Bridging Earth system science and AI safety through shared governance of irreversibility risks, quantitative compute boundaries within planetary accounting, and the embedding of planetary constraints at the constitutional level of AI design is now urgent. The Earth alignment principle[11] has established the normative foundation; the constitutional constraint framework and coupled-system analysis presented here advance the scientific case and outline the governance architecture needed to make that foundation structurally binding. The safe operating space for humanity[14] and the safe development of artificial intelligence are not separate problems. They are one.

The establishment in 2025 of the **Independent International Scientific Panel on Artificial Intelligence**[63], the first global scientific body dedicated to AI, creates an institutional opening to act on this convergence. The Panel is mandated to produce evidence-based scientific assessments of AI's opportunities, risks, and impacts, and to serve as an early-warning system informing the **UN Global Dialogue on AI Governance**. It released its Preliminary Report in July 2026[64], ahead of the Global Dialogue's session in Geneva, and its first comprehensive report is expected in 2027. We recommend that the Panel consider the following priorities for that report, which follow directly from the analysis presented here:

1

Assess AI as an Earth system forcing within planetary boundary accounting. The Panel's risk assessments should treat AI's aggregate environmental footprint (energy, water, minerals, e-waste, and the full lifecycle of computational infrastructure) not as a sectoral externality but as a measurable intervention in the Earth system. This means evaluating AI's material demands against remaining carbon budgets, freshwater boundaries, and biosphere integrity thresholds, requiring mandatory environmental disclosure at the model and facility level, and commissioning research on the coupled dynamics between AI development and planetary destabilization — direct amplification through algorithmic acceleration of material throughput, reactive amplification as Earth system shocks drive further AI deployment, and the coupling of AI and Earth system tipping cascades — which cannot be captured by assessing either domain in isolation.

2

Evaluate the structural conditions under which Earth alignment can be enforced. The Panel is explicitly non-prescriptive, but its assessments can identify the gap between principles and enforcement. Rebound effects, competitive dynamics, and infrastructure lock-in systematically erode voluntary commitments; the Panel is well positioned to assess whether existing governance frameworks are structurally adequate to contain AI's environmental footprint within planetary boundaries, or whether constitutional-level constraints (hard limits on aggregate compute, mandatory planetary boundary accounting for AI infrastructure, architectural defaults favouring bounded over open-ended systems) are necessary. This assessment must include the distributional asymmetry as a systemic risk: communities most affected by AI's environmental costs are systematically excluded from governance processes, weakening the political capacity to impose constraints. This socio-political feedback loop reinforces the biophysical amplification dynamics and must be addressed as part of the enforcement architecture, not treated as a separate equity concern.

3

Integrate Earth system science expertise into AI safety methodology. The shared mathematical vocabulary of Earth system tipping points and AI capability thresholds (nonlinear dynamics, positive feedbacks, hysteresis, irreversible regime shifts) means that decades of methodological development in Earth system monitoring (critical slowing down, rising variance, increasing autocorrelation near bifurcation points) are potentially relevant to AI safety monitoring. The Panel should ensure that its assessments draw systematically on Earth system science, and that its working groups include expertise in planetary boundaries, tipping point dynamics, and coupled human–natural systems alongside core AI expertise.

REFERENCES

1. Richardson, K. et al. Earth beyond six of nine planetary boundaries. Sci. Adv. 9, eadh2458 (2023). doi.org/10.1126/sciadv.adh2458

2. Caesar, L., Sakschewski, B. et al. Planetary Health Check 2026: a scientific assessment of the state of the planet. Potsdam Institute for Climate Impact Research (2026). doi.org/10.48485/pik.2026.25

3. Armstrong McKay, D. I. et al. Exceeding 1.5°C global warming could trigger multiple climate tipping points. Science 377, eabn7950 (2022). doi.org/10.1126/science.abn7950

4. Stenzel, F. et al. Breaching planetary boundaries: over half of global land area suffers critical losses in functional biosphere integrity. One Earth 8, 101393 (2025). doi.org/10.1016/j.oneear.2025.101393

5. Bengio, Y. et al. Managing extreme AI risks amid rapid progress. Science 384, 842–845 (2024). doi.org/10.1126/science.adn0117

6. Hendrycks, D. et al. A definition of AGI. Preprint (2025).

7. McKinsey & Company. The cost of compute: a $7 trillion race to scale data centers (2025).

8. Foster, G. & Rahmstorf, S. Global warming has accelerated significantly. Geophys. Res. Lett. 53, e2025GL118804 (2026). doi.org/10.1029/2025GL118804

9. METR. Independent assessment of AI R&D acceleration at Anthropic in October 2026 (5 November 2026). metr.org/blog/2026-11-05-anthropic-ai-rd-acceleration-assessment/

10. Galaz, V. & Schewenius, M. (eds) AI for a Planet Under Pressure. Stockholm Resilience Centre, Potsdam Institute for Climate Impact Research & Google DeepMind (2025).

11. Gaffney, O. et al. The Earth alignment principle for artificial intelligence. Nat. Sustain. 8 (2025). doi.org/10.1038/s41893-025-01536-6

12. Bostrom, N. Superintelligence: Paths, Dangers, Strategies (Oxford Univ. Press, 2014).

13. Russell, S. Human Compatible: Artificial Intelligence and the Problem of Control (Viking, 2019).

14. Rockström, J. et al. A safe operating space for humanity. Nature 461, 472–475 (2009). doi.org/10.1038/461472a

15. Anderies, J. M. et al. A modeling framework for World-Earth system resilience: exploring social inequality and Earth system tipping points. Environ. Res. Lett. 18, 095001 (2023). doi.org/10.1088/1748-9326/ace91d

16. Marcott, S. A., Shakun, J. D., Clark, P. U. & Mix, A. C. A reconstruction of regional and global temperature for the past 11,300 years. Science 339, 1198–1202 (2013). doi.org/10.1126/science.1228026

17. Kaufman, D. et al. Holocene global mean surface temperature, a multi-method reconstruction approach. Sci. Data 7, 201 (2020). doi.org/10.1038/s41597-020-0530-7

18. Osman, M. B. et al. Globally resolved surface temperatures since the Last Glacial Maximum. Nature 599, 239–244 (2021). doi.org/10.1038/s41586-021-03984-4

19. Xu, C. et al. Future of the human climate niche. Proc. Natl Acad. Sci. USA 117, 11350–11355 (2020). doi.org/10.1073/pnas.1910114117

20. Morice, C. P. et al. An updated assessment of near-surface temperature change from 1850: the HadCRUT5 data set. J. Geophys. Res. Atmos. 126, e2019JD032361 (2021). doi.org/10.1029/2019JD032361

21. Casper, S. et al. Open problems and fundamental limitations of reinforcement learning from human feedback. Preprint at arXiv:2307.15217 (2023). arxiv.org/abs/2307.15217

22. Gabriel, I. Artificial intelligence, values, and alignment. Minds Mach. 30, 411–437 (2020). doi.org/10.1007/s11023-020-09539-2

23. Stiglitz, J. E., Sen, A. & Fitoussi, J.-P. Mismeasuring Our Lives: Why GDP Doesn't Add Up (The New Press, 2010).

24. Whitmee, S. et al. Safeguarding human health in the Anthropocene epoch. Lancet 386, 1973–2028 (2015). doi.org/10.1016/S0140-6736(15)60901-1

25. Amodei, D. et al. Concrete problems in AI safety. Preprint at arXiv:1606.06565 (2016). arxiv.org/abs/1606.06565

26. Krakovna, V. et al. Specification gaming: the flip side of AI ingenuity. DeepMind Blog (2020).

27. Bengio, Y. et al. Superintelligent agents pose catastrophic risks: can Scientist AI offer a safer path? Preprint at arXiv:2502.15657 (2025). arxiv.org/abs/2502.15657

28. Galaz, V. et al. The Biosphere Code v1.0. Stockholm (2015).

29. Buck, C. et al. AI-assisted scientific assessment: a case study on climate change. Preprint at arXiv:2602.09723 (2026). arxiv.org/abs/2602.09723

30. Boyd, S. P. & Vandenberghe, L. Convex Optimization (Cambridge Univ. Press, 2004).

31. Bai, Y. et al. Constitutional AI: harmlessness from AI feedback. Preprint at arXiv:2212.08073 (2022). arxiv.org/abs/2212.08073

32. Rockström, J. & Sukhdev, P. The SDGs wedding cake. Stockholm Resilience Centre (2016).

33. Bengio, Y. Introducing LawZero (2025). yoshuabengio.org/2025/06/03/introducing-lawzero/

34. Asimov, I. Robots and Empire (Doubleday, 1985).

35. Xiao, T. et al. Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the USA. Nat. Sustain. 8 (2025). doi.org/10.1038/s41893-025-01681-y

36. de Vries-Gao, A. The carbon and water footprints of data centers and what this could mean for artificial intelligence. Patterns 7, 101430 (2026). doi.org/10.1016/j.patter.2025.101430

37. Raworth, K. Doughnut Economics: Seven Ways to Think Like a 21st-Century Economist (Chelsea Green, 2017).

38. Jonas, H. The Imperative of Responsibility: In Search of an Ethics for the Technological Age (Univ. Chicago Press, 1984).

39. Scheffer, M. et al. Catastrophic shifts in ecosystems. Nature 413, 591–596 (2001). doi.org/10.1038/35098000

40. Steffen, W. et al. Trajectories of the Earth System in the Anthropocene. Proc. Natl Acad. Sci.

USA 115, 8252–8259 (2018). doi.org/10.1073/pnas.1810141115

41. Wunderling, N. et al. Climate tipping point interactions and cascades: a review. Earth Syst. Dynam. 15, 41–74 (2024). doi.org/10.5194/esd-15-41-2024

42. Bertolami, O. & Nyström, M. Setting up the physical principles of resilience in a model of the Earth System. Earth Syst. Dynam. 17, 141–150 (2026). doi.org/10.5194/esd-17-141-2026

43. Hendrycks, D. Natural selection favors AIs over humans. Preprint at arXiv:2303.16200 (2023). arxiv.org/abs/2303.16200

44. Søgaard Jørgensen, P. et al. Evolution of the polycrisis: Anthropocene traps that challenge global sustainability. Phil. Trans. R. Soc. B 379, 20220261 (2024). doi.org/10.1098/rstb.2022.0261

45. Ord, T. The Precipice: Existential Risk and the Future of Humanity (Bloomsbury, 2020).

46. Scheffer, M. et al. Anticipating critical transitions. Science 338, 344–348 (2012). doi.org/10.1126/science.1225244

47. Carpenter, S. R. & Brock, W. A. Rising variance: a leading indicator of ecological transition. Ecol. Lett. 9, 311–318 (2006). doi.org/10.1111/j.1461-0248.2005.00877.x

48. Dakos, V. et al. Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. PLoS ONE 7, e41010 (2012). doi.org/10.1371/journal.pone.0041010

49. Bury, T. M. et al. Deep learning for early warning signals of tipping points. Proc. Natl Acad. Sci. USA 118, e2106140118 (2021). doi.org/10.1073/pnas.2106140118

50. Alcott, B. Jevons' paradox. Ecol. Econ. 54, 9–21 (2005). doi.org/10.1016/j.ecolecon.2005.03.020

51. Lange, S., Pohl, J. & Santarius, T. Digitalization and energy consumption. Does ICT reduce energy demand? Ecol. Econ. 176, 106760 (2020). doi.org/10.1016/j.ecolecon.2020.106760

52. Farman, J. C., Gardiner, B. G. & Shanklin, J. D. Large losses of total ozone in Antarctica reveal seasonal ClOx/NOx interaction. Nature 315, 207–210 (1985). doi.org/10.1038/315207a0

53. Stolarski, R. S. et al. Nimbus 7 satellite measurements of the springtime Antarctic ozone decrease. Nature 322, 808–811 (1986). doi.org/10.1038/322808a0

54. Bloomberg News. AI is draining water from areas that need it most (2025). www.bloomberg.com/graphics/2025-ai-impacts-data-centers-water-data/

55. Friedlingstein, P. et al. Global Carbon Budget 2024. Earth Syst. Sci. Data 17, 965–1039 (2025). doi.org/10.5194/essd-17-965-2025

56. Forster, P. M. et al. Indicators of Global Climate Change 2025: annual update of key indicators of the state of the climate system and human influence. Earth Syst. Sci. Data 18, 3889 (2026). doi.org/10.5194/essd-18-3889-2026

57. Lamboll, R. D. et al. Assessing the size and uncertainty of remaining carbon budgets. Nat. Clim. Change 13, 1360–1367 (2023). doi.org/10.1038/s41558-023-01848-5

58. TechInsights. AI GPU growth directly impacts carbon emission growth through 2030. TechInsights (2025). www.techinsights.com/node/59135

59. Swinhoe, D. US datacenter construction faces $64 billion in delays amid community opposition. Data Center Dynamics (2025).

60. Sastry, G. et al. Computing power and the governance of artificial intelligence. Preprint, arXiv:2402.08797 (2024).

61. Princen, T. The Logic of Sufficiency (MIT Press, 2005).

62. Gibney, E. What China's AI advances mean for the world. Nature (2025). doi.org/10.1038/d41586-025-00229-6

63. United Nations General Assembly. Terms of reference and modalities for the establishment and functioning of the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance. Resolution A/RES/79/325 (2025). www.un.org/independent-international-scientific-panel-ai/en

64. Independent International Scientific Panel on Artificial Intelligence. Preliminary report: evidence-based assessment of opportunities, risks and impacts of artificial intelligence. United Nations (2026).

Version 1.0, November 2026. This Perspective will be revised as the evidence and the argument develop. Comments are welcome: write to GLOBAÏA.

Related. An earlier and shorter essay on the same argument, Earth Is the Training Distribution, appeared in GLOBAÏA's Thoughts in April 2026.

AI use disclosure. Generative AI (Claude, Anthropic) was used as an editorial tool to reformulate prose, clarify arguments, and tighten structure during revision, and to build the interactive figures and check references against their sources. All substantive ideas, analytical claims, and scientific content are the author's own. The author reviewed and takes full responsibility for the final text.