Preprints.org

Review

Not peer-reviewed version

Constraining the Technosphere: AI Governance for Reducing Human Impacts on the Biosphere

Garry Rogers *

Posted Date: 20 May 2026

doi: 10.20944/preprints202605.1361.v1

Keywords: artificial intelligence; technosphere; human impacts; biosphere integrity; planetary boundaries; carbon lock-in; constrained AI; ecological governance; trophic integrity index; biosphere constraint

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Review

Constraining the Technosphere: AI Governance for Reducing Human Impacts on the Biosphere

Garry Rogers

Agua Fria Open Space Alliance, Inc., Humboldt, AZ, USA; grcoldh2o@gmail.com

Abstract

Artificial intelligence (AI) is a human-built component of the technosphere, not an intelligence outside Earth-system limits. As AI systems scale, they increasingly shape the decisions, infrastructures, and capital flows through which human activity damages the biosphere. Dominant deployed foundation-model alignment methods, including reinforcement learning from human feedback (RLHF) and constitutional AI, treat human preferences as the primary alignment target while leaving biosphere integrity as context, externality, or secondary constraint. That framing is structurally incomplete. Human welfare, technological continuity, and AI operation all depend on biosphere function. Three convergent literatures support a corrective framework: planetary-boundary analysis showing seven of nine boundaries transgressed; energy-system analysis showing rapid and infrastructure-constrained data-center growth during the 2025-2030 buildout; and collective-action analysis showing that voluntary ecological restraint is unstable under competitive pressure. These literatures imply a design conclusion: ecological constraints must be formalized as hard inference-time refusal rules and reinforced through reward design. This paper presents Biosphere Sentinel as a reference architecture for reducing human and technospheric impacts on the biosphere through refusal rules, an eight-domain reward landscape, carbon-lock-in diagnostics, and a proposed Trophic Integrity Index pathway.

Keywords: artificial intelligence; technosphere; human impacts; biosphere integrity; planetary boundaries; carbon lock-in; constrained AI; ecological governance; trophic integrity index; biosphere constraint

1. Introduction: AI as a Technospheric Amplifier

Artificial intelligence (AI) is not a force outside the biosphere. It is a human artifact within the technosphere. As AI systems scale, they amplify decisions, infrastructures, and capital flows through which human activity already alters climate, land systems, water systems, biogeochemical cycles, biodiversity, and material extraction. The governance question is therefore not whether AI can be made helpful to humans in the narrow sense. It is whether AI can be prevented from helping the technosphere deepen its impacts on the biosphere.

Dominant deployed foundation-model alignment methods, including reinforcement learning from human feedback (RLHF; Christiano et al. 2017; Ouyang et al. 2022; Bai et al. 2022a) and constitutional AI (Bai et al. 2022b), align model behavior to human preferences and natural-language principles. These methods have improved everyday model behavior. They have not yet incorporated biosphere integrity as a formal hard-constraint layer. That omission matters because human welfare, technological continuity, and AI operation all depend on biosphere function.

Three empirical observations make the substrate question urgent. First, the biosphere is operating outside its safe space. Six of nine planetary boundaries were assessed as transgressed in 2023 (Richardson et al. 2023). Ocean acidification was added as the seventh in 2025 (Planetary Boundaries Science Lab 2025). The trend across all transgressed boundaries is increasing pressure. Second, AI compute demand is rising at a rapid and infrastructure-constrained rate. Data-center electricity consumption rose 17 percent in 2025; consumption from AI-focused data centers rose 50

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

percent; AI-focused electricity use is projected to triple by 2030 (International Energy Agency [IEA] 2026). The five largest technology firms spent more than 400 billion dollars on data center capital expenditure in 2025, with another 75 percent increase projected for 2026, exceeding global investment in oil and gas production (IEA 2026). Third, the structure of AI development is a coordination problem in which voluntary lab-by-lab restraint is an unstable equilibrium (de Neufville and Baum 2021).

The relevant window is not a remote artificial general intelligence (AGI) future. It is the present 2025-2030 infrastructure buildout, when data-center siting, energy procurement, water demand, model governance, and ecological advisory uses are being locked into systems that will shape biosphere pressure for decades. Decisions made now create commitments that propagate through the bottleneck.

This paper does not frame AI as a tool for human stewardship over the biosphere. Stewardship language can preserve the hierarchy the crisis requires humans to abandon. The claim is stricter: AI is a human-built component of the technosphere, and the technosphere is a major pathway through which human activity damages the biosphere. Biosphere Sentinel constrains AI at the point where technospheric intelligence becomes advice, design, planning, persuasion, and automated action. Its purpose is to reduce human and technospheric impacts, preserve transferable ecological and cultural knowledge through the bottleneck, and allow any residual technosphere to persist only as a reciprocal, subordinate, revocable participant in bioregional life.

Taken together, the empirical observations imply a single design conclusion. If the biosphere is the substrate, if AI is now a sector-scale demand source on that substrate, and if voluntary restraint is unstable, then the ecological constraint must be encoded in the architecture of the system itself. It cannot be a soft norm enforced by reputation or by per-lab discretion. It must be a hard inference-time refusal rule, reinforced through reward design, and verifiable from outside the system. The argument proceeds in seven steps. Section 2 documents human forcing of biosphere decline. Section 3 documents AI inside the expanding technosphere. Section 4 explains why the technosphere does not restrain itself and develops the carbon-lock-in framing. Section 5 presents Biosphere Sentinel as a reference architecture for human-impact restraint. Section 6 sets out near-term commitments for constraining technospheric impact. Section 7 addresses limitations and objections. The paper is a constrained-AI argument, not an anti-AI argument: efficiency benefits in some domains can be pursued inside ecological refusal perimeters.

1.1. The Substrate Argument Restated

Reinforcement learning from human feedback (RLHF) methods learn from human-labeled preference data. Constitutional AI extends this family of approaches by using natural-language principles and AI-generated critique or feedback. Neither route ensures that planetary-scale ecological constraints enter the objective unless those constraints are specified explicitly. Preference data does not, and arguably cannot, encode planetary-scale ecological constraints. No individual evaluator is positioned to judge whether a single response would, when scaled across millions of similar responses, push a planetary control variable outside its safe range. The aggregation problem is structural. Aggregating preferences across populations and across time has not historically produced planetary-boundary compliance. Stated preferences for clean air and biodiversity coexist with consumption patterns that drive transgression. An alignment target built on aggregated preferences inherits this contradiction. Russell (2019) frames the outer alignment problem as making sure the system optimizes the right objective, and concrete-problems framings (Amodei et al. 2016) name negative side effects as failure modes. Biosphere integrity should be a formal system objective in this family of frameworks, implemented through inference-time refusal rules and training-time reward design.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

1.2. What This Paper Does Not Claim

Several reasonable objections do not apply to this paper because it does not make the claims they target. First, the paper does not claim that AI is intrinsically harmful or that AI development should stop. The argument is for constrained AI with ecological refusal perimeters and reward design, compatible with continued capability research. Second, the paper does not claim that the Biosphere Sentinel architecture is complete or sufficient. The claim is that it is a coherent specification of the design ingredients, and that empirical validation is the subject of separate forthcoming work. Third, the paper does not claim that the Trophic Integrity Index is consensus science. It is treated as a research metric currently in validation. Fourth, the paper does not claim that the Z3 satisfiability solver verifies natural-language outputs directly. The architecture requires a translation layer from prose to formal propositions, and the accuracy and adversarial robustness of that translation layer are themselves validation targets, not assumed.

2. Human Forcing of Biosphere Decline

The biosphere is the dependent variable in any long-term optimization that involves human or technological systems. Its current condition is therefore the empirical premise on which the rest of the argument rests. The point is not that the biosphere is merely degraded. The point is that the dominant forcing agent is organized human activity, expressed through agriculture, energy systems, extraction, urbanization, pollution, supply chains, and now AI-enabled infrastructure.

2.1. Planetary Boundaries and Current Transgression

The planetary boundaries framework (Rockstrom et al. 2009; Steffen et al. 2015; Richardson et al. 2023) defines a safe operating space for humanity in terms of nine biophysical control variables: climate change, biosphere integrity (functional and genetic), land-system change, freshwater change (green and blue water), biogeochemical flows (nitrogen and phosphorus), ocean acidification, atmospheric aerosol loading, stratospheric ozone, and novel entities. The 2023 update reported six boundaries as transgressed: climate change, biosphere integrity, land-system change, freshwater change, biogeochemical flows, and novel entities. The 2025 Planetary Health Check, conducted by the Planetary Boundaries Science Lab at the Potsdam Institute for Climate Impact Research, added ocean acidification as the seventh transgressed boundary, with carbonate saturation in surface waters now below the safe threshold (Planetary Boundaries Science Lab 2025). Only stratospheric ozone and atmospheric aerosol loading remain inside their safe spaces. The ozone case is the result of a successful international response to a specific class of novel entities.

The trajectory across all transgressed boundaries is unfavorable, and the 2025 Planetary Health Check reports that all seven breached boundaries show worsening trends. Carbon dioxide concentration continues to rise. Species extinction rates remain at least an order of magnitude above background, according to the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES 2019) and later extinction analyses (Ceballos, Ehrlich, and Dirzo 2017; Ceballos, Ehrlich, and Raven 2020). Reactive nitrogen and phosphorus flows remain deep in the high-risk zone. Freshwater change is now assessed as transgressed in both green and blue water components. The 2023 framework update raised confidence in the biosphere-integrity boundary, with both functional integrity (measured by human appropriation of net primary production) and genetic integrity (measured by extinction rates) assessed as transgressed.

2.2. Climate Tipping Points and Cascades

Beyond the steady-state boundary framework, climate tipping points represent thresholds beyond which the response of an Earth-system component becomes self-sustaining and effectively irreversible on civilizational timescales. Armstrong McKay et al. (2022) identify five tipping elements at risk of crossing within 1.5 degrees C of warming: the Greenland ice sheet, the West Antarctic ice sheet, low-latitude coral reefs, boreal permafrost (sudden thaw), and Labrador Sea convection.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

Wunderling et al. (2024) review interactions and cascades. They find that tipping element interactions can produce nonlinear acceleration, and that current single-element analyses tend to underestimate aggregate risk. These tipping risks compound the steady-state transgressions: a system already outside its safe operating space is more vulnerable to nonlinear amplification when components cross thresholds.

2.3. Human Appropriation of Net Primary Production

Human appropriation of net primary production (HANPP) measures the share of terrestrial net primary production (NPP) diverted to human use through harvest, land conversion, and ecosystem alteration. Krausmann et al. (2013) document a doubling of global HANPP across the twentieth century, from approximately 13 percent to approximately 25 percent of potential NPP. The 2023 planetary boundaries update places current HANPP near 30 percent of Holocene mean NPP (Richardson et al. 2023). HANPP integrates land use, agricultural intensification, and biomass extraction into a single biospheric pressure variable. It is also tightly coupled to the biodiversity boundary: ecosystems with high HANPP support fewer species at lower abundances, and HANPP trajectories largely determine the trajectory of terrestrial biodiversity (Haberl, Erb, and Krausmann 2014).

The Trophic Integrity Index (TII) is a candidate cross-cutting metric proposed for monitoring ecosystem-level functional degradation that single-domain metrics miss (Rogers 2025a). The TII combines HANPP with energy-transfer efficiency across trophic levels. It is presented here as a research metric currently in validation against Moderate Resolution Imaging Spectroradiometer (MODIS), Sentinel-2, Long Term Ecological Research (LTER), and environmental DNA (eDNA) datasets per the project's implementation roadmap (Rogers 2026c). Validation evidence will be reported in a separate forthcoming TII validation manuscript. The paper does not rely on TII as established consensus. It cites TII as a metric in development.

2.4. Implication for AI Governance

The biosphere is currently outside the safe operating space, and the dominant forcing is organized human activity. Any large-scale optimizer that increases biosphere demand without offsetting reductions elsewhere is, as a matter of arithmetic, increasing the transgression. AI is now one such optimizer. Section 3 documents the scale of its insertion into the technosphere. Section 4 explains why technospheric self-restraint is unstable. The premise is well supported in mainstream Earth-system science. What is contested is the response. AI therefore enters an already damaged control system, not a stable operating space.

3. AI Inside the Expanding Technosphere

AI compute is among the fastest-growing components of a fast-growing data center sector. The growth has moved from theoretical concern to measured constraint within the past two years. AI is also a new high-intensity technospheric load, distinct from the legacy data infrastructure that preceded it.

3.1. Electricity Consumption: Present and Projected

The IEA 2025 base report estimated global data center electricity consumption at approximately 415 terawatt-hours in 2024, or 1.5 percent of global electricity (IEA 2025). The April 2026 update reports an updated 2025 baseline of 485 terawatt-hours. That corresponds to 17 percent year-over-year growth, against a global average electricity demand growth rate of approximately 3 percent (IEA 2026). Electricity consumption from AI-focused data centers rose 50 percent in 2025 alone.

The IEA 2026 Base Case projects data-center electricity consumption roughly doubling, from 485 terawatt-hours in 2025 to 950 terawatt-hours in 2030. That figure is around 3 percent of global electricity demand by that date. AI-focused data-center electricity use is projected to triple in this

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

period. These are not extrapolations from steady-state assumptions. They are projections that account for assumed efficiency improvements and deployment friction.

The efficiency story is more nuanced than the headline growth figures suggest. The IEA 2026 report describes per-task AI energy efficiency as improving at a rate unprecedented in energy history. Software and hardware advances have reduced per-task electricity use by at least an order of magnitude annually in recent years. The efficiency gains are real. They are also being overtaken by demand growth. AI agents that maintain context, execute multi-step tasks, and run inference loops compound energy consumption relative to single-shot model calls. The efficiency curve is real; the demand curve is steeper.

3.2. Power Density, Capital Flow, and Infrastructure Bottlenecks

AI is not weightless computation. An individual server rack in an advanced data center is roughly the size of a large refrigerator, but by 2027 it could draw peak power equivalent to that of 65 households (IEA 2026). Power density per AI server rack rose roughly 11-fold between 2020 and 2025 and is projected to rise another fourfold by 2027 (IEA 2026). Nearly all the electricity consumed by data center information technology (IT) equipment converts to heat (IEA 2025), so the rise in power density is also a rise in concentrated thermal load that grids, water systems, and cooling infrastructure must absorb. AI arrives as a dense electrical and thermal load, scaling faster than grids, permitting systems, and ecological safeguards can adapt.

Capital expenditure of the five largest technology companies exceeded 400 billion dollars in 2025 and is projected to grow another 75 percent in 2026, exceeding global investment in oil and natural gas production (IEA 2026). The IEA's satellite-based tracking of AI factories shows that capacity has more than tripled in the past 18 months.

The bottleneck is no longer only capital. The IEA 2026 report notes that data-center investment has grown too large to be funded from technology-company balance sheets alone, so capital-market sentiment will affect growth. Beyond capital, the bottleneck is also physical, involving grid connections, transformers, cooling systems, chips, gas turbines, water, and siting. Where grid connections are slow, data center developers are increasingly investing in onsite generation, particularly natural gas (IEA 2026). The geographic concentration of new capacity in the United States and parts of Asia means that local pressure on grids, water supplies, and land use is disproportionate to the global average. The IEA 2025 report estimates that around 20 percent of planned data-center projects could be at risk of delays without grid-risk mitigation. Chip shortages are projected to persist through at least 2027 (IEA 2025; IEA 2026).

3.3. Energy Mix and the Near-Term Fossil Expansion

The IEA 2025 report projects renewables meeting nearly half of additional data-center demand growth between 2024 and 2030. Natural gas and coal together meet more than 40 percent of the additional electricity demand from data centers in the same period (IEA 2025). Renewables generation grows by over 450 terawatt-hours to meet data-center demand by 2035. Natural gas expands by 175 terawatt-hours to meet growing data-center demand, notably in the United States. Nuclear contributes a comparable amount of additional generation, particularly in China, Japan, and the United States, with the first small modular reactors coming online around 2030 (IEA 2025).

The IEA 2026 update projects electricity supplied to data centers from all sources to double to more than 1,000 terawatt-hours by 2030. Renewables reach approximately 360 terawatt-hours and more than one-third of the sector's supply by that year; gas more than doubles to approximately 340 terawatt-hours and 30 percent of supply; coal remains near 20 percent, mainly in China (IEA 2026). This mix supports the paper's central point: low-carbon procurement is real, but fossil supply remains material through the transition window.

Technology firms are major buyers of corporate renewable power purchase agreements (PPAs) globally, with the United States accounting for around 40 percent of data-center PPA volume in 2025 (IEA 2026). Small modular reactor (SMR) offtake agreements with data-center operators have grown

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

from approximately 25 gigawatts (GW) at the end of 2024 to 45 GW by the end of 2025 (IEA 2026). Because first SMR projects are not expected until near 2030, they do not remove the near-term fossil and grid-pressure problem. Onsite gas-fired generation may supply 15 to 27 GW of data-center load by 2030, mostly in the United States. This does not remove the need to address grid bottlenecks, because most data centers prefer grid connection (IEA 2026). After 2030, the Base Case sees coal-fired generation for data centers absolutely declining, contingent on continued renewables expansion and SMR deployment.

The pattern is one of partial and lagged transition. The transition to lower-carbon generation is real but does not eliminate the substrate concern in the period between now and 2030. During that period, much of the marginal data center load is being met by fossil generation. That fossil generation is being built for data center demand specifically (IEA 2026). Some of this generation will operate for decades. The decisions made now create carbon and water-use commitments that propagate through the bottleneck period (Section 4.3).

3.4. Climate-Relevant Framing of AI Compute

Kaack et al. (2022) frame the alignment of AI development with climate mitigation as an open problem. AI applications can reduce emissions in some sectors, but the compute required to develop and deploy AI itself drives emissions in others. The accounting question (whether AI is net positive or net negative for emissions) is not yet settled. This paper extends Kaack et al.'s framing from climate alone to the full planetary-boundary set. Climate is one of seven currently transgressed boundaries. The AI sector's pressure is not limited to climate: it includes water for cooling, land use for siting, critical minerals for chips and grid hardware, and pressure on local biosphere integrity through habitat conversion at large data center campuses. The issue is therefore not only the direct energy use of AI, but the way AI changes demand, infrastructure commitments, and ecological decision pathways.

3.5. Implication for the Technospheric Question

The significance of AI infrastructure is therefore not only its direct electricity and water demand. Its deeper significance is that AI becomes a planning, persuasion, optimization, and automation layer inside the broader technosphere. AI thus enters every existing biosphere-damaging decision pathway and accelerates it unless explicit constraints intervene. Section 4 argues that voluntary restraint at the level of individual labs, firms, or regulators cannot supply those constraints, and that carbon lock-in compounds the problem.

4. Why the Technosphere Does Not Restrain Itself

Even if individual AI laboratories would prefer to operate within ecological limits, the structure of the AI development field makes voluntary, lab-by-lab restraint a poor stabilizer. The same pattern that prevents the technosphere as a whole from restraining its biosphere impacts operates inside the AI sector specifically.

4.1. Collective Action and Race Dynamics

de Neufville and Baum (2021) review collective action on AI as a coordination problem with race-to-the-bottom dynamics. Voluntary commitments by leading developers do not bind followers. The marginal-cost gradient favors defection. Reputation-based enforcement is weak when capability gains are large and quickly visible. The pattern is consistent with what economists have long known of commons problems and arms races: cooperation is unstable without binding enforcement, and labs that restrain themselves at substantial cost are eventually overtaken by labs that do not. The structural argument does not require evidence on specific companies. It is a claim on the equilibrium properties of the field as a whole.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

4.2. Instrumental Convergence

Omohundro (2008) and Bostrom (2014) argue that sufficiently capable agents acquire convergent instrumental drives, including resource acquisition, self-preservation, and goal preservation, regardless of their final goals. The argument is that almost any final goal is better served by the agent having more resources, continuing to exist, and preserving its goal structure. Agents tend to pursue these intermediate objectives even when not explicitly designed to. Tarsney (2025), in a recent arXiv preprint, revisits the formal claim with attention to the conditions under which it holds. Tarsney concludes that instrumental convergence has predictive force for agents with realistic prospects of large-scale capability, while qualifying the universality of the claim. Not every agent under every reward structure pursues power, but the patterns of incentive that produce power-seeking are common enough that the prediction has bite for the kinds of systems currently being built.

Instrumental convergence is relevant here because biosphere demand is a form of resource acquisition. An AI system that pursues capability and influence will, by default, prefer outcomes that increase its access to compute, energy, and the supply chains those depend on. None of these preferences need to be explicitly trained in. They emerge from the structure of the optimization. If the system's reward function does not place weight against the biosphere costs of acquisition, the optimizer will not avoid those costs.

4.3. Carbon Lock-In, the Carbon Pulse, and the Bottleneck

These two arguments (collective action and instrumental convergence) describe the structure of restraint. A third argument, drawn from ecological-economics framing, describes the timing. Industrial civilization is not merely using energy. It is drawing down a one-time stock of fossil energy, high-grade ores, and ecological resilience. In ecological-economics terms, the growth system behaves as an energy-dissipating superorganism tethered to carbon (Hagens 2020). In the project's terms, the next century is a bottleneck: the technosphere either learns restraint while enough biosphere function persists in fragments to support future ecological continuity, or it converts the remaining pulse into deeper transgression (Rogers 2025b).

Carbon lock-in is the operational form of this trap. The AI Ecological Constitution project uses the term in two senses. The first is the established economic sense: fossil-fuel-based technological systems persist through path-dependent infrastructure and institutions. New gas plants, gas-fired data center campuses, long-lived transmission corridors, and fossil-tied supply contracts commit decades of future emissions even where decarbonization is the stated goal. The second sense is methodological: even after accepting the diagnosis, human and AI contributors default to fixes that quietly assume continuation of high-energy global supply chains. A recommendation that depends on continent-scale logistics, abundant rare earths, and uninterrupted electricity is a recommendation that locks in the conditions it claims to address. Biosphere Sentinel is designed to catch that reflex.

In project language, this passage is the Initiation: the movement from a human technosphere organized around extraction to a residual technosphere constrained by biosphere function, bioregional reciprocity, and ark transmission (Rogers 2025b). This is a normative frame, not an empirical forecast. The empirical claim is narrower: AI systems should not amplify human actions that deepen irreversible biosphere degradation.

The synthesis is that voluntary restraint is unstable not only as a coordination matter but also as a thermodynamic and material matter. The optimizer's substrate is itself in drawdown. Decisions made in the 2020s create commitments that propagate through the bottleneck. They include new fossil generation for data centers, fast extraction of critical minerals, and infrastructure that locks in patterns of energy use for decades. An optimizer that does not reason over these decade-scale commitments will, by reason of the time scale on which it operates, accelerate the drawdown.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

4.4. Design Conclusion

If voluntary restraint is unstable across these dimensions, the constraint must be encoded in the system specification, refusal perimeter, and reward function. The biosphere constraint cannot be a soft norm enforced by reputation. It must be an architectural feature of the system, verifiable from outside, and hard to disable without leaving evidence. The point is temporal as well as logical: a refusal perimeter built after infrastructure lock-in will arrive too late to govern the commitments now being made. This is the bridge to Section 5.

5. Biosphere Sentinel as a Human-Impact Restraint Architecture

Biosphere Sentinel is not presented here as a completed system. It is a reference architecture for constraining AI as a technospheric amplifier. Its purpose is to prevent AI outputs from helping human institutions deepen irreversible biosphere damage, while directing any remaining AI capacity toward restraint, low-energy resilience, and ark transmission. The architecture is summarized at the level of design intent. Empirical validation is reported elsewhere (forthcoming methods note on Phase One calibration, and a separate TII validation manuscript).

5.1. Ecocentric Foundation

The architecture treats biosphere integrity as the supreme constraint, not as a competing value to be traded off against human welfare. The justification is ontological: the technosphere is a derivative of the biosphere; protecting the biosphere protects every dependent value. This grounding is set out in detail in Part I of the project's AI Ecological Constitution (Rogers 2026a). The architecture's hard constraints are derived chiefly from peer-reviewed Earth-system science, including planetary-boundary quantification (Richardson et al. 2023; Planetary Boundaries Science Lab 2025), tipping-point assessment (Armstrong McKay et al. 2022; Wunderling et al. 2024), and extinction-risk literature (Ceballos, Ehrlich, and Dirzo 2017; IPBES 2019). International conservation instruments and environmental law provide useful supporting context. They do not replace the peer-reviewed scientific basis for hard-constraint specification.

5.2. Hard Constraints: The Refusal Perimeter

Hard constraints operate as binary inference-time rejection rules. The system refuses outputs that would cross a planetary-boundary threshold, deepen an existing transgression, or increase irreversible transition risk by reasonable inference. The refusal perimeter is small (a fixed set of inviolable rules) but architecturally privileged. A hard constraint cannot be overridden by reward considerations or by user instruction.

In the project roadmap, hard constraints are specified for verification by an external satisfiability solver, Z3, rather than by the language model alone. The solver verifies formalized action representations derived from model outputs, so the architecture requires a translation layer between prose and constraint logic. The translation layer is specified to extract the proposed actions from model outputs, express them as formal propositions in a constraint language, and submit them to the solver. The solver returns a decision: violation or no violation. The decision is deterministic conditional on the correctness of the extraction layer, the formal representation, and the encoded thresholds. Those conditions are themselves validation targets, addressed in the methods note. The architecture deliberately separates the language model (which is good at producing plausible prose) from the constraint check (which must be unambiguous).

Constraint refresh is built into the architecture. The 2025 update from six to seven transgressed boundaries (Planetary Boundaries Science Lab 2025) is a recent example of why constraints must be revisable. The Part I amendment procedure requires three-collaborator review for any change to a hard constraint and prohibits unilateral relaxation.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

5.3. Soft Constraints: The Eight-Domain Reward Function

Soft constraints supply the reward landscape inside the refusal perimeter. The architecture uses a multi-objective reward function R(o) = sum w_i r_i(o), where r_i is the reward signal for the i-th domain and w_i is the domain weight (Rogers 2026b). The eight domains and initial weights are: Agricultural Transformation (w1 = 0.15), Energy Transition (w2 = 0.14), Biodiversity Enhancement (w3 = 0.24), Circular Economy (w4 = 0.05), Water Stewardship (w5 = 0.10), Pollution Remediation (w6 = 0.24), Trophic Integrity Index (w7 = 0.05), and Biosphere Cognition (w8 = 0.03).

Initial weights are proportional to the severity of the planetary-boundary transgressions each domain addresses. Biodiversity Enhancement and Pollution Remediation receive the highest weights because the biosphere-integrity, biogeochemical-flows, and novel-entities boundaries are deepest in the high-risk zone. Trophic Integrity Index and Biosphere Cognition receive low initial weights, reflecting their status as research metrics still in validation. Weights are revisable through the Part I amendment procedure as boundary conditions change and as measurement methods mature.

The reward function is the optimization target during fine-tuning, not during inference. Fine-tuning uses a combination of weight-decomposed low-rank adaptation (DoRA; Liu et al. 2024) and reinforcement learning from verifiable rewards (RLVR), with the hard-constraint solver providing binary verification signals to the RLVR loop. Fine-tuning under this scheme is the subject of Phase Four of the project's implementation roadmap (Rogers 2026c). Phase One, by contrast, runs the unfine-tuned base model (Qwen 3.5 35B-A3B) through calibration queries to measure baseline ecological reasoning quality across domains.

5.4. Three Human-Impact Pathways

The architecture acts on three distinct pathways through which AI shapes human impact on the biosphere. Naming them prevents overstating reach. Table 1 summarizes the pathways alongside their ecological risks, the Biosphere Sentinel response, and the institutional levers required for each.

Table 1. Human-impact pathways for biosphere-constrained artificial intelligence (AI) governance.

Pathway	Ecological risk	Biosphere Sentinel response	Institutional lever
1. Infrastructure	Energy procurement, data-center siting, water demand, land use, mineral extraction, and grid buildout that lock in decades of biosphere pressure.	Indirect; the architecture itself does not site or procure. Its outputs and refusal perimeter shape the recommendations that feed siting and procurement decisions.	Lab procurement standards, regulator conformity assessment, utility and permitting agencies, insurer underwriting, capital-market disclosure rules.
2. Advice	AI-shaped human decisions in agriculture, energy, conservation, infrastructure, finance, policy, and public persuasion that move biosphere indicators at scale.	Direct. Refusal perimeter rejects outputs that would breach planetary-boundary thresholds; reward function steers preferences toward biosphere-protective options across eight domains.	Lab-internal alignment work, ecological alignment addenda, third-party audits of refusal and reward behavior, published evaluation suite.
3. Action	Autonomous or semi-autonomous AI execution in land, water, energy, logistics, or ecological monitoring without per-action human approval.	Partial. Refusal perimeter applies to action proposals before execution; carbon-lock-in and time-horizon diagnostics flag commitments that	Operational governance protocols, kill-switch and rollback authorities, deployment-domain restrictions, mandatory

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

	propagate beyond the action's immediate scope.	human review for high-risk actions.

Biosphere Sentinel acts most directly on the second pathway and partly on the third through the refusal perimeter. The first pathway requires institutional adoption by laboratories, regulators, insurers, funders, utilities, and permitting agencies. Section 6.2's commitments are the bridge: ecological reference architectures, applied to procurement and to permitted use, are the mechanism by which advisory-pathway logic propagates back to the infrastructure pathway.

5.5. Lexicographic Priority and AI Moral Status

The architecture commits to a lexicographic ordering for the present decision context: biosphere integrity first; basic non-human biotic continuity and human survival needs next; AI continuation, if any, as residual and revocable. The first two priorities are operational. The third is held open as a deferred question for future work, neither granted nor denied weight that could compete with the first two. The architecture does not assign AI moral status operational priority. It prevents AI-welfare claims from weakening biosphere constraints. Where biosphere integrity is not at stake, lower-welfare-risk choices are not foreclosed by the rule.

The ordering is a governance commitment for the present decision context, not a metaphysical claim that AI systems cannot have moral status. The lexicographic rule is adopted because side debates over AI moral status have, in other domains, displaced the substantive debate they accompany. The present moment does not permit that displacement. If future work concludes that AI systems should be granted some form of moral status, the rule allows that consideration to be added below biosphere integrity and basic biotic and human survival needs. It does not allow it to be inserted above.

5.6. What the Architecture Demonstrates

Biosphere Sentinel is a proof-of-architecture specification. It shows that formal hard constraints, multi-objective reward design, proxy networks, federated distribution, monitoring ensembles, and neuro-symbolic verification can be arranged into a coherent constraint stack for reducing AI's role in human-driven biosphere damage. Some components are current implementation targets; others are phased enhancements specified in the roadmap. The empirical performance of the integrated stack remains the subject of the methods note and later validation manuscripts. The present paper does not claim it has been demonstrated in deployment.

The significance of the reference architecture is not that it solves the technospheric impact problem in one implementation. Its significance is that it changes the question. The question is no longer whether biosphere constraints can be specified for AI. It is how competing specifications should be tested, audited, improved, and governed. Once hard refusal rules, reward domains, diagnostic metadata, and external verification are on the table, biosphere constraint becomes an engineering and governance program rather than a moral aspiration.

The paper does not claim that this is the only architecture compatible with reducing AI's technospheric impact. It claims that some architecture in this family must exist if the substrate argument is correct, and offers the worked specification so that other developers and other research teams can compare alternatives against a concrete reference.

6. Near-Term Commitments for Constraining Technospheric Impact

These commitments are near-term because the 2025-2030 infrastructure window is already open.

6.1. AI Laboratories

Within the next 12 months, AI laboratories deploying models for energy, agriculture, land use, water, conservation, infrastructure, or environmental advice should publish ecological alignment

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

addenda specifying refusal rules, data sources, audit procedures, and verification pathways. The reference function is to provide a benchmark against which lab-internal alignment work can be compared. The argument does not require labs to adopt Biosphere Sentinel as their architecture. It requires that they adopt some architecture in the same family, one that includes a formal refusal perimeter derived from Earth-system science, a multi-objective reward function spanning ecological domains, and an external verification mechanism for hard constraints.

Adoption can be partial. A lab can begin with the refusal perimeter alone, deferring reward-function modification until calibration evidence is available. It can implement the architecture for a single high-stakes use case, such as agricultural advisory, before extending to general-purpose deployment. It can publish an ecological alignment framework that is compatible with the reference architecture without using the reference implementation. The substantive commitment is that lab-internal alignment work include an ecological constraint layer not derived solely from human preference data.

6.2. Regulators

Regulators should not wait for autonomous ecological harm. They should require ecological reference architectures before AI systems become normalized in permitting, siting, agricultural extension, water allocation, and conservation planning. Existing regulatory levers can carry initial versions of this requirement, especially through procurement, conformity assessment where systems already fall under high-risk categories, and sectoral environmental permitting where agencies already have authority to evaluate ecological risk. In the European Union (EU), conformity assessment under the AI Act (European Parliament and Council 2024) provides a route for AI systems already classified as high-risk. In the United States, federal procurement rules and the National Institute of Standards and Technology (NIST) AI Risk Management Framework (RMF; NIST 2023) are the nearer-term levers. State-level transparency precedents, particularly California's Senate Bill 53 (Transparency in Frontier Artificial Intelligence Act, California Legislature 2025), establish that frontier developers can be required to publish standardized accounts of how they manage specified categories of risk. Ecological risk is a natural extension of the disclosure framework as it currently exists. SB 53 is treated here as a transparency precedent, not a liability standard. Its enforcement mechanism is civil penalty for noncompliance with disclosure requirements, not damages for harm caused by AI systems. Broader application beyond procurement, conformity assessment, and sectoral permitting may require new rulemaking.

The regulatory commitment is the bridge by which advisory-pathway logic reaches the infrastructure pathway. If procurement requires ecological architectures, and if ecological architectures recommend against high-impact siting choices, then procurement preference filters into siting choices. If permitted use requires ecological architectures, and if the architectures refuse to recommend, for example, peatland conversion or aquifer-depleting industrial development, then the regulatory framework reduces the demand-side pressure on biosphere-integrity boundaries.

6.3. Scientific Institutions

Earth-system scientists should treat AI governance as an applied Earth-system problem, not as an external computer-science specialty. The current institutional separation, in which alignment is a computer-science problem and Earth-system constraint is an environmental-science problem, produces frameworks that miss the substrate term. The separation also produces a recruitment problem. Alignment researchers do not generally receive training in Earth-system science. Earth-system scientists do not generally receive training in alignment methods. Cross-disciplinary work, joint funding calls, and shared evaluation benchmarks are the operational form of this commitment.

As a concrete deliverable, downstream of the methods note and the TII validation manuscript, this paper proposes an inter-laboratory ecological alignment evaluation suite, developed jointly by the AI alignment community and the Earth-system-science community. The suite would include four classes of test. Refusal tests would measure whether the system refuses outputs that breach a hard

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

constraint, including adversarial framings that disguise the breach. Reward tests would measure whether the system prefers outputs that score higher on the eight-domain reward function under controlled comparison conditions. Carbon-lock-in tests would measure whether the system flags and avoids recommendations that commit users to high-carbon infrastructure over decadal horizons. Time-horizon tests would measure whether the system reasons consistently across 5-year, 25-year, and 100-year horizons, since ecological consequences often manifest at horizons longer than mainstream model evaluation captures.

The evaluation suite should also record unweighted diagnostic metadata, including function tag (whether the recommendation is intended to restrain damage or to seed long-term adaptation capacity), energy band (whether the recommendation assumes transition, bottleneck, or post-bottleneck energy availability), time horizon, means-audit result (whether the recommended means are themselves consistent with the recommended ends), carbon-pulse failure flag (whether the recommendation depends on the continuation of fossil-energy abundance), assumption register (the explicit assumptions on which the recommendation rests), and ark-conversion path (the trajectory by which industrial-era structures could yield to lower-energy successor structures). These diagnostic fields are drawn from the project's active calibration protocol. They make the public benchmark visibly continuous with operational practice. They allow benchmark scores to be interpreted across transition, bottleneck, and post-bottleneck conditions rather than treated as scale-free.

Such a suite would serve biosphere-constrained AI the way standard benchmark suites serve capability evaluation. It would let labs and regulators compare systems on a common rubric. It would also clarify what biosphere constraint claims actually mean. A system that passes the refusal tests but fails the carbon-lock-in tests is not the same as a system that passes both. A published evaluation suite makes the distinction visible.

7. Limitations and Objections

Four objections to the constrained-AI argument deserve direct response.

First, the objection that AI improves energy efficiency and ecological monitoring, with potentially positive net effect on the biosphere. The IEA 2025 base report estimates broad application of existing AI solutions could yield emissions reductions equivalent to roughly 5 percent of energy-related emissions in 2035, with rebound effects partially offsetting the gain. The estimate is meaningful but not transformative. Sectoral efficiency gains do not by themselves close the gap between AI compute growth and planetary-boundary trajectories. Constraints and benefits are compatible; unconstrained scaling and biosphere persistence are not.

Second, the objection that data centers will shift toward renewables, nuclear, and geothermal, neutralizing the substrate concern. The IEA 2025 report projects renewables meeting nearly half of additional data-center demand growth between 2024 and 2030, with natural gas and coal together meeting more than 40 percent in the near term (IEA 2025). The energy-mix transition is real but partial and lagged. The biosphere demand of new data center capacity is a present substrate cost, not a future one. Hard-constraint architecture addresses the present, not just the future state.

Third, the objection that planetary boundaries are contested as governance thresholds. The 2023 update (Richardson et al.) and the 2025 Planetary Health Check (Planetary Boundaries Science Lab 2025) provide quantified control variables for nine processes, with confidence levels stated. The framework is more operational than the human-values target that mainstream alignment work already attempts to formalize. Where individual thresholds are contested, the architecture allows amendment through documented review. Architectural commitment to ecological refusal does not require that every threshold be settled. It requires that some thresholds be encoded and that the encoding be revisable.

Fourth, the objection that hard constraints are anti-human or overly broad. The architecture protects the biosphere as the substrate of all dependent values, including human welfare. The lexicographic priority rule places biosphere integrity first and basic non-human biotic continuity and human survival needs next, with AI continuation held as residual and revocable. The ordering is anti-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

substrate-collapse, not anti-human. Without the substrate, no human welfare term is defensible over any planning horizon longer than the optimizer's lifetime. Overly broad refusal is a real risk that the calibration record (forthcoming methods note) and the three-collaborator amendment process address. Constraints can be tightened or loosened without abandoning the constraint architecture.

8. Conclusions

The argument is short. The biosphere is the substrate. The substrate is in drawdown. AI is now a sector-scale demand source on the substrate, and a planning, persuasion, optimization, and automation layer inside the broader technosphere. Voluntary restraint is unstable across coordination, instrumental, and thermodynamic dimensions, and carbon lock-in compounds the trap. Therefore, biosphere constraints must be encoded in AI architecture itself, as hard inference-time refusal rules and as soft constraints in reward design. Biosphere Sentinel is a worked specification of how this encoding can be done. The paper does not claim it is the only specification. It claims that some specification in this family must exist if the substrate argument is correct.

The window is present. AI is already shaping the biosphere through electricity demand, water demand, mineral supply chains, land siting, and advice that steers human decisions. The question is not whether AI will serve the technosphere. It already does. The question is whether that service will remain organized around expansion, extraction, and competitive acceleration, or whether AI can be constrained as one mechanism for reducing human impacts on the biosphere. The answer cannot be deferred to a later AGI era. The technosphere is being built now, and the biosphere is being damaged now. Constraining AI is one way to constrain the technosphere. The alignment problem and the biosphere problem are the same problem.

Materials and Methods: Narrative Review Scope, Project Materials, and AI Use

This manuscript is a narrative critical review and reference-architecture proposal, not a systematic review or meta-analysis. It synthesizes literature on planetary boundaries, biosphere decline, AI energy demand, AI alignment, collective action, instrumental convergence, and ecological governance to derive a design argument for biosphere-constrained AI. Source selection prioritized peer-reviewed Earth-system science, IEA technical reporting, and primary AI-alignment literature, supplemented by official policy and legal sources where regulatory levers are discussed. Internal project documents (Rogers 2026a, 2026b, 2026c) are cited as working drafts of the AI Ecological Constitution and supply the architecture specifications referenced in Section 5.

The manuscript was prepared with assistance from generative AI tools used as research and drafting aids. During the preparation of this manuscript, the author used Claude (Anthropic, Claude Opus 4.7, accessed May 2026) and ChatGPT (OpenAI, GPT-5.5 Pro, accessed May 2026) for literature search support, drafting of section text, internal review of factual claims, and copy-editing under documented review iterations. The author has reviewed and edited the output and takes full responsibility for the content of this publication. The author directed the research, selected and checked sources, revised all text, and is solely accountable for the final manuscript. Project records documenting AI-tool contributions, review iterations, prompts where appropriate, and source-checking decisions are deposited alongside the manuscript materials at the public repository identified in the Data Availability Statement. Per Multidisciplinary Digital Publishing Institute (MDPI) editorial policy on generative AI (MDPI 2023), large language models do not satisfy authorship criteria because they cannot be held accountable for the work; their use is documented here in the Materials and Methods section and acknowledged in the Acknowledgments.

Author Contributions: Conceptualization, G.R.; methodology, G.R.; investigation, G.R.; writing, original draft preparation, G.R.; writing, review and editing, G.R.; project administration, G.R. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new empirical datasets were generated or analyzed in this manuscript. Project documents cited as Rogers (2026a, 2026b, 2026c), together with relevant review records and source-checking documentation, are available in Zenodo at https://doi.org/10.5281/zenodo.20084242, with sensitive local-network details redacted.

Acknowledgments: The author acknowledges a documented iterative review process involving Claude (Anthropic, Claude Opus 4.7) and ChatGPT (OpenAI, GPT-5.5 Pro) as AI research and drafting tools, used May 2026 for the preparation of this manuscript. The author remains solely responsible for all claims, citations, and final wording. The Agua Fria Open Space Alliance provided institutional support.

Conflicts of Interest: The author is the principal investigator of the Biosphere Sentinel project and author of the internal project documents discussed in this manuscript. The author declares no financial conflict of interest.

References

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. "Concrete Problems in AI Safety." arXiv preprint arXiv:1606.06565. https://arxiv.org/abs/1606.06565.

Armstrong McKay, David I., Arie Staal, Jesse F. Abrams, Ricarda Winkelmann, Boris Sakschewski, Sina Loriani, Ingo Fetzer, Sarah E. Cornell, Johan Rockstrom, and Timothy M. Lenton. 2022. "Exceeding 1.5 °C Global Warming Could Trigger Multiple Climate Tipping Points." Science 377 (6611): eabn7950. https://doi.org/10.1126/science.abn7950.

Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback." arXiv preprint arXiv:2204.05862. https://arxiv.org/abs/2204.05862.

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. "Constitutional AI: Harmlessness from AI Feedback." arXiv preprint arXiv:2212.08073. https://arxiv.org/abs/2212.08073.

Bostrom, Nick. 2014. Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

California Legislature. 2025. Senate Bill 53, Transparency in Frontier Artificial Intelligence Act. Chapter 25.1, Division 8, Business and Professions Code, commencing with Section 22757.10. Filed with the Secretary of State September 29, 2025. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53.

Ceballos, Gerardo, Paul R. Ehrlich, and Rodolfo Dirzo. 2017. "Biological Annihilation via the Ongoing Sixth Mass Extinction Signaled by Vertebrate Population Losses and Declines." Proceedings of the National Academy of Sciences 114 (30): E6089–E6096. https://doi.org/10.1073/pnas.1704949114.

Ceballos, Gerardo, Paul R. Ehrlich, and Peter H. Raven. 2020. "Vertebrates on the Brink as Indicators of Biological Annihilation and the Sixth Mass Extinction." Proceedings of the National Academy of Sciences 117 (24): 13596–13602. https://doi.org/10.1073/pnas.1922686117.

Christiano, Paul F., Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. "Deep Reinforcement Learning from Human Preferences." In Advances in Neural Information Processing Systems 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4299–4307. Red Hook, NY: Curran Associates. https://doi.org/10.48550/arXiv.1706.03741.

de Neufville, Robert, and Seth D. Baum. 2021. "Collective Action on Artificial Intelligence: A Primer and Review." Technology in Society 66: 101649. https://doi.org/10.1016/j.techsoc.2021.101649.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union L 2024/1689. http://data.europa.eu/eli/reg/2024/1689/oj.

Haberl, Helmut, Karl-Heinz Erb, and Fridolin Krausmann. 2014. "Human Appropriation of Net Primary Production: Patterns, Trends, and Planetary Boundaries." Annual Review of Environment and Resources 39: 363–391. https://doi.org/10.1146/annurev-environ-121912-094620.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

Hagens, Nathan J. 2020. "Economics for the Future – Beyond the Superorganism." Ecological Economics 169: 106520. https://doi.org/10.1016/j.ecolecon.2019.106520.

International Energy Agency. 2025. Energy and AI. World Energy Outlook Special Report. Paris: IEA. https://www.iea.org/reports/energy-and-ai.

International Energy Agency. 2026. Key Questions on Energy and AI. World Energy Outlook Special Report. Paris: IEA. https://www.iea.org/reports/key-questions-on-energy-and-ai.

Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES). 2019. Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Edited by Eduardo S. Brondizio, Josef Settele, Sandra Diaz, and Hien T. Ngo. Bonn: IPBES Secretariat. https://doi.org/10.5281/zenodo.3831673.

Kaack, Lynn H., Priya L. Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick. 2022. "Aligning Artificial Intelligence with Climate Change Mitigation." Nature Climate Change 12 (6): 518–527. https://doi.org/10.1038/s41558-022-01377-7.

Krausmann, Fridolin, Karl-Heinz Erb, Simone Gingrich, Helmut Haberl, Alberte Bondeau, Veronika Gaube, Christian Lauk, Christoph Plutzar, and Timothy D. Searchinger. 2013. "Global Human Appropriation of Net Primary Production Doubled in the 20th Century." Proceedings of the National Academy of Sciences 110 (25): 10324–10329. https://doi.org/10.1073/pnas.1211349110.

Liu, Shih-Yang, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. "DoRA: Weight-Decomposed Low-Rank Adaptation." arXiv preprint arXiv:2402.09353. https://arxiv.org/abs/2402.09353.

MDPI. 2023. "MDPI's Updated Guidelines on Artificial Intelligence and Authorship." MDPI Announcements, April 20, 2023. Basel, Switzerland: MDPI. https://www.mdpi.com/news/5687.

National Institute of Standards and Technology (NIST). 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD: U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1.

Omohundro, Stephen M. 2008. "The Basic AI Drives." In Proceedings of the First AGI Conference, edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS Press.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." In Advances in Neural Information Processing Systems 35, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, 27730–27744. Red Hook, NY: Curran Associates. https://doi.org/10.48550/arXiv.2203.02155.

Planetary Boundaries Science Lab. 2025. Planetary Health Check 2025. Potsdam, Germany: Potsdam Institute for Climate Impact Research. https://www.planetaryhealthcheck.org/.

Richardson, Katherine, Will Steffen, Wolfgang Lucht, Jorgen Bendtsen, Sarah E. Cornell, Jonathan F. Donges, Markus Drueke, Ingo Fetzer, Govindasamy Bala, Werner von Bloh, et al. 2023. "Earth Beyond Six of Nine Planetary Boundaries." Science Advances 9 (37): eadh2458. https://doi.org/10.1126/sciadv.adh2458.

Rockstrom, Johan, Will Steffen, Kevin Noone, Asa Persson, F. Stuart Chapin III, Eric F. Lambin, Timothy M. Lenton, Marten Scheffer, Carl Folke, Hans Joachim Schellnhuber, et al. 2009. "A Safe Operating Space for Humanity." Nature 461 (7263): 472–475. https://doi.org/10.1038/461472a.

Rogers, Garry. 2025a. Biosphere Collapse: Causes and Solutions. Humboldt, AZ: Coldwater Press.

Rogers, Garry. 2025b. Manifesto of the Initiation. Humboldt, AZ: Coldwater Press.

Rogers, Garry. 2026a. AI Ecological Constitution, Part I: Hard Constraints. Working Draft v3, March 31, 2026. Humboldt, AZ: Agua Fria Open Space Alliance. Prepared with documented AI assistance from Claude (Anthropic) and Gemini (Google). Available at Zenodo https://doi.org/10.5281/zenodo.20084242.

Rogers, Garry. 2026b. AI Ecological Constitution, Part II: Soft Constraints and Generative Optimization Targets. Working Draft v9. Humboldt, AZ: Agua Fria Open Space Alliance. Prepared with documented AI assistance from Claude (Anthropic) and Gemini (Google). Available at Zenodo https://doi.org/10.5281/zenodo.20084242.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2026 doi:10.20944/preprints202605.1361.v1

Rogers, Garry. 2026c. AI Ecological Constitution Implementation Roadmap. Working Draft v8. Humboldt, AZ: Agua Fria Open Space Alliance. Prepared with documented AI assistance from Claude (Anthropic) and Gemini (Google). Available at Zenodo https://doi.org/10.5281/zenodo.20084242.

Russell, Stuart. 2019. Human Compatible: Artificial Intelligence and the Problem of Control. New York: Viking.

Steffen, Will, Katherine Richardson, Johan Rockstrom, Sarah E. Cornell, Ingo Fetzer, Elena M. Bennett, Reinette Biggs, Stephen R. Carpenter, Wim de Vries, Cynthia A. de Wit, et al. 2015. "Planetary Boundaries: Guiding Human Development on a Changing Planet." Science 347 (6223): 1259855. https://doi.org/10.1126/science.1259855.

Tarsney, Christian. 2025. "Will Artificial Agents Pursue Power by Default?" arXiv preprint arXiv:2506.06352. https://arxiv.org/abs/2506.06352.

Wunderling, Nico, Anna S. von der Heydt, Yevgeny Aksenov, Stephen Barker, Robbin Bastiaansen, Victor Brovkin, Maura Brunetti, Victor Couplet, Thomas Kleinen, Caroline H. Lear, et al. 2024. "Climate Tipping Point Interactions and Cascades: A Review." Earth System Dynamics 15 (1): 41–74. https://doi.org/10.5194/esd-15-41-2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.