

Reckoning with Ecology, Earth Systems, and Resource-Environmental Constraints in the Age of AGI

Compute Expansion, Energy and Resources, and the Material Boundaries of Civilization

Reckoning with Ecology, Earth Systems, and Resource-Environmental Constraints in the Age of AGI

Compute Expansion, Energy and Resources, and the Material Boundaries of Civilization

Ecology, Earth Systems, and Resources and Environment: Standalone In-depth Professional Paper for External Circulation V1.1

Author: Zijun Fu | Produced by: Civilization Leap Research Group

Evidence cutoff: August 26, 2026

Abstract translated from the Chinese

Artificial intelligence is often described as a digital capability that can be replicated almost without limit. Yet every real-world AI system depends on chips, servers, data centers, power grids, cooling, water, land, materials, communications networks, construction, and equipment replacement and disposal. What truly calls for a renewed reckoning in the AGI era is not the single question of whether AI consumes electricity, but how computational demand becomes coupled to actual Earth systems over the long term as intelligent capabilities gradually become shared infrastructure for research, production, education, health care, governance, and security: whether gains in efficiency per unit of computation can offset expansion in aggregate demand; whether loads whose global share appears limited create high pressure on local grids and watersheds; whether reducing consumption of one resource increases another environmental cost; how equipment replacement, materials, and critical minerals enter life-cycle impacts; and whether AI itself can, in turn, reduce environmental pressure through grid dispatch, materials discovery, and climate and energy optimization.

Drawing on the established disciplinary foundations of ecology, Earth-system science, energy and power systems, water resources, industrial ecology, life-cycle assessment, material flows, and resources and environment research, this paper establishes a strict separation among levels of argument. Category A comprises verifiable data-center electricity use, regional loads, water use, official statistics, life-cycle methods, and facts about real infrastructure. Category B comprises competing judgments that still depend on scenarios and assumptions, including future AI energy use, water use, carbon emissions, rebound effects, hardware lifetimes, and emissions-reduction potential. Category C comprises cross-domain causal chains such as “expanding AI capabilities → expansion of compute and infrastructure → energy/water/material pressures → regional ecological and resource consequences → price, employment, and governance responses → changes that feed back into AI expansion.” Category F explicitly undertakes the normative commitment that “every civilizational vision of future intelligence abundance must withstand the test of the real material world.” This paper does not automatically turn Category A resource facts into an argument against growth, nor permit ecology to determine ultimate value priorities on behalf of economics, political science, law, or civilizational direction.

As of 2026, the International Energy Agency's updated assessment suggests that global data-center electricity consumption could rise from approximately 485 TWh in 2025 to approximately 950 TWh in 2030, accounting for around 3% of global electricity consumption at that time, with faster growth in AI-dedicated data centers. A

bottom-up equipment model from Lawrence Berkeley National Laboratory in the United States gives a reference scenario of 649 TWh for US data-center electricity consumption in 2030, approximately 11.8% of total US electricity consumption, while explicitly providing a broad uncertainty range.[1–3] These figures are not certain predictions of civilization's future. They demonstrate a methodological fact: AI's environmental impacts must be tied to time, place, system boundaries, the energy mix, equipment, and utilization. Water-resource research shows that water consumption for the same computational workload can vary by more than four orders of magnitude because of server efficiency, grid water intensity, utilization, cooling technology, climate, and other factors. Recent integrative research also indicates that reducing water consumption and lowering carbon emissions may involve trade-offs under some conditions.[6,10] A single figure for “water consumed per prompt” or “electricity consumed per model” therefore cannot sustain a complete ecological judgment.

The paper ultimately argues that intelligence abundance cannot be established merely by a decline in the marginal cost of computation. It must also pass real-world tests of electricity availability, local grid capacity, water and ecological sustainability, materials and supply chains, life cycles, and recoverability. At the same time, environmental constraints must not be misrepresented as a requirement that civilization halt technological progress. The real question is whether AI development can move from “greater scale” toward “greater real utility per unit of environmental cost, stronger regional adaptation, longer hardware life cycles, greater resource circularity, more dispatchable loads, and net improvements to energy and ecological systems themselves.” This is a refutable and revisable professional research agenda, not a predetermined verdict on civilizational direction issued in the name of ecology.

Keywords: AGI era; ecology; Earth system; data centers; energy systems; water resources; critical minerals; life-cycle assessment; rebound effect; civilizational carrying capacity

Original English Abstract

Artificial intelligence is often described as a digital capability that can be replicated at very low marginal cost. Yet every real deployment depends on chips, servers, data centres, electricity grids, cooling systems, water, land, materials, networks, construction, and equipment turnover. The relevant environmental question in the age of AGI is therefore not simply whether AI consumes electricity. It is how expanding computation becomes coupled to real Earth systems once advanced AI functions as shared infrastructure across science, production, education, health care, government, and security.

This paper draws on ecology, Earth-system science, energy and power systems, hydrology, industrial ecology, life-cycle assessment, material-flow analysis, and resource-environment research. It distinguishes established or verifiable facts and formal data (A), unsettled mechanisms and scenario-dependent estimates (B), cross-domain structural hypotheses (C), and the author's explicit normative commitments (F). It argues that unit efficiency cannot be equated with lower aggregate resource use; a small global electricity share can coexist with severe local grid or watershed constraints; reducing water, carbon, land, or material impacts can involve trade-offs; and operational energy alone is insufficient for assessing the life-cycle footprint of AI.

The paper treats “intelligence abundance” as a physical-ecological proposition as well as an economic or technical one. Any claim that advanced AI can create abundance must remain answerable to electricity availability, grid integration, water and ecological conditions, material supply, embodied impacts, equipment turnover, and recoverability. At the same time, environmental constraints do not themselves determine the final social or political choice. AI may also improve power-system flexibility, energy innovation, material discovery, climate analysis, and resource efficiency. The objective is therefore not to oppose technological expansion, but to identify the conditions under which intelligence can scale without externalizing irreversible material and ecological costs.

Keywords: AGI era; ecology; Earth system; data centres; energy systems; water resources; critical minerals; life-cycle assessment; rebound effect; civilizational carrying capacity

Research Position, Object Boundaries, and Discipline of Argument

This is an in-depth professional paper on ecology, Earth systems, and resources and environment. Its independent responsibility is to examine how energy, grids, water, land, materials, critical minerals, manufacturing, emissions, waste, regional ecology, and Earth-system feedbacks are measured, explained, and constrained once AI and data centers enter the real material world. It treats neither “compute” as a purely abstract quantity nor environmental problems as merely a carbon-emissions inventory. Its minimum object of study is a material chain: what hardware performs a computational task; where that hardware operates; what electricity and cooling it uses; where its electricity and water originate; how equipment is manufactured and replaced; what land and materials infrastructure occupies or consumes; how it is disposed of; and which watersheds, grids, ecosystems, and global environmental processes receive these local pressures.

This paper does not judge chip, model, software, or system performance on behalf of AI and computer science; decide prices, investment, ownership, or resource allocation on behalf of economics; decide who governs electricity grids, water, land, and minerals on behalf of political science and law; determine the ultimate norms of intergenerational responsibility and environmental justice on behalf of ethics; or decide “where development should lead” on behalf of the paper on civilizational direction. The professional responsibility of ecology and resources and environment science is to establish real material conditions up to the point at which other disciplines can re-enter and assess them.

Table 1. A/B/C/F argument categories and the boundaries of ecological expertise

Category	Typical content in this paper	What it can support	What it cannot automatically establish
A Relatively stable facts / formal materials	Measured/statistical data-center electricity use, grid structure, water use and WUE, official resource statistics, established LCA methods	Factual judgments within specified locations, times, technologies, and system boundaries	AI's overall net environmental benefit or the necessary correctness of a policy
B Frontier questions / competing explanations	Future AI energy and water use, equipment lifetimes, rebound effects, net emissions-reduction potential, material substitution	Comparisons of scenarios, mechanisms, and external validity	A converged global conclusion
C Cross-domain structural judgments	Compute expansion → resource pressure → prices/governance → feedback changing AI's trajectory; AI optimization → resource efficiency → lower environmental pressure	Refutable cross-domain causal hypotheses	A complete causal loop already directly established by a single discipline
F Normative commitments	All “intelligence abundance” must withstand material-world testing; irreversible ecological damage cannot become invisible because of digital benefits	Why this paper requires complete boundaries and actual carrying capacity	A substitute for final economic, legal, political, or civilizational value priorities

The “AGI Era” Does Not Await Fully Realized AGI

This paper uses “the AGI era” to describe an infrastructural condition already emerging in advance: advanced AI is embedding generation, analysis, coding, reasoning, tool use, and some agent capabilities in an increasing number of real-world tasks, thereby driving the expansion of dedicated accelerated computing, data centers, grid connections, cooling, and supply chains. When genuine AGI will arrive, and whether machine subjecthood or superintelligence will emerge, are not premises on which this paper depends. If AI capabilities, algorithmic efficiency, hardware architectures, or demand patterns change substantially in the future, the paper's quantitative judgments must be updated. Its research responsibility to examine how computation must occur through the real material world will nevertheless remain.

Contents

Introduction | The Stronger Digital Capabilities Become, the More We Must See the Material World on Which They Depend

1 | Established Disciplinary Foundations: What Ecology, Earth Systems, and Resources and Environment Already Know

2 | Objects of Study: From a Single Computation to the Complete Material Chain

3 | Electricity: A Limited Global Share Can Coexist with High Local Pressure

4 | Power Grids: Annual Electricity Supply Does Not Mean Connectability, Stability, or Affordability

5 | Water: Why One Figure for “AI Water Consumption” Is Far from Enough

6 | Carbon and Climate: Electricity Sources, Time and Place, and Life-Cycle Boundaries

7 | Materials and Critical Minerals: Intelligent Infrastructure Also Has Upstream Supply Chains

8 | Chips, Manufacturing, and Embodied Environmental Impacts: Environmental Accounts Beyond Operation

9 | Electronic Waste and Circularity: Is Faster Replacement the Only Route to Efficiency?

10 | Land, Ecology, and Regional Carrying Capacity: Global Averages Cannot Conceal Local Realities

11 | Efficiency and Rebound: Why Aggregate Use May Still Grow When Each Unit Uses Less

12 | AI May Also Reduce Environmental Pressure: From Load to Environmental Tool

13 | Multiobjective Trade-offs: Energy, Water, Carbon, and Materials Cannot Stand In for One Another

14 | How Earth-System Boundaries Enter AI Research without Becoming an Overall Ecological Score

15 | Real-World Conditions for “Intelligence Abundance”: From Digital Possibility to Civilizational Benefits

16 | Interdisciplinary Handoffs: Who Must Re-enter Once Facts Have Been Established?

17 | What Facts Could Change This Paper: The Strongest Objections and a Refutable Research Agenda

Conclusion | There Is No Intelligence Detached from Earth, Only Intelligence That Lives Better on Earth

Further Reading and Entry Points for Collaborative Research

Appendix A | Register of Source–Claim Boundaries

Appendix B | AI Resources and Environment Stress-Test Matrix

Declaration of Generative AI Use

References

Version Note

Introduction | The Stronger Digital Capabilities Become, the More We Must See the Material World on Which They Depend

One illusion that AI readily creates is the confusion of “digital” with “immaterial.” Model weights can be copied, software updated remotely, and inference interfaces extended around the world within seconds. It is therefore easy to imagine intelligence growth as constrained only by algorithms and capital. Yet every actual training and inference operation takes place on servers; servers depend on chips, racks, power, cooling, communications, and buildings; electricity comes from particular grids; water comes from particular watersheds; hardware comes from particular mines, smelters, wafer fabs, and supply chains; and discarded equipment returns to real waste systems. The “cloud” has never been placeless.

The IEA's 2025 Energy and AI report estimated that global data centers consumed approximately 415 TWh of electricity in 2024, around 1.5% of global electricity use. Its 2026 update estimated 2025 consumption at approximately 485 TWh and maintained a central scenario of approximately 950 TWh by 2030, around 3% of global electricity consumption, while noting substantially faster growth in electricity use by AI-dedicated data centers.[1–2] These figures support neither the exaggeration that “AI will consume the world's electricity” nor the careless dismissal that “3% makes it insignificant.” Electricity systems are strongly spatial: the same global share may be concentrated in a small number of nodes, cities, and regions, creating very different pressures on construction timelines, prices, emissions, and reliability.

There is already real-world evidence for this distinction between aggregate quantities and local conditions. Ireland's Central Statistics Office reported in 2026 that data centers accounted for 23% of national metered electricity consumption in 2025, compared with 5% in 2015.[36] Virginia's 2024 JLARC study likewise identified data centers as the main driver of rapid future electricity-demand growth and found that meeting unconstrained demand would require very large additions to generation and transmission infrastructure.[37] Neither case can be generalized into a global conclusion. Both nevertheless suffice to demonstrate that environmental and resource research must examine global totals and regional carrying capacity together.

Digital capabilities can be replicated almost instantaneously; power grids, watersheds, materials, and ecosystems cannot.

This paper therefore begins not with the overarching question of whether “AI is green,” but with a series of local questions that professional inquiry can establish: how much electricity is available, when, and where; where water is withdrawn and how it is consumed; at what stages materials enter; how frequently equipment is replaced; where emissions occur; how land and ecological impacts are distributed; whether aggregate demand continues to grow after efficiency improves; and how much real benefit AI can deliver when applied to energy and the environment themselves. Only when these questions have been established separately is civilization entitled to proceed to discuss the actual feasibility of so-called intelligence abundance.

1 | Established Disciplinary Foundations: What Ecology, Earth Systems, and Resources and Environment Already Know

1.1 Ecology First Studies Relationships and Carrying Capacity Rather than Assigning Civilization an Overall Score

Ecology examines relationships between organisms and their environments, populations and communities, energy and material cycles, and disturbance and recovery. One of its most important lessons is that impacts often depend on place, scale, timing, and system state: the same resource appropriation or emissions may have different consequences in different ecosystems. Ecological research can identify pressures on carrying capacity, vulnerabilities, and irreversible risks, but ecological indicators alone cannot determine all social

values. This paper inherits that disciplinary restraint: environmental facts must enter civilizational judgment, but must not monopolize it.

1.2 Earth-System Science Addresses Cross-Scale Coupling and Conditions for Global Stability

The planetary-boundaries framework proposed by Rockström and colleagues, its update by Steffen and colleagues, and Richardson and colleagues' further assessment in 2023 all seek to examine climate, the biosphere, land, freshwater, biogeochemical cycles, and “novel entities” together in relation to Earth-system stability.[22–24] The 2023 update assessed six boundaries as having been transgressed. This is an enormously illuminating and still-evolving scientific framework, not a table of thresholds for granting construction permits to individual data centers. This paper uses it to remind researchers that local resource pressures can accumulate through multiple Earth-system processes, while emphatically refusing to convert “planetary boundaries” mechanically into scores for AI projects.

1.3 Climate Science Provides Causal and Risk Boundaries Rather than a Ban on a Particular Technology

The IPCC Sixth Assessment Synthesis Report makes clear that human activities have unequivocally caused global warming; that climate, ecosystems, biodiversity, and human systems are interdependent; and that some risks increase with warming and may have long-term or irreversible consequences.[25] This gives data-center electricity sources, construction speed, and life-cycle emissions real climatic significance. Yet the mere fact that an AI application “has emissions” cannot by itself determine whether it should be prohibited. Alternative pathways, system benefits, emissions intensity, and other social objectives must be compared.

1.4 Industrial Ecology and Life-Cycle Assessment: Look Beyond the Operational Stage

Life-cycle assessment (LCA) has long addressed environmental impacts from resource extraction through production, use, and end-of-life treatment by defining goals and scope, compiling inventories, assessing impacts, and interpreting results.[19–20] Industrial ecology understands production systems within flows of materials, energy, and waste and cycles among industries, emphasizing that resource extraction, manufacturing, and disposal must not be excluded from the system. Research on metal life cycles further shows substantial differences in the resource and environmental burdens of different metals: “total material mass” cannot substitute for material types and supply-chain impacts.[30–31] ISO 14046 additionally provides life-cycle principles and requirements for water footprints.[21] OECD policy research on the “AI footprint” likewise explicitly distinguishes the direct environmental impacts of AI computation and equipment from the indirect costs and benefits generated by AI applications, recommending assessment beyond operational energy use and better measurement standards and transparency.[33] These established methods mean that reporting training electricity alone, or the energy consumed by a single inference, does not constitute a complete environmental assessment of AI.

1.5 Resources and Environment Research Reminds Us That Scarcity Can Shift

The UNEP International Resource Panel's 2024 Global Resources Outlook reports that global material use has more than tripled over roughly the past 50 years, and that resource extraction and processing are themselves major sources of greenhouse gases, particulate-matter health impacts, land use, biodiversity loss, and water stress.[15] AI is neither the sole nor the main source of global resource growth. As intelligent infrastructure expands, however, it enters existing competition surrounding the energy transition, urban construction, minerals, and manufacturing. The “digital economy” therefore cannot be regarded as an escape from the material economy; it represents a structural transformation of material demand.

2 | Objects of Study: From a Single Computation to the Complete Material Chain

Discussions of AI's environmental impacts frequently conflate different objects: “a model,” “a training run,” “a prompt,” “a data center,” “the entire data-center industry,” and “all digital infrastructure” are not the same object of research. In recalibrating global data-center energy estimates in 2020, Masanet and colleagues already warned that server efficiency, virtualization, infrastructure efficiency, and changing workloads can substantially alter aggregate estimates.[5] LBNL's 2024 US data-center energy-use report subsequently brought servers, storage, networks, facility efficiency, and different load types into a single bottom-up accounting framework, providing the direct baseline for the 2026 update.[4] This separation of objects becomes even more important in the generative AI era.

Table 2. Six levels of objects in AI environmental research and their evidentiary stopping points

Object level	What it includes	Main evidence	What it cannot automatically establish
Task/workload	A training, inference, retrieval, generation, or agent-chain execution	Task duration, tokens/samples, compute, hardware utilization	Annual environmental impacts of a data center
Hardware/server	GPUs/accelerators, CPUs, storage, networks, and server configurations	Power, performance, lifetime, utilization	Full life cycles of grids, water use, or materials
Data-center facility	Power supply and distribution, cooling, buildings, backup power, and IT equipment as a whole	PUE, WUE, load profiles, facility water use	Regional grid and watershed impacts
Grid/watershed/region	Generation mix, transmission and distribution, withdrawals, ecology, and local infrastructure	Marginal/average electricity, water stress, capacity, emissions, and ecological conditions	Economic or political legitimacy
Supply chain/life cycle	Minerals, smelting, chip manufacturing, buildings, transport, replacement, disposal	Material flows, embodied environmental impacts, waste, and circularity	Operational efficiency cannot substitute for this level
Global Earth system	Large-scale processes involving climate, the biosphere, freshwater, land, and more	Cumulative pressure, boundaries, and long-term risks	Cannot be mechanically converted into individual project permits

2.1 Four Types of System Boundary Must Be Disclosed

The first is the temporal boundary: an hourly peak, annual electricity use, equipment lifetime, and a 2030 scenario are entirely different things. The second is the spatial boundary: server rooms, campuses, grid nodes, watersheds, countries, and the globe must not be conflated. The third is the life-cycle boundary: whether only operational electricity is considered or chip manufacturing, buildings, transport, and disposal are also included. The fourth is the attribution boundary: data centers simultaneously support conventional cloud services, storage, video, enterprise computing, and AI, while operators generally do not disclose a sufficiently granular breakdown of AI and non-AI loads. Any “AI environmental figure” that fails to specify these four boundaries has limited interpretability.

2.2 Efficiency Indicators Are Useful but Must Not Exceed Their Remit

ISO/IEC 30134-2:2026 defines PUE as a standardized key indicator of data-center energy efficiency, while ISO/IEC 30134-9:2022 defines WUE to quantify data-center water-consumption intensity during use.[34–35]

Both are important, but address only facility energy efficiency and operational water intensity, respectively. Lower PUE does not mean lower carbon emissions; lower WUE does not mean a lower full-life-cycle water footprint; and neither establishes whether an AI service produces sufficient social value. Indicators must return to the questions they were designed to answer.

3 | Electricity: A Limited Global Share Can Coexist with High Local Pressure

3.1 The Most Reliable Current Figures Remain Ranges and Scenarios

In 2025, the IEA estimated global data-center electricity consumption in 2024 at approximately 415 TWh and projected approximately 945 TWh by 2030 in its base case. Its 2026 update raised the 2025 estimate to approximately 485 TWh and the 2030 figure to approximately 950 TWh, suggesting that electricity consumption by AI-dedicated data centers could roughly triple between 2025 and 2030.[1–2] Such figures should be interpreted as energy-system scenarios under explicit assumptions, not facts about “how much electricity AGI must consume.” Hardware efficiency, AI adoption, capital markets, supplies of chips and transformers, construction permits, and grid connections will all affect the actual trajectory.

LBNL's 2026 US update models bottom-up data including equipment shipments, annual electricity use per device, cooling simulations, facility types, and locations. It gives a reference scenario of 649 TWh in 2030, approximately 11.8% of US electricity use, together with a scenario range of 9.5%–15.3% and a broader combined uncertainty interval of 521–843 TWh.[3] The value of this study is not simply whether its number is higher or lower. It breaks a “forecast” into testable assumptions: how much equipment is installed, how long AI chips last, what idle power and utilization look like, and how efficient facilities are.

3.2 Global Averages Can Conceal High Local Concentrations

Data-center electricity consumption is geographically concentrated. The IEA notes that approximately half of US data-center capacity is concentrated in five regional clusters, with shares in local electricity markets far exceeding the global average.[1] Ireland's statistic that data centers accounted for 23% of national metered electricity consumption in 2025 further demonstrates that researchers cannot dismiss local problems by saying the global share is only a few percentage points.[36] Conversely, Ireland's 23% cannot be used to claim that every country will reach the same level.

3.3 “Is There Enough Electricity?” Is Not a Global Proposition

The real questions for electricity systems include at least annual energy supply, peak power, ramp rates, nodal capacity, transmission and distribution, generation mix, reliability, reserves, and construction timelines. Ample aggregate annual generation in a country does not mean a particular campus can connect hundreds of megawatts of continuous load within two years. The ability to build substantial renewable generation in a region does not mean generation hours naturally coincide with AI load hours. AI's energy constraint is therefore a question of availability across time and space, not a division involving global electricity totals.

4 | Power Grids: Annual Electricity Supply Does Not Mean Connectability, Stability, or Affordability

Data centers are generally built faster than power grids can expand. The IEA notes that data centers can enter service within two or three years, whereas generation, transmission, transformers, and grid planning often require longer. Its 2026 update further emphasizes that supply-chain bottlenecks for transformers, gas turbines, chips, and IT equipment have already reduced the likelihood of near-term high-growth scenarios being realized.[1–2] Thus, “sufficient generation technology will exist in the future” does not automatically establish that “a particular region can provide a connection in time.”

4.1 The Virginia Case: Load Growth Can Rewrite Infrastructure Pathways

Virginia's 2024 JLARC study finds that under an unconstrained-demand scenario, the state's electricity demand could double over the next decade, with data centers as the principal driver. Meeting that demand would require major additions to generation and transmission, and even accommodating half the unconstrained growth would present substantial construction challenges.[37] This is a regional planning study, not a global conclusion about AI. It establishes that concentrated large loads cause new data centers to change electricity-system investment pathways and resource mixes. Environmental analysis must incorporate these indirect infrastructure consequences.

4.2 Data Centers May Also Become Flexible Loads

Treating data centers as permanently inflexible loads may likewise underestimate technological possibilities. Granderson and colleagues' 2026 synthesis identifies four mechanisms for flexibility: shifting computational loads in time or location, adjusting facility infrastructure, energy storage, and on-site generation. They note that demand flexibility cannot replace long-term additions to generation but can relieve near-term interconnection and peak pressures.[12] Colangelo and colleagues conducted a field demonstration on a 256-GPU cluster in a hyperscale cloud facility in Phoenix, Arizona, using software coordination to reduce power by 25% during a three-hour peak period while maintaining their defined AI quality-of-service guarantees, without relying on hardware modifications or energy storage.[11] This is important positive A/B evidence: some AI loads do have potential for grid interaction. It cannot be generalized to all workloads or all regions.

Table 3. Five questions that move beyond “electricity consumption” into the actual grid

Electricity issue	The actual question	What cannot substitute for it
Annual electricity supply	How much electricity can be produced in a year?	May conceal peak, nodal, and seasonal constraints
Peak power	How much instantaneous capacity is required at the tightest moment?	Does not establish annual energy use or emissions
Nodal/transmission capacity	Can the load be connected at a particular location?	Cannot be replaced by national generation totals
Flexibility	Can loads shift time, reduce power, or migrate?	Depends on task latency, data location, SLAs, and security
Reliability/reserves	Can operation continue through faults and extreme weather?	May require more on-site fuel, storage, and additional infrastructure

5 | Water: Why One Figure for “AI Water Consumption” Is Far from Enough

Water issues are particularly susceptible to misleading headline numbers. Data-center water use has at least three pathways: direct consumption for facility cooling; indirect consumption in electricity generation; and upstream life-cycle water footprints associated with chips, equipment, buildings, and other inputs. ISO/IEC WUE addresses only water consumption during the data-center use stage and is not equivalent to a full-life-cycle water footprint.[21,35]

5.1 Water Consumption for the Same Workload Can Differ by More than Four Orders of Magnitude

Lei, Lu, Shehabi, and Masanet's review and assessment of workload-level water consumption found that water use for the same workload can differ by more than 10,000-fold under different conditions. Water-consumption factors associated with each kWh of server electricity can differ by more than 1,000-fold, while server workload efficiency itself can differ approximately tenfold. Principal determinants include server efficiency, grid water-consumption factors, utilization, cooling systems, infrastructure efficiency, climate, the proportion

of idle servers, and replacement cycles.[6] This finding directly rejects a common communication practice: presenting “water consumed per prompt” measured for a particular model, location, and cooling system as a universal AI constant.

Earlier work by Mytton and by Siddik and colleagues had already shown that data-center water impacts require consideration of both direct on-site consumption and indirect electricity-related consumption and differ substantially by region.[7–8] Xiao and colleagues' 2025 modeling of US AI-server scenarios for 2024–2030 provides further broad ranges for water and carbon, explicitly identifying efficiency, grid decarbonization, and spatial distribution as sources of deep uncertainty.[9] For future AI water footprints, conditional ranges are therefore often a more honest expression of modeling results than point predictions.

5.2 Saving Water and Reducing Carbon Can Sometimes Conflict

There is no single “green” cooling solution independent of location. Chien and colleagues' 2026 review notes that strategies to reduce data-center water use can, in some cases, increase CO₂-equivalent emissions. Water and carbon may therefore be adversely coupled and require joint optimization that considers the energy mix, climate, and facility design.[10] This does not mean water cannot be conserved. It means that the lowest WUE does not automatically indicate the best overall environmental performance.

5.3 Water Must Be Assessed by Watershed and Time Period

The same consumption of one million cubic meters of water has different ecological and social consequences in a highly water-stressed watershed, a water-abundant area, the dry season, or the rainy season. Water-resource research must examine withdrawal sources, consumptive use, reuse rates, river and groundwater conditions, ecological water requirements, seasonal fluctuations, and competition with other users. National average water quantities cannot substitute for local watershed carrying capacity, and a company's annual “water replenishment” pledge does not automatically establish that withdrawals have been equivalently restored in time and space.

6 | Carbon and Climate: Electricity Sources, Time and Place, and Life-Cycle Boundaries

6.1 Electricity Consumption Is Not Carbon Emissions

Emissions from the same 1 kWh depend on the generation mix, time, location, and whether average or marginal emissions factors are used. In 2025, the IEA estimated data-center-related indirect CO₂ emissions at approximately 180 Mt, rising with electricity consumption in its base case, while emphasizing that their share of global emissions remains limited and regional electricity mixes differ substantially.[1] Training and inference energy use must therefore be tied to specific grid emissions factors. Directly equating electricity consumption with an equivalent increase in pollution is not rigorous.

6.2 Operational Emissions Are Not Full-Life-Cycle Emissions

Chip manufacturing, servers, data-center construction, transformers, backup systems, and frequent hardware replacement all entail embodied environmental impacts. Chien and colleagues' review notes that hardware and facility manufacturing for large AI data centers can account for a substantial share, and under particular life-cycle boundaries even more than half, of emissions. It discusses reducing these impacts through longer hardware use, reuse, and low-carbon design.[10] Plociennik and colleagues observe that systematic LCA of AI systems remains at an early stage: operational electricity or CO₂ equivalents alone are inadequate, and indicators including critical raw materials, water, and resource depletion should also be included.[18] These studies share a direction of expanding system boundaries, rather than claiming that a universal AI carbon-footprint constant already exists.

6.3 AI Expands Demand but May Also Help Reduce Emissions in Other Sectors

The IEA also assesses AI's potential to reduce emissions in power grids, buildings, industry, transport, methane monitoring, and energy innovation. It identifies potentially large reductions under widespread-adoption scenarios, while explicitly noting that present barriers involving adoption, data, skills, infrastructure, and rebound effects prevent these benefits from materializing automatically.[1] Net climate impact must therefore compare AI's own additional emissions with real emissions reductions enabled by AI, without treating theoretical technological potential as a net benefit already achieved.

7 | Materials and Critical Minerals: Intelligent Infrastructure Also Has Upstream Supply Chains

The material character of AI infrastructure extends beyond electricity and water. Servers, chips, power supply and distribution, storage, and networks require copper, aluminum, silicon, and various critical or strategic minerals. The IEA's 2025 Global Critical Minerals Outlook shows that the energy transition itself has already driven rapid growth in critical-mineral demand, while supply concentration, project lead times, and export restrictions present real risks.[13] Its 2026 update and accompanying critical-minerals dataset further show that refining remains highly concentrated for many minerals and that supply growth and processing capacity are heavily concentrated in a few countries. Specific minerals and years must be checked individually against the dataset.[14,32] AI thus enters not an empty resource ledger, but a materials system already driven jointly by the energy transition, electrification, and industrial security.

7.1 Data Centers May Exert Nonlinear Pressure on a Few Strategic Materials

The IEA's 2025 Energy and AI report estimated that data-center demand for gallium could amount to approximately 10% of global supply in 2030 and noted that China then accounted for approximately 95% of global gallium refining. The IEA's 2026 Global Critical Minerals Outlook further shows that China, as the largest refining country, still accounts for more than 90% of global gallium supply.[1,14] These figures do not mean that "AI will exhaust gallium." They show that even a material constituting a small share of total data-center mass can become a bottleneck for expansion and security if its supply chain is highly concentrated, expansion is constrained by its by-product character or specialized processing capacity, and substitution options are limited. Total mass is not a sufficient proxy for strategic importance.

7.2 Materials Must Be Analyzed Together with Substitution, Recycling, and Technological Pathways

Material demand is not a fixed coefficient. Chip processes, packaging, optical interconnects, power electronics, cooling, storage, and grid design all alter the material mix; recycling rates, secondary materials, equipment lifetimes, and modular design determine demand for new extraction. Resource research must disaggregate "how much is needed" by technological pathway and scenario rather than treating today's bill of materials as permanent. Conversely, any claim that new materials will automatically solve bottlenecks must provide evidence on performance, scale, recycling, toxicity, and supply chains.

8 | Chips, Manufacturing, and Embodied Environmental Impacts: Environmental Accounts Beyond Operation

AI environmental research that looks only at server outlet power misses semiconductor manufacturing and equipment replacement. Chip manufacturing involves ultrapure water, energy, chemicals, clean rooms, upstream materials, and complex global supply chains. Process nodes, fabrication sites, electricity conditions, and water conditions differ substantially. Without data on a comparable basis, this paper does not assign a fixed water-consumption or carbon-emissions figure to each AI chip. Instead, it stresses that manufacturing must be included within LCA system boundaries.

Plociennik and colleagues explicitly noted in 2025 that AI life-cycle assessment still lacks mature, unified methods. Existing energy calculators generally cover operational energy and carbon conversion but are inadequate for full impacts involving critical raw materials, hardware production, maintenance, and end-user devices.[18] The OECD in 2022 also recommended more transparent, standardized, comparable AI environmental measurement and a distinction between direct and indirect environmental effects.[33] Together these studies show that the most important task today is not pursuing an apparently precise “overall AI environmental score,” but disclosing missing data and boundaries.

8.1 Hardware Replacement Cycles Are Themselves Environmental Variables

Faster chips can lower electricity use per task. Yet if economic optimization continually shortens server replacement cycles, upstream manufacturing and waste burdens may rise at the same time. Conversely, excessively extending the life of inefficient equipment may increase operational electricity consumption. The optimal replacement cycle is therefore a life-cycle question, not a slogan that “newer is greener” or “older is more environmentally friendly.” Chien and colleagues discuss the potential for reused or lower-performance hardware to reduce embodied emissions in some inference applications, but applicability still depends on task-performance requirements and electricity conditions.[10]

9 | Electronic Waste and Circularity: Is Faster Replacement the Only Route to Efficiency?

Global electronic waste is already a rapidly growing waste stream. The ITU/UNITAR Global E-waste Monitor 2024 estimates that approximately 62 billion kilograms of e-waste were generated globally in 2022, with only 22.3% documented as formally collected and recycled. It projects substantial further growth by 2030 under existing trends.[16] AI servers are only one additional component; global e-waste cannot all be attributed to AI.

Using material-flow and computing-capacity scenarios, Wang and colleagues in 2024 estimated that cumulative generative-AI-related e-waste during 2020–2030 could reach 1.2–5.0 million tonnes across different scenarios. They suggested that circular-economy strategies could substantially reduce this waste stream.[17] This is Category B scenario research and must not be rewritten as “AI will certainly produce five million tonnes of waste by 2030.” Its real research question is how server replacement rates, geopolitical supply restrictions, equipment reuse, modular repair, recycling, and cross-generation compatibility jointly determine material outcomes.

9.1 Circularity Does Not End with Recycling Waste at the End of the Chain

More mature circular pathways include extending the life of usable equipment; moving high-end training hardware into inference or lower-load tasks; modular upgrades; repair; component reuse; higher material-recovery rates; and reducing hard-to-recycle composites at the design stage. Circular strategies must simultaneously meet cybersecurity, data destruction, reliability, and performance requirements. Ecology cannot ignore security and service continuity merely because circularity is preferable; engineering cannot ignore material consequences merely because faster retirement appears more efficient.

10 | Land, Ecology, and Regional Carrying Capacity: Global Averages Cannot Conceal Local Realities

A data-center campus itself may occupy limited land, but the substations, transmission, generation, storage, roads, and water infrastructure it induces expand its spatial footprint. Siting may also involve wetlands, farmland, wildlife habitats, historic landscapes, or densely populated areas. Virginia's 2024 JLARC study considers land, natural resources, residential noise, local tax revenues, and energy infrastructure within a

single regional analysis, demonstrating precisely why data-center environmental impacts should not be reduced to the electricity meter.[37]

10.1 Ecological Impacts Must Start with Place

Ecosystems do not operate within corporate organizational boundaries. Even when a facility has low on-site emissions, real consequences remain if new transmission crosses sensitive habitat, withdrawals change dry-season flows, or backup generation increases local air pollution. The IPBES 2019 Global Assessment reports that since 1970, land-use change has had the largest relative negative direct impact on the state of nature in terrestrial and freshwater ecosystems.[26] This global context explains why infrastructure siting, associated transmission, and indirect land changes must enter ecological review. It cannot, however, establish that a particular data center has already caused biodiversity loss: project attribution still requires evidence on location, baselines, impact pathways, and cumulative effects. Conversely, siting on brownfields, in existing industrial areas, or in low-water-stress locations while using lower-impact energy and recycled water may substantially reduce environmental pressure. The central task of ecological research is not to pronounce data centers “harmful” or “harmless,” but to locate environmental consequences within specific spatial processes.

10.2 Compact Regions Reveal Carrying-Capacity Boundaries Earlier

Singapore's 2024 Green Data Centre Roadmap explicitly recognizes data centers as electricity- and water-intensive infrastructure. Under land and resource constraints, it proposes unlocking at least 300 MW of additional capacity in the near term, with efficiency and green-energy deployment as conditions for expansion. [38] This is not a policy template that can be copied into every country. It does demonstrate that regional resource conditions directly shape the scale, technology, and governance pathways of AI infrastructure.

11 | Efficiency and Rebound: Why Aggregate Use May Still Grow When Each Unit Uses Less

AI environmental discussions readily confuse two different quantities: resource intensity per task and total resource use. Hardware, models, quantization, sparsification, caching, and facility efficiency can reduce electricity use for the same task. Yet as efficiency lowers costs, user populations, task frequency, context length, video generation, agent chains, and new applications may expand simultaneously. Lower consumption per unit does not automatically equal lower aggregate consumption.

11.1 Rebound Is a Contested Empirical Question Rather than a Law

Energy economics has long debated rebound effects. Sorrell's 2009 review notes that the strong version of “Jevons' paradox” is difficult to establish empirically, while partial rebound induced by efficiency gains has theoretical and empirical support, with its magnitude depending on prices, income, technology, and market structure.[27] This paper therefore treats neither “more efficient AI necessarily uses more electricity” nor “falling energy per token necessarily reduces total industry energy use” as a law.

The IEA's 2026 update describes a real coexistence of trends: power requirements per AI task are declining rapidly, but AI use and energy-intensive applications are expanding faster, so total data-center electricity consumption is still projected to grow.[2] This is precisely why environmental research must report unit efficiency, utilization, task scale, total service volume, and aggregate resource use together.

Table 4. Why efficiency improvements cannot be directly converted into aggregate environmental outcomes

Efficiency change	Direct benefit	Rebound or trade-off requiring empirical testing
Higher hardware efficiency per unit	Lower electricity consumption for the same workload	Deployment grows as equipment becomes cheaper or more powerful

Efficiency change	Direct benefit	Rebound or trade-off requiring empirical testing
More efficient model inference	Less compute for the same output	Longer contexts, more frequent calls, multistep agent execution
Improved PUE/WUE	Lower facility resource use per unit of IT load	Total campus-capacity expansion offsets unit improvements
More low-carbon electricity	Lower emissions per unit of electricity	Total load growth, displacement of other electricity uses, more material-intensive construction
Longer equipment lifetimes	Less manufacturing and disposal	Lower operational efficiency of older hardware increases use-phase electricity consumption

12 | AI May Also Reduce Environmental Pressure: From Load to Environmental Tool

An environmental paper that lists only AI's resource costs would also distort the picture. Sustainable-AI research already explicitly distinguishes “AI for sustainability” from “the sustainability of AI.” The former examines which environmental problems AI can help solve; the latter examines the resource and environmental costs of building and operating AI itself. The two cannot simply cancel each other out.[28] AI can contribute to grid forecasting and dispatch, renewable-generation forecasting, equipment maintenance, building energy savings, industrial control, methane detection, weather and climate analysis, new materials, and energy-storage research. Rolnick and colleagues systematically survey emissions-reduction applications of machine learning in electricity, transport, buildings, industry, climate prediction, and social systems. The IEA's 2025 report likewise devotes its other half to AI's potential role in energy optimization, innovation, and security.[1,29]

12.1 Positive Effects Must Move from Technological Potential to Real Net Outcomes

An algorithm that reduces energy use by 10% in an experiment does not imply 10% net energy savings across an industry. Real net outcomes also depend on deployment rates, costs, organizational adoption, infrastructure replacement, behavioral changes, and rebound. AI's discovery of a new battery material does not mean it has been commercialized. Improved grid forecasting does not mean the grid has sufficient transmission or market mechanisms to use it. Environmental benefits must also follow the complete causal chain from capability to deployment to real-world outcomes.

12.2 Data Centers Themselves Can Participate in Power-System Optimization

Colangelo and colleagues' 256-GPU field experiment and LBNL's review of flexible data centers provide real-world evidence for turning AI loads from pure consumers into partially adjustable resources.[11–12] If increasing numbers of training, batch-processing, or latency-tolerant workloads can shift time or location according to grid conditions, data centers may reduce additional peak-capacity requirements, improve renewable-energy absorption, and ease local constraints. Yet low-latency online inference, high-security tasks, data sovereignty, and network restrictions may substantially limit this flexibility.

12.3 Both the Positive and Negative Chains Must Be Retained

Table 5. Bidirectional causality between AI and the environment, rather than one-way destiny

Pathway	Causal chain
Negative possibility	AI capability → more computation → infrastructure → energy/water/material pressures → ecological and climatic consequences
Positive possibility	AI capability → forecasting/control/discovery → better energy and resource efficiency → lower environmental pressure
Ultimate net outcome	Depends on the two chains' scale, location, timing, deployment, rebound, and substitution, rather than an inherent “nature” of AI

13 | Multiobjective Trade-offs: Energy, Water, Carbon, and Materials Cannot Stand In for One Another

An environmental system rarely has only one objective. Air cooling may reduce direct water consumption while increasing electricity demand; evaporative cooling may improve energy efficiency while increasing water consumption; large-scale storage construction improves accommodation of renewable energy while adding mineral and manufacturing demand; and extending hardware life reduces embodied environmental impacts while potentially increasing operational electricity use. Chien and colleagues' 2026 review explicitly places water-carbon trade-offs on the agenda for AI data-center design.[10]

13.1 No “Overall Green AI Score”

This paper rejects compressing energy, water, carbon, land, materials, biodiversity, and social impacts into an opaque overall score. The reason is not that these indicators cannot be compared, but that value weights, regional boundaries, and irreplaceable losses must not be silently predetermined by technical specialists. Professional environmental analysis is better suited to providing a transparent multidimensional evidence table specifying each indicator's units, scope, thresholds, data quality, regional significance, and trade-offs, after which economics, ethics, law, politics, and affected people must participate in the choice.

Table 6. Why multidimensional environmental evidence must be retained separately

Dimension	Typical indicators	Principal discipline	Prohibited automatic substitution
Electricity/energy	TWh, MW, load profiles, PUE	Energy systems and grids	Cannot stand in for carbon, cost, or social value
Water	Withdrawals/consumption, WUE, water footprint, watershed stress	Water resources/hydrology	Cannot use national averages alone
Climate	CO ₂ e, average/marginal carbon intensity, life-cycle emissions	Climate/energy/LCA	Cannot stand in for biodiversity and water
Materials	Mass, criticality, supply concentration, circularity rate	Industrial ecology/resources and environment	Cannot consider total tonnage alone
Ecology/land	Habitats, disturbance, recovery, local irreversibility	Ecology/environmental science	Cannot automatically create political authorization

14 | How Earth-System Boundaries Enter AI Research without Becoming an Overall Ecological Score

Earth-system science offers an important reminder about scale: global environmental risks are not a list of infinitely substitutable resources. Climate, the biosphere, freshwater, land, and nitrogen and phosphorus cycles interact, and some changes involve thresholds, feedbacks, and irreversibility.[22–25] As additional industrial activity, AI infrastructure must be situated within these aggregate pressures rather than presumed naturally detached from Earth systems because it belongs to a digital industry.

14.1 Planetary Boundaries Are Not Project-Permitting Lines

Steffen and colleagues explicitly discuss cross-scale interactions and regional heterogeneity, while Richardson and colleagues' update likewise emphasizes systemic relationships among multiple boundaries.[23–24] Thus, “a global boundary has been transgressed” cannot be directly converted into “this data center must not be built.” Conversely, a project's tiny global share does not establish that its local impacts are irrelevant. Project-level assessment requires evidence about place, ecology, watersheds, grids, and life cycles; the global framework explains cumulative consequences and the systemic context.

14.2 Ecological Irreversibility Must Be Distinguished from Substitutability

Some environmental losses can be mitigated through restoration, alternative locations, or technological modification. Others are highly irreversible or have recovery periods far longer than the project's benefit period. Professional research should identify these differences rather than preemptively erase all harms by claiming that compensation is possible. At the Category F normative level, this paper argues that irreversible ecological damage cannot disappear from the accounts because AI generates digital benefits. Whether a specific trade-off should be accepted nevertheless requires ethics, law, politics, and affected people to participate together.

15 | Real-World Conditions for “Intelligence Abundance”: From Digital Possibility to Civilizational Benefits

“Intelligence abundance” is readily misunderstood to mean that as model inference becomes cheaper, civilization will enter a state of unlimited resources. In fact, falling marginal costs of intelligent services address only some constraints. Intelligence at scale must also pass through electricity, chips, networks, construction, water, land, materials, supply chains, and ecological carrying capacity. Scarcity does not disappear when one capability becomes cheaper; it may shift to new bottlenecks.

15.1 First Test: Does Efficiency Truly Improve Faster than Aggregate Demand Grows?

If algorithmic and hardware efficiency improve over the long term faster than total task demand grows, AI's aggregate resource pressure may stabilize or even fall. If agents, multimodality, video, and broad automation expand demand faster, total resource use will still grow. This question must be answered using real workload, hardware-shipment, and energy data rather than technological optimism or pessimism.

15.2 Second Test: Can Growth Enter the Grid without Crowding Out Other Basic Needs?

When data centers become major loads, the real civilizational question is not whether they may use electricity. It is whether added capacity arrives in time, is reliable and low-emission, and has clearly assigned cost responsibility; whether residents and industry bear opaque infrastructure costs; and whether it conflicts with critical public services, housing, electrification, and industrial development. These consequences first require joint reassessment by energy research, economics, and political science. Ecology cannot adjudicate them alone.

15.3 Third Test: Do Water, Materials, and Ecology Have Sustainable Alternative Pathways?

Even after electricity problems are resolved, water, chip materials, critical minerals, land, and waste may remain bottlenecks. Genuinely robust intelligent infrastructure requires low-water or locally appropriate cooling, resource circularity, longer periods of useful equipment value, diversified supply chains, ecologically appropriate siting, and recoverable design. No single breakthrough can substitute for the whole material chain.

15.4 Fourth Test: Can AI's Environmental Benefits Exceed Its Own Additional Burdens?

If AI substantially improves energy systems, materials development, climate adaptation, and resource efficiency, its net environmental outcome may be far better than a static estimate counting only data-center loads. Conversely, if AI mainly generates energy-intensive, low-value demand while positive environmental applications remain technical demonstrations over the long term, the opposite may occur. This paper retains “net environmental benefit” as a Category C civilizational-level judgment awaiting verification rather than settling it in advance.

Intelligence abundance does not mean “unlimited compute”; it means that after intelligence expands, the real world remains capable of sustainably supporting human life.

16 | Interdisciplinary Handoffs: Who Must Re-enter Once Facts Have Been Established?

This paper's professional responsibility stops at ecology, energy, resources, and Earth systems. Once resource facts enter prices, rights, politics, security, technology, or civilizational purposes, they must be handed over for renewed assessment. Handoff does not mean transferring conclusions to a “higher discipline”; it prevents disciplinary overreach.

Table 7. Handoffs of judgment between this ecology/Earth-systems paper and neighboring disciplines

Neighboring discipline	Its independent responsibility	Handoff with this paper
AI/computer science	Models, chips, algorithms, loads, system efficiency, and technical feasibility	This paper reassesses those engineering demands when they enter grids, water, materials, and life cycles
Systems science	General feedbacks, thresholds, cascades, resilience, and lock-in	This paper provides specific Earth, energy, water, and ecological facts; general patterns cannot replace actual measurement
Physics/information/spacetime	Physical energy use, heat, causality, and basic constraints	Fundamental physical constraints do not equal actual grid, resource, and ecological carrying capacity
Economics/political economy	Prices, investment, ownership, returns, and resource allocation	Economics must continue assessing ecological scarcity and environmental externalities once they enter its domain; prices cannot in turn prove ecological safety
Political science	Infrastructure power, state capacity, resource governance, and public authorization	Ecological facts cannot automatically confer governing authority
Ethics/law	Environmental rights, intergenerational responsibility, procedure, fairness, and harm	This paper does not use ecological indicators to settle ultimate conflicts of rights
International relations/strategic security	Geopolitical dependencies involving critical minerals, energy, and infrastructure	Strategic specialists must reassess how supply concentration enters strategic stability

Neighboring discipline	Its independent responsibility	Handoff with this paper
Sociology/anthropology	Local communities, land, culture, acceptance, and consequences for life	"Local impacts" cannot be represented solely by engineering costs
Life sciences	Biodiversity, biological systems, and impacts on life	A renewed handoff is required when ecosystem changes involve mechanisms of life

16.1 Five Explicit Prohibitions on Overreach

- Do not directly infer from "limited ecological carrying capacity" that a particular economic system or growth regime must be correct.
- Do not use electricity, water, or carbon indicators to prescribe model architectures and system guarantees on behalf of AI and computer science.
- Do not automatically use ecological thresholds to determine who is entitled to allocate energy, water, land, and minerals.
- Do not mechanically extrapolate resource pressure in one region into a conclusion that AI is globally unsustainable.
- Do not use "better environmental performance" to settle all civilizational values concerning survival, rights, happiness, and meaning.

17 | What Facts Could Change This Paper: The Strongest Objections and a Refutable Research Agenda

17.1 Strongest Objection One: Efficiency Gains May Rapidly Resolve AI Resource Pressures

This is an opposing position the paper must take seriously. Data-center history has included periods when efficiency improvements offset demand growth. Masanet and colleagues in 2020 used more granular technological data to challenge some excessively high extrapolations of global data-center energy use then circulating.[5] If future algorithms, hardware, model compression, and facility efficiency cause resource intensity per unit to decline persistently and substantially faster than service demand expands, the paper must substantially weaken its claim that resource constraints will become an important civilizational variable.

17.2 Strongest Objection Two: AI May Become a Technology with Net Environmental Benefits

If real energy savings and emissions reductions from AI in grids, energy innovation, materials, industrial processes, transport, buildings, climate adaptation, and ecological monitoring consistently exceed the additional environmental burden of AI infrastructure at scale and with demonstrable attribution, it would be wrong to view AI primarily as a source of resource pressure. Rolnick and colleagues and the IEA offer many entry points for positive applications.[1,29] This objection does not require denying the material nature of data centers; it requires a shift toward research on net impacts.

17.3 Strongest Objection Three: Many Current Local Conflicts May Primarily Reflect Existing Failures of Infrastructure Governance

Grid queues, housing costs, water-rights conflicts, land approvals, and community distrust may not be unique to AI; they may reflect existing problems in planning, pricing mechanisms, infrastructure investment, and governance capacity. If rigorous counterfactual studies show that, under equivalent added-load conditions, data centers create no distinct environmental increment relative to other industrial loads, this paper should return the corresponding issues to energy, economics, and governance research rather than rename old problems with the label "AI."

17.4 Evidence That Could Require the Paper to Narrow, Be Rewritten, or Lose Its Basis for a Standalone Treatment

Table 8. Conditions for refuting or withdrawing this paper's claims

Evidence direction	If future evidence shows the following	How this paper should change
Unit resource intensity	Long-term declines substantially outpace aggregate AI demand, with total pressure falling consistently globally and in major regions	Narrow the “expansion pressure” claim
Grid carrying capacity	Additional AI loads can be reliably accommodated through flexibility, distributed resources, and rapid grid expansion without substantially increasing costs/emissions	Reduce the importance assigned to grid bottlenecks
Water	Direct and indirect consumption remains low over the long term in major deployment regions and becomes decoupled from local water stress	Narrow the water-resource chapter
Materials/waste	Critical materials are highly substitutable, supply diversifies rapidly, equipment lifetimes lengthen, and circularity rises substantially	Weaken judgments about materials and e-waste
Net environmental benefit	Attributable AI energy savings, emissions reductions, and ecological benefits consistently exceed its additional environmental burden	Rewrite the overall direction as research on net positive effects
Independent factual domain	AI engineering, systems science, and economics can address all remaining problems without overreach	Reassess the justification for this paper as a standalone work

17.5 Professional Research That Can Be Undertaken Directly in the Next Round

- Establish a reproducible multiscale data chain linking tasks, servers, facilities, grids, watersheds, and life cycles, avoiding extrapolation from isolated indicators to an overall footprint.
- Compare water and electricity consumption for real workloads across regions, disclosing server efficiency, utilization, cooling, grid, and climate conditions.
- Establish separate load profiles and flexibility limits for AI training, online inference, batch processing, video generation, and agent systems.
- Continuously calibrate 2030 electricity scenarios against actual chip shipments, facility interconnections, canceled projects, and utilization data rather than merely adding announced capacity.
- Build AI life-cycle inventories including chip manufacturing, servers, buildings, backup systems, replacement, and disposal, and disclose data gaps.
- Conduct watershed-scale research on withdrawals, consumption, and seasonality in highly water-stressed regions rather than reporting only company-wide annual global water volumes.
- Compare full-life-cycle crossover points among hardware life extension, redeployment to less demanding uses, modular upgrades, and energy savings from new hardware.
- Test real net energy savings from AI in grid and energy optimization, distinguishing algorithmic potential, actual deployment, and rebound.
- Hand local grid, watershed, and ecological results together with evidence on communities, prices, employment, and international supply chains to neighboring disciplines for continued investigation.

Conclusion | There Is No Intelligence Detached from Earth, Only Intelligence That Lives Better on Earth

For the first time, AI enables humanity to replicate some previously expensive capacities for knowledge, analysis, and generation at very low marginal cost. This is a real civilizational opportunity. Yet marginal digital costs are not complete real-world costs. Every computation passes through chips and servers; every data center sits within land, a grid, and a watershed; and every generation of hardware originates in materials and ultimately enters circular use or waste. The more technology resembles an “intangible capability,” the greater the responsibility of professional research to place it back in the real world.

This does not lead to an anti-technology conclusion. On the contrary, if AI can make grids more flexible, accelerate energy and materials research, improve precision in equipment maintenance, and raise resource-use efficiency, it may expand rather than reduce civilization's ecological carrying capacity. What must be rejected are two automatic inferences: “digital means no material cost,” and “material cost means development should not occur.” Both bypass professional evidence.

Ecology, Earth-system science, and resources and environment research can make real boundaries visible: where the issue concerns annual totals and where local nodes; what has been measured and what remains a scenario; where water–carbon–material trade-offs arise; which losses are repairable and which may be irreversible; whether aggregate use truly declines as unit efficiency improves; and whether environmental benefits created by AI actually materialize. Once the facts become clear, economics, politics, law, ethics, and civilizational direction can re-enter.

Every civilizational vision of future intelligence abundance must withstand the test of the real material world.

Civilizational maturity means not escaping Earth, but continuing to understand what we depend on, change, consume, and restore as our capabilities grow stronger, and ensuring that technology ultimately expands human life rather than leaving invisible costs to other regions, other lives, or the future.

Further Reading and Entry Points for Collaborative Research

This paper has answered how AI infrastructure enters electricity, water, materials, land, climate, waste, and Earth systems; how far particular environmental figures remain valid; and where automatic extrapolation must stop. Continuing further is not intended to keep readers within the same theoretical system, but to let questions follow real causal pathways.

Upward | Why Do These Material Boundaries Ultimately Become Civilizational Questions?

Continue to Reckoning with the Disciplines in the Age of AGI: Disciplinary Boundaries, Interactions, and a Shared Research Agenda, V2.4, and the real-world convergence paper Research on the Prospects and Risks of Artificial Intelligence Applications in the AGI Era: From Capability Expansion to Civilizational Consequences, V1.2. The former explains why disciplinary facts must remain independent while reconnecting to a common causal chain; the latter follows how energy and compute pass through research, industry, governance, and life to generate feedback. The civilizational scale is not a superior tribunal: ecological facts can still force the overall theory to narrow or be rewritten.

<https://www.civitas.top/en/academic/papers/cl-aca-005/> | <https://www.civitas.top/en/academic/papers/cl-aca-022/>

Downward | Turning This Paper's Judgments into Verifiable Reality

Continue into energy systems, electrical engineering, water resources, industrial ecology, life-cycle assessment, material flows, critical minerals, regional environmental monitoring, and audits of real data centers. The most valuable next step is not another review asserting that “AI has environmental impacts,” but reproducible multiscale data: real loads, real watersheds, real hardware replacement, real LCA, real flexibility, and real ecological outcomes.

Leftward | What Happens When Resource Facts Enter Other Causal Chains?

Continue into AI and computer science to reassess algorithms, chips, and system efficiency; economics to study prices, investment, and distribution; political science and law to study infrastructure authorization, water rights, land, and cost responsibility; international relations and strategic security to study dependence on critical minerals, chips, energy, and infrastructure; and sociology and anthropology to study local communities, acceptance, and distributional consequences. Where biodiversity and living systems are involved, the life sciences must re-enter.

<https://www.civitas.top/en/academic/papers/cl-aca-015/> | <https://www.civitas.top/en/academic/papers/cl-aca-014/> | <https://www.civitas.top/en/academic/papers/cl-aca-010/> | <https://www.civitas.top/en/academic/papers/cl-aca-009/> | <https://www.civitas.top/en/academic/papers/cl-aca-019/> | <https://www.civitas.top/en/academic/papers/cl-aca-013/> | <https://www.civitas.top/en/academic/papers/cl-aca-012/>

Rightward | If You Disagree, First Test These Strongest Opposing Directions

Prioritize competing positions such as “AI's environmental pressures are overstated,” “efficient computing will substantially reduce aggregate demand,” “nuclear power, renewables, and new grids can eliminate electricity bottlenecks,” “water-saving technologies can rapidly remove water constraints,” “rebound effects are weak,” “critical materials are substitutable and supply can expand quickly,” and “AI's net contribution to energy and climate will be substantially positive.” Opposition is not a formal balancing exercise. Whenever these pathways acquire stronger real-world evidence, this paper must change.

Forward | Which Professional Work Could Directly Change This Paper?

New measured data, negative findings, cross-regional replication, watershed studies, grid simulations and field experiments, material-supply data, complete life-cycle inventories, equipment-retirement and circularity records, and research on net energy savings and rebound can all weaken, remove, or rewrite the paper's judgments. The most valuable participation is not agreement, but professional work sufficient to change the theory.

The Civilizational Future Chain | What World Might People Live in If These Problems Are Better Solved?

To explore the world that might emerge when “technological capability remains constrained by real material boundaries while genuinely serving human life” extends into work, family, education, communities, public life, and long-term civilization, continue to the Living Civilization trilogy and the future-civilization thread in the five-volume series. These offer visions of future life that can be examined, revised, and rejected; they are not evidence for this paper's ecological or Earth-system facts. A vision is not evidence. Research must nevertheless ultimately answer where we hope these facts, taken together, will lead civilization.

<https://www.civitas.top/en/documents/p04-living-001/> | <https://www.civitas.top/en/documents/p02-five-001/>

Civilization Leap Five-Volume Series: <https://www.civitas.top/en/documents/p02-five-001/> | Public gateway: <https://www.civitas.top/en/documents/p01-entry-006/> | Complete library: <https://www.civitas.top/en/library/>

The aim is not for more people to finish reading Civilization Leap, but for more people to complete Civilization Leap together.

Appendix A | Register of Source–Claim Boundaries

Table A1. Principal sources and the scope of claims they support

Source	Claims adopted in this paper	Explicit stopping point
IEA 2025/2026 [1–2]	Global data-center electricity, AI drivers, energy supply, scenarios, and bottlenecks	Supports current assessments and 2030 scenarios; not a certain future
LBNL 2026 [3]	Bottom-up model and scenario range for US data-center electricity consumption in 2030	Applies only to its US equipment/facility assumptions; should be updated with shipments and utilization
LBNL water research [6]	Determinants of workload water consumption and differences exceeding 10,000-fold	Supports rejecting a single water-per-task constant; does not imply extreme differences at every facility
Xiao et al. [9]	Water–carbon scenarios and spatial differences for US AI servers	Model scenarios, not measured national totals
Chien et al. [10]	Synthesis and trade-offs concerning energy, water, and embodied environmental impacts of AI data centers	Review supports multiobjective methods; specific shares depend on system boundaries
Colangelo et al. [11]	Field experiment on load flexibility with 256 GPUs	Demonstrates that certain workloads are adjustable; not generalizable to all AI loads
IEA critical minerals [13–14]	Material demand and supply concentration	Indicates possible material bottlenecks; does not imply inevitable AI shortages
E-waste Monitor [16]	Global e-waste quantities and recycling context	Not AI-specific waste
Wang et al. [17]	Generative-AI e-waste scenarios for 2020–2030	Category B scenarios, not certain facts for 2030
Planetary boundaries [22–24]	Global Earth-system risk framework	Not an automatic permitting threshold for individual projects
IPCC/IPBES [25–26]	Climate, biodiversity, and systemic-risk context	Does not replace attribution specific to AI
Ireland/Virginia/Singapore [36–38]	Real cases of regional carrying capacity	Establishes regional differences and policy responses only; not globally generalizable

Appendix B | AI Resources and Environment Stress-Test Matrix

Table B1. Resources and environment stress-test matrix

Scenario	Primary environmental dimension	Required measurement	Discipline that must re-enter	Main failure mode
Large new AI campus	Electricity	Nodal capacity/transmission/generation/peaks	Regional energy systems	Sufficient total electricity, but delayed connections or higher costs or emissions
Cooling in a highly water-stressed region	Water	Direct consumption/indirect electricity-related water/seasonal watershed stress	Hydrology/environmental engineering	Low WUE but higher electricity use or carbon; annual replenishment mismatched to the dry season
Accelerated hardware replacement	Materials/e-waste	Lifetimes, manufacturing, redeployment, recycling	Industrial ecology/LCA	Better unit energy efficiency but higher embodied environmental impacts and waste
Supporting low-carbon energy	Climate/land/materials	Actual electricity sources, transmission, storage, construction	Energy/ecology/materials	Lower operational carbon but greater land/material pressures
Flexible AI loads	Grid	Task latency, migration, SLAs, data location	Power systems/AI engineering	Technically adjustable loads, but high-security or real-time tasks cannot be adjusted
AI for energy savings	Net environmental benefits	Baselines, attribution, deployment rates, rebound	Energy/economics/organizations	Experimental savings cannot enter real systems or are offset by additional demand
Concentration of critical minerals	Materials/security	Demand shares, supply concentration, substitution, inventories	Resources/international relations	Small material volumes but strategically important supply shocks
Ecologically sensitive siting	Land/biodiversity	Habitats, recoverability, cumulative infrastructure	Ecology/life sciences/social research	Low-impact server rooms but supporting infrastructure imposes high ecological costs

Declaration of Generative AI Use

Generative AI may have assisted in organizing materials and literature leads, designing the structure, drafting and polishing some passages, cross-checking, verifying citation leads, and checking layout during the preparation of this paper. AI output is not treated as empirical data, an authoritative source, or independent evidence. The author reviews and approves the research questions, theoretical propositions, normative judgments, selection of materials, use of citations, and final text, and assumes final academic responsibility. AI is not listed as an author.

This paper adds a discipline specific to resources and environment: no energy-use, water-use, carbon-emissions, material-demand, critical-mineral-share, ecological-threshold, or Earth-system conclusion supplied by generative AI may be inserted directly into the text. All quantitative material must be traced to original research, official statistics, standards, professional databases, or verifiable technical reports, retaining time, location, system boundaries, methods, and scenario conditions. At formal submission, details of specific tools, model versions, use dates, and scope of use should be added according to the target journal's requirements.

References

- [1] International Energy Agency (IEA). (2025). Energy and AI. Paris: IEA. <https://www.iea.org/reports/energy-and-ai>
- [2] International Energy Agency (IEA). (2026). Key Questions on Energy and AI. Paris: IEA. <https://www.iea.org/reports/key-questions-on-energy-and-ai>
- [3] Smith, S. J., Hubbard, A., Newkirk, A., Ganeshalingam, M., Holecek, B., Sartor, D. A., Mills, M., & Shehabi, A. (2026). United States Data Center Energy Usage Report: 2025 Update. Lawrence Berkeley National Laboratory. <https://doi.org/10.71468/P1RP4F>
- [4] Shehabi, A., Smith, S. J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M. A. B., Holecek, B., Koomey, J. G., Masanet, E. R., & Sartor, D. A. (2024). 2024 United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory. <https://doi.org/10.71468/P1WC7Q>
- [5] Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). Recalibrating global data center energy-use estimates. *Science*, 367(6481), 984–986. <https://doi.org/10.1126/science.aba3758>
- [6] Lei, N., Lu, J., Shehabi, A., & Masanet, E. R. (2025). The water use of data center workloads: A review and assessment of key determinants. *Resources, Conservation and Recycling*, 219, 108310. <https://doi.org/10.1016/j.resconrec.2025.108310>
- [7] Mytton, D. (2021). Data centre water consumption. *npj Clean Water*, 4, 11. <https://doi.org/10.1038/s41545-021-00101-w>
- [8] Siddik, M. A. B., Shehabi, A., & Marston, L. (2021). The environmental footprint of data centers in the United States. *Environmental Research Letters*, 16, 064017. <https://doi.org/10.1088/1748-9326/abfba1>
- [9] Xiao, T., Nerini, F. F., Matthews, H. D., Tavoni, M., & You, F. (2025). Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the USA. *Nature Sustainability*, 8, 1541–1553. <https://doi.org/10.1038/s41893-025-01681-y>
- [10] Chien, A. A., Gupta, U., Ren, S., Sriraman, A., & Tomlinson, B. (2026). Strategies and design for increasing AI sustainability. *Nature Reviews Clean Technology*. <https://doi.org/10.1038/s44359-026-00195-w>
- [11] Colangelo, P., Coskun, A. K., Megrue, J., et al. (2026). AI data centres as grid-interactive assets. *Nature Energy*, 11, 254–261. <https://doi.org/10.1038/s41560-025-01927-1>
- [12] Granderson, J., Hoffman, I. M., Holecek, B., Crowe, E., Smith, S. J., & Mims Frick, N. (2026). Integrating AI Data Centers with the Power Grid. *The Bridge*. <https://doi.org/10.20357/B7X61S>
- [13] International Energy Agency (IEA). (2025). Global Critical Minerals Outlook 2025. Paris: IEA. <https://www.iea.org/reports/global-critical-minerals-outlook-2025>
- [14] International Energy Agency (IEA). (2026). Global Critical Minerals Outlook 2026. Paris: IEA. <https://www.iea.org/reports/global-critical-minerals-outlook-2026>
- [15] United Nations Environment Programme, International Resource Panel. (2024). Global Resources Outlook 2024: Bend the Trend – Pathways to a Liveable Planet as Resource Use Spikes. Nairobi: UNEP. <https://www.unep.org/resources/Global-Resource-Outlook-2024>
- [16] International Telecommunication Union (ITU) & United Nations Institute for Training and Research (UNITAR). (2024). The Global E-waste Monitor 2024. Geneva/Bonn. https://www.itu.int/hub/publication/d-gen-e_waste-01-2024/
- [17] Wang, P., Zhang, L.-Y., Tzachor, A., & Chen, W.-Q. (2024). E-waste challenges of generative artificial intelligence. *Nature Computational Science*, 4, 818–823. <https://doi.org/10.1038/s43588-024-00712-6>
- [18] Plociennik, C., Watjanatepin, P., Van Acker, K., & Ruskowski, M. (2025). Life Cycle Assessment of Artificial Intelligence Applications: Research Gaps and Opportunities. *Procedia CIRP*, 135, 924–929. <https://doi.org/10.1016/j.procir.2025.01.079>

- [19] International Organization for Standardization (ISO). (2006). ISO 14040: Environmental management — Life cycle assessment — Principles and framework.
- [20] International Organization for Standardization (ISO). (2006). ISO 14044: Environmental management — Life cycle assessment — Requirements and guidelines.
- [21] International Organization for Standardization (ISO). (2014). ISO 14046: Environmental management — Water footprint — Principles, requirements and guidelines.
- [22] Rockström, J., Steffen, W., Noone, K., et al. (2009). A safe operating space for humanity. *Nature*, 461, 472–475. <https://doi.org/10.1038/461472a>
- [23] Steffen, W., Richardson, K., Rockström, J., et al. (2015). Planetary boundaries: Guiding human development on a changing planet. *Science*, 347(6223), 1259855. <https://doi.org/10.1126/science.1259855>
- [24] Richardson, K., Steffen, W., Lucht, W., et al. (2023). Earth beyond six of nine planetary boundaries. *Science Advances*, 9(37), eadh2458. <https://doi.org/10.1126/sciadv.adh2458>
- [25] IPCC. (2023). Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report. Geneva: IPCC. <https://doi.org/10.59327/IPCC/AR6-9789291691647>
- [26] IPBES. (2019). Global Assessment Report on Biodiversity and Ecosystem Services. Bonn: Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. <https://doi.org/10.5281/zenodo.3831673>
- [27] Sorrell, S. (2009). Jevons' Paradox revisited: The evidence for backfire from improved energy efficiency. *Energy Policy*, 37(4), 1456–1469. <https://doi.org/10.1016/j.enpol.2008.12.003>
- [28] van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 1, 213–218. <https://doi.org/10.1007/s43681-021-00043-6>
- [29] Rolnick, D., Donti, P. L., Kaack, L. H., et al. (2022). Tackling Climate Change with Machine Learning. *ACM Computing Surveys*, 55(2), Article 42. <https://doi.org/10.1145/3485128>
- [30] Nuss, P., & Eckelman, M. J. (2014). Life cycle assessment of metals: a scientific synthesis. *PLOS ONE*, 9(7), e101298. <https://doi.org/10.1371/journal.pone.0101298>
- [31] Frosch, R. A., & Gallopoulos, N. E. (1989). Strategies for manufacturing. *Scientific American*, 261(3), 144–152.
- [32] International Energy Agency (IEA). (2026). Critical Minerals Dataset, 2026 edition. Paris: IEA. <https://www.iea.org/data-and-statistics/data-product/critical-minerals-dataset>
- [33] OECD. (2022). Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint. OECD Digital Economy Papers, No. 341. <https://doi.org/10.1787/7babf571-en>
- [34] ISO/IEC. (2026). ISO/IEC 30134-2:2026 Information technology — Data centres key performance indicators — Part 2: Power usage effectiveness (PUE).
- [35] ISO/IEC. (2022). ISO/IEC 30134-9:2022 Information technology — Data centres key performance indicators — Part 9: Water usage effectiveness (WUE).
- [36] Central Statistics Office Ireland. (2026). Data Centres Metered Electricity Consumption 2025. Statistical Release, 7 July 2026. <https://www.cso.ie/en/releasesandpublications/ep/p-dcmec/datacentresmeteredelectricityconsumption2025/>
- [37] Joint Legislative Audit and Review Commission (JLARC), Commonwealth of Virginia. (2024). Data Centers in Virginia. Richmond, VA. <https://jlarc.virginia.gov/landing-2024-data-centers-in-virginia.asp>
- [38] Infocomm Media Development Authority (IMDA), Singapore. (2024). Green Data Centre Roadmap. <https://www.imda.gov.sg/-/media/imda/files/news-and-events/media-room/media-releases/2024/05/green-dc-roadmap.pdf>

Version Note

V1.1 makes the necessary controlled revisions following an independent professional review of V1.0, the first complete professional version: it specifies the 256-GPU grid-interaction field demonstration more precisely; explicitly connects IPBES biodiversity evidence to the land and ecological siting chapter; separates the 2025 scenario for data-center gallium demand from the 2026 update on critical-mineral supply concentration; corrects the DOI of Xiao and colleagues' 2025 paper; and adds stable entry points for dynamic and formal sources. The theoretical core, object of study, A/B/C/F categories, established disciplinary foundations, interdisciplinary handoffs, prohibitions on overreach, refutation conditions, upward/downward/leftward/rightward plus forward and civilizational-future chain, and generative AI use declaration remain unchanged. This same-version cross-paper consistency update only synchronizes the version of the real-world convergence paper and the unified public entry points; it does not change ecological judgments. V1.1 is the current professional research baseline.