

An MIT Exploration of Generative AI • From Novel Chemicals to Opera

The Climate and Sustainability Implications of Generative AI

**Noman Bashir¹ Priya Donti^{2,3} James Cuff⁴ Sydney Sroka¹ Marija Ilic^{2,3}
Vivienne Sze^{4,5,6,7} Christina Delimitrou⁸ Elsa Olivetti^{1,9}**

¹MIT Climate and Sustainability Consortium, Massachusetts Institute of Technology, Cambridge, MA,

²Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA,

³Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA,

⁴Office of Research Computing and Data, Massachusetts Institute of Technology, Cambridge, MA,

⁵Research Lab of Electronics, Massachusetts Institute of Technology, Cambridge, MA,

⁶Microsystems Technology Laboratories, Massachusetts Institute of Technology, Cambridge, MA,

⁷and MIT Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA,

⁸MIT Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA,

⁹Department of Materials Science & Engineering (DMSE), Massachusetts Institute of Technology, Cambridge, MA

MIT

Published on: Mar 27, 2024

URL: <https://mit-genai.pubpub.org/pub/8ulgrckc>

License: [Creative Commons Attribution-NonCommercial 4.0 International License \(CC-BY-NC 4.0\)](https://creativecommons.org/licenses/by-nc/4.0/)

ABSTRACT

The rapid expansion of generative artificial intelligence (Gen-AI) is propelled by its perceived benefits, significant advancements in computing efficiency, corporate consolidation of artificial intelligence innovation and capability, and limited regulatory oversight. As with many large-scale technology-induced shifts, the current trajectory of Gen-AI, characterized by relentless demand, neglects consideration of negative effects alongside expected benefits. This incomplete cost calculation promotes unchecked growth and a risk of unjustified techno-optimism with potential environmental consequences, including expanding demand for computing power, larger carbon footprints, shifts in patterns of electricity demand, and an accelerated depletion of natural resources. This prompts an evaluation of our currently unsustainable approach toward Gen-AI's development, underlining the importance of assessing technological advancement alongside the resulting social and environmental impacts. Presently, efforts to boost computing sustainability largely focus on efficiency improvements, including enhancing hardware energy efficiency, refining artificial intelligence algorithms, and improving the carbon efficiency of executing computing workloads through spatiotemporal workload shifting. In the presence of relentless demand and prioritization of economic growth, this siloed focus on efficiency improvements results instead in increased adoption without fundamentally considering the vast sustainability implications of Gen-AI. We argue that responsible development of Gen-AI requires a focus on sustainability beyond only efficiency improvements and necessitates benefit–cost evaluation frameworks that encourage (or require) Gen-AI to develop in ways that support social and environmental sustainability goals alongside economic opportunity. However, a comprehensive value consideration is complex and requires detailed analysis, coordination, innovation, and adoption across diverse stakeholders. Engaging stakeholders, including technical and sociotechnical experts, corporate entities, policymakers, and civil society, in a benefit–cost analysis would foster development in directions that are most urgent and impactful while also reducing unsustainable practices.

▶ 0:00 / 56:32 ————— 🔊 ⋮

🔊 Listen to this article

1. Unfettered Growth and Its Drivers

Generative artificial intelligence (AI) (Gen-AI) has become a ubiquitous, global phenomenon in modern society, with significant increases in the number and diversity of use cases being implemented ([Chui et al. 2023](#)). This specific, albeit disruptive, set of machine learning (ML) algorithms has captured the mind share and the focus of researchers, scientists, and corporations alike ([Chui et al. 2023](#)), comparable to the Klondike gold rush at the dawn of the twentieth century ([Berton 2011](#)). Many predict this as the end of the AI winter and

the dawn of a new age of intelligence ([Entefy 2023](#); [Knight 2023](#)). However, the reality behind this excitement and unbounded growth is significantly more complex and multifaceted. The growth of Gen-AI is driving increased electricity demand, which runs counter to the massive efficiency gains that are needed to achieve net-zero greenhouse gas emissions by 2050 in energy-related sectors (alongside simply “greening the grid”) ([IEA 2021](#)). Our current capacity to build sustainably also cannot keep pace with the datacenter construction necessary to support Gen-AI. Furthermore, this explosive growth exacerbates supply chain issues, affecting essential goods and services that rely on the tech industry and potentially creating macroeconomic impacts.

The rapid expansion of Generative AI is reflected in the rising demands on data centers. The datacenter capacity under construction in North America, measured using the datacenter power requirement, increased from 2,688 MW at the end of 2022 to 5,341 MW at the end of 2023 ([Beets and Hartnett 2024](#)). This is in addition to the existing demand for datacenters that are expected to add a staggering 12,000 MW of co-location capacity ([Uptime 2024](#)), exceeding the demand of over 70% of countries ([GlobalEconomy.com 2021](#)). Based on conservative estimates, datacenters' energy demand increased from 194 TWh in 2010 to 204 TWh in 2018 ([Masanet et al. 202a](#)) — a 6% growth in energy consumption despite 6x increase in datacenter capacity, primarily due to massive increases in energy efficiency gains. However, the global electricity consumption of datacenters rose to 460 TWh in 2022. Depending on the efficiency improvements, cryptocurrency trends, and artificial intelligence demand, the global electricity consumption of datacentres is expected to range between 620 — 1,050 TWh in 2026, with the base case for demand at just over 800 TWh ([IEA 2024](#)), reflecting a demand that outpaces even the most optimistic forecasts of technological improvements. It is worth noting that the current pace of energy consumption growth is potentially under-reflective of actual demand, given limitations on the availability of AI chips, multi-year-long lead times for datacenter equipment ([Uptime 2024](#)), and power availability constraints ([Beets and Hartnett 2024](#)).

A contemporaneous alignment of several factors beyond simply Gen-AI’s perceived benefits, including consolidation of AI power, limited regulatory oversight, and efficiency improvements, may explain this unfettered and unbounded growth. By dissecting the enablers of this growth, we can identify the key stakeholders who are instrumental in shaping the trajectory of Gen-AI development. This analysis is crucial for understanding the broader implications of Gen-AI’s proliferation, including its societal, ethical, and environmental impacts. Furthermore, recognizing these driving forces provides essential context for the discussions in Section 3, where we delve into the responsibilities and actions these stakeholders can take to foster a sustainable and ethical Gen-AI ecosystem.

1.1 Perceived Benefits of Generative AI

Hugely popular implementations of a specific type of Gen-AI model, the Generative Pre-trained Transformer (GPT), have allowed rapid inference packaged in easy-to-use interfaces, capturing widespread attention globally. By inputting simple text prompts, a user is instantly presented with what appears to be a magical output in response to any conceivable question. From a user’s perspective, it may seem that endless knowledge,

images, and information can now be automatically and immediately generated at their fingertips. Beyond leisurely use, there is genuine optimism that Gen-AI will unlock large-scale efficiency gains, productivity enhancements, and innovations ([Knight 2023](#)). The hype surrounding Gen-AI implies that no sector will forgo Gen-AI's transformative benefits ([Chui et al. 2023](#)).

Despite this potential, there is a growing realization that not every Gen-AI application will be inherently beneficial or realize its anticipated advantages ([Nature editorial authors 2024](#); [Bender et al. 2021](#)). Stakeholders, ranging from technology giants to startups, heavily invest in this technology, betting on its transformative impact. However, this investment is often predicated on optimism about Gen-AI's capabilities and utility, which may not always align with practical outcomes. In the worst case, Gen-AI's perceived benefits can be used as an argument for limited regulation of its direct impacts or delayed adoption of alternative sustainable approaches across sectors, even when the expected benefits could be minor relative to the environmental impact ([Dauvergne 2022](#); [Ipsen et al. 2019](#)). Gen-AI applications can also be actively counterproductive to society, such as by facilitating the spread of misinformation or delaying the retirement of fossil fuel-based power plants ([Kaack et al. 2022](#)). This highlights the need for a balanced perspective, acknowledging the potential, limitations, and risks of Gen-AI in societal applications. As technology evolves, its exact benefits and impact remains the subject of keen observation and ongoing assessment.

1.2 Consolidation of AI Capabilities

The quality of Gen-AI algorithms today correlates with the size of required computing systems, necessitating a larger number of or more powerful sets of computers for more sophisticated data sets and models. Because the correlation appears simple, a few for-profit organizations, notably AI giants, are deploying more datacenters and computing infrastructure than ever before. For instance, in the third quarter of 2023, Microsoft and Meta each bought three times more NVIDIA graphics processing units (GPU) than Amazon and Google, which acquired 50,000 units each ([Norem 2023](#)). The larger the investment, the higher will be the perceived quality and potential return on investment. Furthermore, AI giants continue consolidating AI power through strategic partnerships ([Ward and Lung 2023](#)), such as Microsoft with OpenAI and Amazon with Anthropic. This leads to global growth in model training and increased use of consumer-facing web-based inference engines. Capital, equipment, resources, and energy are effectively viewed as fuel to meet the demands of Gen-AI algorithms, driving an insatiable need for more of these inputs, predominantly procured and/or fulfilled by these AI powerhouses.

This unprecedented computational power also allows for handling the large, diverse datasets that are crucial for the competitive edge of major tech companies in AI development ([Clarke 2023](#); [Kak and West 2023](#)). For instance, Google's use of web data for BERT ([Devlin et al. 2018](#)) and OpenAI's utilization of varied text sources for GPT models ([OpenAI 2023](#)) demonstrate the impact of large datasets on AI advancements. This data control, coupled with the proprietary nature of such data ([OpenAI 2023](#)), is central to sophisticated model development and a significant focus of the industry's AI policy. Moreover, the concentration of computing

power and data enables these giants to attract top AI talent, as seen in the significant year-on-year increase in AI personnel at companies such as Amazon, Microsoft, and Meta ([glass.ai 2023](#)). At the end of 2022, these companies employed more than 33,000 AI personnel, with almost 8,000 people in the roles of core AI research scientists ([glass.ai 2023](#)), fostering a self-sustained ecosystem of innovation. We note that on the one hand consolidation could enable efficiency improvements and centralize reporting or auditing activities. On the other hand, consolidation also means that it is challenging to devise effective incentives and increase the social sustainability of the technology, and it comes with negative implications for the equity of technological trajectories and outcomes.

1.3 Limited Regulatory Oversight

The unfettered growth in Gen-AI has notably outpaced global regulatory efforts, leading to varied and insufficient oversight of its socioeconomic and environmental impact ([Jelinek, Wallach, and Kerimi 2021](#); [Scherer 2015](#); [West 2023](#)). Many countries are passing regulatory measures, including those of the European Union (EU), South Korea, Brazil, Singapore, and the United States ([Kremer et al. 2023](#)). The EU has passed the EU AI Act that aims to address risks to health, safety, the environment, and fundamental rights ([“Artificial Intelligence Act” 2023](#); [EU Council 2023](#)). Although the EU act calls for robust and transparent accounting of emissions for AI systems, like many other jurisdictions and sectors, there is not yet an explicit call to limit emissions ([“Artificial Intelligence Act” 2023](#)). The Greenhouse Gas Protocol, an essential international framework for managing emissions, fails to adequately address the technology sector’s unique challenges ([Becker et al. 2022](#); [Mytton 2020](#)), leading to substantial underreporting of greenhouse gas (GHG) emissions by technology companies ([Klaaßen and Stoll 2021](#)). Other social and environmental impacts, such as water usage, receive even less regulatory attention ([Coeckelbergh 2021](#); [Kak and West 2023](#)).

Fully characterizing the social and environmental impacts of Gen-AI is complex and hinders targeted regulations. Recent US initiatives, such as the White House executive order on climate-related financial risk ([Executive Order 14030 2021](#)) and the CHIPS Act ([“Chips and Science Act, H.R.4346” 2021](#)), highlight these issues but lack comprehensive guidelines for Gen-AI’s broader impacts. The CHIPS Act, focusing on semiconductor manufacturing, only indirectly addresses the technology industry’s environmental responsibilities. Furthermore, even if social and environmental legislation is present, its implementation may differ across regions, posing unique challenges for regulators and risk management professionals in light of Gen-AI’s rapid evolution ([Cihon, Maas, and Kemp 2020](#); [Kremer et al. 2023](#)). This scarcity of actionable regulations for Gen-AI’s impacts limits effective oversight, inadvertently enabling rapid AI technology adoption without sufficient environmental and social accountability. Adequate and targeted regulations require a comprehensive and comparative evaluation capability that weighs Gen-AI’s potential societal benefits against the costs of Gen-AI’s unfettered growth.

1.4 Efficiency Improvements

The rapid performance growth of general purpose GPU, sophisticated low-latency and high-bandwidth communications ([Pierce 2020](#); [Dean et al. 2012](#)), and specialized hardware ([Jouppi et al. 2017](#)) have collectively enabled the execution of large multibillion parameter models in significantly reduced time ([Dally 2023](#); [Perry 2018](#)). Key to this transformation is the transformer architecture ([Vaswani et al. 2017](#)) that enables AI models to scale to billions of parameters with a sublinear increase in the associated computational costs. Further enhancements, such as Mixture-of-Experts models have optimized the computational efficiency of models and allowed them to scale to trillions of parameters with manageable computational requirements ([Du et al. 2021](#); [Shazeer et al. 2017](#)). Efficient model architectures, combined with distributed training techniques, such as pipeline parallelism ([Huang et al. 2019](#); [Dean et al. 2012](#); [Abadi et al. 2016](#); [Brown et al. 2020](#)), have facilitated the creation of giant Gen-AI models. This progress is supported by improvements in ML training, including Adam optimization ([Kingma and Ba 2014](#)), dropout ([Srivastava et al. 2014](#)), batch normalization ([Ioffe and Szegedy 2015](#)), and the emergence of transfer learning and fine-tuning ([Devlin et al. 2018](#); [Krizhevsky, Sutskever, and Hinton 2012](#)). Furthermore, hyperscale datacenter growth, capabilities, and efficiency improvements, such as parallel processing, have allowed significant increases in the implementation of large Gen-AI models, triggering further unbounded growth.

Unfortunately, these efficiency gains have not reduced Gen-AI's overall energy consumption because of the implications of rebound effects and the Jevons paradox ([Hintemann 2018a](#); [Masanet et al. 2020a](#)). Furthermore, the expansion of Gen-AI tools such as ChatGPT and Bard, along with 33% of the global population still being offline ([ITU 2023](#)), indicates that demand and energy use is far from saturation. To address the climate impact of this demand, academic and industrial research has focused on enhancing the energy efficiency ([Bashir et al. 2023](#); [Katal, Dahiya, and Choudhury 2023](#); [Mammen et al. 2023](#); [Patterson et al. 2022](#); [Wu et al. 2022](#); [Li et al. 2023](#)) and carbon efficiency of computational technologies ([Acun et al. 2023](#); [Bashir et al. 2021](#); [Hanafy et al. 2023](#); [Lechowicz et al. 2024](#); [Radovanović et al. 2023](#); [Chakrabarty et al. 2023](#); [Switzer et al. 2023](#); [Thiede et al. 2023](#); [Wiesner et al. 2021](#); [Sukprasert et al. 2024](#)). However, whereas efficiency gains remain critical, a narrow focus on them does not solve the problem; instead, it can exacerbate it by encouraging further (unbounded) growth ([Birhane et al. 2022](#); [Giampietro and Mayumi 2018](#); [Wright et al. 2023](#)).

As discussed, the environmental and socioeconomic sustainability implications of Gen-AI are complex. While Gen-AI has the potential to provide tangible benefits for various sectors and applications, its unfettered growth incurs significant costs. The current approach of growing the Gen-AI sector to satisfy every imaginable application considers neither what benefits have actually been realized in practice nor the extensive societal costs. We call for the sustainable development of Gen-AI and propose a comparative benefit-cost evaluation framework as a potential approach toward responsible development in Gen-AI.

2. Need for Comparative Evaluation Capability

Recent legislative and regulatory efforts are sharpening the focus on the sustainability of Gen-AI's growth. The newly adopted rules by the Securities and Exchange Commission (SEC) compel public companies to disclose material climate-related risks and their Scope 1 and Scope 2 emissions, a step towards greater transparency for all stakeholders, and especially investors ([SEC 2024](#)). Concurrently, a bill proposed by Senator Ed Markey is pushing for a closer examination of AI's environmental footprint, signaling a legislative intent to steer AI development toward sustainable practices ([Artificial Intelligence Environmental Impacts Act of 2024](#)). These measures align with the broader evaluation framework proposed in our work, which aims to encompass a spectrum of costs—including sustainability—and benefits of Gen-AI across diverse stakeholders and sectors. This shift towards integrating sustainability in Gen-AI's discourse and policy may inform more balanced and responsible technological progress.

To instill accountability for more sustainable Gen-AI practices, stakeholders must assemble to provide guidance and inform decision-making that weighs societal benefits against societal costs (sustainability and otherwise), to shape further development. Here, we outline elements of an evaluation framework, and in Section 3, we describe stakeholders' roles in development and execution of such a framework. As much as possible, this guidance should offer a comprehensive, comparative evaluation capability that includes costs and perceived benefits across multiple stakeholders (individual, organizational, or regional actors), sectors, and contexts as others have indicated ([Richards et al. 2023](#)). Although the costs of Gen-AI extend beyond sustainability costs alone (including, e.g., costs related to labor, privacy, and copyright infringement ([Luccioni 2023](#)), here, we provide insight (primarily) on sustainability-related costs (we use the term cost to specifically delineate negative impacts from benefits or positive impacts).

The first-order costs associated with Gen-AI relate to direct computing-related impacts ([Kaack et al. 2022](#)) from cradle to grave. These include the materials used for everything from individual semiconductors to datacenter infrastructure, manufacturing processes and distribution, energy associated with powering the computing devices, and waste management at the end of product life. These computing-related impacts result in energy, water, and materials use as well as emissions to land, air, and water, which can lead to depletion of natural resources and damage to human health and ecosystems. Accounting for these costs is typically done through life cycle assessment (LCA) ([Finkbeiner et al. 2006](#)), which includes steps to define a study goal and scope, account for the life cycle inventory at each step, assess impact along a defined set of metrics, and interpret the results. LCA provides basic guidelines to perform such an analysis but leaves much open to the practitioner. Stakeholders for a particular sector then assemble to define specific rules for a product or service of interest, termed “product category rules” ([Ingwersen and Subramanian 2014](#)).

Although study differences can make summary statements challenging, there is contemporary consensus that the highest computing-related costs are in the manufacturing (or embodied) phase and use (or operational) phase, including datacenter cooling and infrastructure ([Itten et al. 2020](#); [Clément, Jacquemotte, and Hilty](#)

[2020](#)). More specifically, embodied impacts are driven by the fabrication of integrated circuits including high-performance processors and high-density memory semiconductor devices. Water consumption is correlated with electricity consumption throughout the life cycle and device manufacture. Whether cost is higher for embodied versus operational phases is determined by whether computations are occurring on a datacenter or an edge device. Computing sector GHG emissions will likely shift more toward embodied emissions dominance for both consumer devices and datacenters as computing hardware becomes increasingly efficient, software and algorithms are optimized, and cleaner energy is used to power datacenter operations ([Das and Mao 2020](#); [Belkhir and Elmeligi 2018](#)). For this reason, measuring impact not only for carbon and energy, but also across multiple environmental metrics will become critical (for example, the EU Product Environmental Footprint guidelines recommend nineteen impact categories). Sustainability costs also include unintended consequences resulting from the use of models for an immediate application or broadly at a system or structural-level (LCA practitioners use the term consequential or indirect) costs. In this latter category, others have suggested that sustainability costs result from extending (or locking in) impact-intensive technologies, sectors, and energy sources; accelerating consumption through consumer behavior; and miseducation through faster spread of climate-negative information ([Kaack et al. 2022](#)). System-level impacts induced by Gen-AI include those relevant to any form of innovation, including rebound, rematerialization, learning and scale effects, technology evolution, or cultural shifts. Costs dominate the computing-related impacts, whereas immediate applications and system-level impacts can result in benefits ([Rolnick et al. 2022](#)) and/or costs depending on the specifics of the application.

A framework to assess intended benefits (mentioned in Section 1) and costs (mentioned above) must balance quantitative data with qualitative assessments to aim for a comprehensive evaluation of Gen-AI's impact. This is a highly complex undertaking, but even steps toward this capability will promote necessary transparency and discourse among stakeholders. Here, we suggest three framework elements, including defining the scope and boundaries, developing baseline and scenarios, and building data inventory for accounting alongside examples of where a starting set of existing frameworks could be leveraged, adopted, or adapted. Articulating a complete framework is beyond the scope of this document. Instead, we provide some starting points for deep collaborative investigation by the Gen-AI and industrial ecology communities and their broader stakeholder groups.

2.1 Scope and Boundaries

We suggest that estimating benefit and cost should align with LCA methodology (but leverage methods beyond LCA) and begin by articulating the analysis goal and scope, which provides the structure for the materials, processes, or products considered. Through this exercise, one would define the intended audience and desired metrics, specify a unit of analysis (to enable comparison), as well as outline the conceptual, geographic, and temporal boundaries.

An essential criterion for comparability is the unit of analysis, based on the desired performance (termed functional unit in LCA) and the amount of a particular product or service (a volume of paper bag for carrying groceries, for example) required to meet that performance (termed “reference flow” in LCA) ([Andrae and Andersen 2010](#)). As with every dimension of an LCA, there are several ways to develop useful functional units, and recent efforts to make them comparable recommend a practiced approach for screening multiple options ([Furberg, Arvidsson, and Molander 2022](#)). Within Gen-AI, for example, a unit of analysis could be a query result. Ultimately, this functional unit is translated to a reference flow, such as floating-point operation per unit of power, to deliver that result, which is directly informed by the type of hardware, the type of task, the model used, and the specific characteristics of a dataset (including product power draw and system infrastructure and inefficiencies) ([Debus et al. 2023](#)).

Defining the boundaries of a study completely and consistently is essential for comparative evaluation, but it is also tremendously complex ([Furberg, Arvidsson, and Molander 2022](#)). This is particularly true for Gen-AI, given the interdependency of supply chains and cascading implications of applications. The period that a particular study covers (temporal boundaries) should be short in duration, and the choice of these boundaries should be revisited frequently, given the pace at which Gen-AI is developed and adopted. The geography that a study represents (geographic boundaries) can be guided by spatial delineation where data are currently collected and then push for a broader or more granular scope, depending on the study goal. The most challenging boundaries are conceptual in nature (as indicated by the levels described above: computing-related and application-related, including immediate application and system level) ([Kaack et al. 2022](#)). Within LCA, the boundaries that account for application-related impact (both for benefit and cost) are informed by accounting for “what has changed” based on the introduction of a product, technology, or service. This is described by the LCA community as a consequential analysis and accounts for what has been “displaced” ([Weidema 2003](#)). More specifically, such an analysis defines the main affected marginal player in both the short and long term. Determining this marginal impact (i.e., what is the displaced or competing product) is the most important aspect of such an analysis. The conceptual boundary, therefore, includes material and energy flows directly or indirectly affected by the change. Examples of where consideration of marginal effects is critical include cases of constrained resources (steep supply curves, where each additional new supply is much more expensive than the previous), where timing and scale matter (i.e., electricity use for which a new plant will be built or turned on), and when growth trajectories are high ([“Life Cycle Inventory Analysis,” 2021](#)). The pervasive challenge of limited data availability, spatial nature of datacenter energy consumption, and diversity in study scope argue for emphasizing narrow approaches to consistent boundary definition over global assessments. One recent idea is to develop boundaries that enable what the authors term “relational footprinting” that aligns with discrete (and perhaps more measurable) geographic, spatial, technical, and social units ([Pasek, Vaughan, and Starosielski 2023](#)). This relationship-based approach resonates with the need to bring affected communities into a conversation regarding the system-level costs of Gen-AI ([Debnath et al. 2023](#)).

2.2 Baselines and Scenarios

The largest methodological need for effective benefit–cost determination for Gen-AI is to develop specific baselines from which comparisons can be made. For each Gen-AI application, communities of practitioners and users should specify what business as usual was before the introduction of a new algorithmic capability. As Gen-AI is a subset of ML capabilities more broadly, one would need to differentiate capabilities that data-driven approaches have enabled in terms of quality enhancement or increased productivity versus those resulting from Gen-AI specifically. A baseline would demonstrate Gen-AI application benefits and costs over business as usual, whereas a scenario would provide comparisons across multiple Gen-AI approaches ([Norris et al. 2021](#)). This seemingly intractable task should be initiated through sets of collectively articulated baselines and scenarios by application, geography, and time period that evolve with study needs and capture broad cost implications.

The ML community is well-versed in the value of benchmarks for making comparisons in model performance. Although Gen-AI use cases are far from the data and task delineation of formal ML benchmarking, we must move the community toward this level of specificity to accurately understand the benefit–cost implications of Gen-AI. Initial precedents are emerging in sustainability-relevant domains such as transportation ([Vinitsky et al. 2018](#)), scientific discovery ([Fung et al. 2021](#)), and soil carbon ([Wijewardane et al. 2016](#)). And guidance can be sourced from processes to develop climate change scenarios, such as the development of Shared Socioeconomic Pathways ([Riahi et al. 2017](#)) or sector-specific road-mapping efforts.

Once baselines are established by application, scenarios account for what would have happened without the actor being an agent of change ([Norris et al. 2021](#)). Scenarios are typically defined by variables that change from one point in time to the next and are coupled with a narrative justifying the change ([Fauré et al. 2017](#)). These variables should be defined relative to the scope and boundaries of the study and articulate whether a variable change is explicit to a Gen-AI benefit or cost. These scenarios should be dynamic, as further innovations and mitigation efforts continue ([Börjesson Rivera et al. 2014](#)). Scenarios should be framed alongside geographically relevant strategies for GHG emissions reduction, for example, to understand whether benefits persist beyond proposed emissions reductions, and should consider multiple forms of potential costs related to extending emissions-intensive sectors or spreading misinformation. These baselines could define thresholds for a particular application (i.e., metrics that will indicate where Gen-AI has a clear benefit for a particular sector). This could leverage notions from planetary and social carrying capacity ([Steffen et al. 2015](#)). Crossing these boundaries or thresholds increases the risk of generating large-scale abrupt or irreversible changes. Scenarios, however, would define a set of explanatory variables that define several key variants for a Gen-AI application (for example: hardware, model type, task, etc.).

To date, the assessment of benefits within the information communication and technology industry broadly has focused on extrapolation from individual case study baselines, such as telework or the use of smart metering ([Ligozat et al. 2022](#); [Rasoldier et al. 2022](#)). Challenges emerge in the impossibility of extrapolating from case

studies, as overly optimistic extrapolation factors are often used and negative effects are incompletely accounted for (including the potential for induced demand or rebound) ([Roussilhe, Ligozat, and Quinton 2023](#)). As these authors state, improvements to this approach for Gen-AI would limit extrapolation, provide transparency in assumptions, and perform case studies according to more random sampling ([Roussilhe, Ligozat, and Quinton 2023](#)). In addition, lessons could be drawn from carbon market validation frameworks, acknowledging the limitation of assigning monetary value to many of the broader system-level implications discussed throughout this document ([Knox-Hayes 2016](#)). The learnings and pitfalls arising from standards that have been developed for verification and validation within voluntary carbon markets provide a precedent for determining the value of interventions ([MacKenzie 2009](#)). To demonstrate any potential carbon reduction, projects offered within voluntary frameworks must be additional (projects must reduce emissions that would not otherwise be cut), verifiable, immediate (as emissions happen today, projects that occur imminently are more valuable), and durable (i.e., permanent—CO₂ emissions stay in the atmosphere for a century or more, requiring the offset of an equivalent amount of emissions for at least that long).

2.3 Data and Inventory

The goal and scoping exercise identifies the materials and processes that will be considered, which defines the inflows and outflows that must be quantified. This inventory step is an accounting exercise that includes primary data collection by life cycle stage or process step but relies on background datasets assembled by a variety of practitioners with varying regional, temporal, and technology relevance. Given the unprecedented pace of Gen-AI development, the data that support this accounting must align with findable, accessible, interoperable, and reuseable (FAIR) data principles, and the methods to determine impacts must be transparent and available ([Wilkinson et al. 2016](#)). These measures can enable users to estimate impact across a level playing field to drive traceable and auditable reporting. There is a significant role for industry associations and government agencies to increase the effectiveness of efforts in data collection.

Standards for quantifying organizational emissions have been developed through the Global Reporting Initiative and Carbon Disclosure Project since the 2000s, including accounting for carbon emissions, water footprint, and other metrics ([Matisoff, Noonan, and O'Brien 2013](#)). The impact is tracked as a direct impact (scope 1), purchased electricity (scope 2), and associated upstream and downstream impacts (scope 3). Many have cited the challenges with these reporting initiatives, but there are relevant details to leverage. More recently, the Science Based Targets initiative has outlined sector-specific guidelines for how quickly companies must reduce GHG emissions, with emphasis in the power and industrial sectors, along with significant guidance around land use emissions ([Science Based Targets 2019](#)). Although a full description of inventory calculation is beyond the scope of our commentary here, we provide a few specific insights into the impact of computing broadly and Gen-AI in particular.

2.3.1 Impact of Gen-AI

The nature of Gen-AI means we must focus attention on primary data collection and inventory specification associated with the use phase (or operational impact). The operational impact of Gen-AI is so extensive and dynamic that we assert a benefit–cost analysis must separate elements within the use phase inventory. As others have outlined, this should individually account for model development and training (tuning model variants through hyperparameter search, for example) as well as inference (at lower energy cost but with the most frequent occurrence). The difficulty in accounting for overall usage patterns cannot be overstated, as it depends on knowledge of "model type and size, hardware and resource utilization, training requirements, and frequency" of each disaggregated step ([Ligozat et al. 2022](#); [Luccioni, Viguiet, and Ligozat 2022](#)). Even the burden of data generation and storage itself should be considered within the scope of an analysis. Foundation models may be at least a factor of two more energy consumptive than more task-specific models, although standardizing comparisons is challenging ([Bommasani et al. 2021](#)). Another opportunity is to increase the use of relevant priors that might focus model development more efficiently but could increase multimodal algorithm needs (generative and multimodal tasks may consume ten times more energy per inference) ([Luccioni, Jernite, and Strubell 2023](#)). Beyond the use phase, maintaining relevant hardware performance may increase upstream burdens in manufacturing as devices become more specialized (resulting in slower manufacturing learning rates) or hardware experience more dynamic use patterns (resulting in shorter device lifetimes) ([Bell 2023](#)). Specific attention should be paid to the impact of Gen-AI on hardware specialization, device lifetime, and operating system integration, among other factors. The data associated with applications would initially include that derived from specific case studies for quality of search result or prediction, for example, but would broaden into metrics that include individuals and communities impacted, and e.g., how search behavior changes over time.

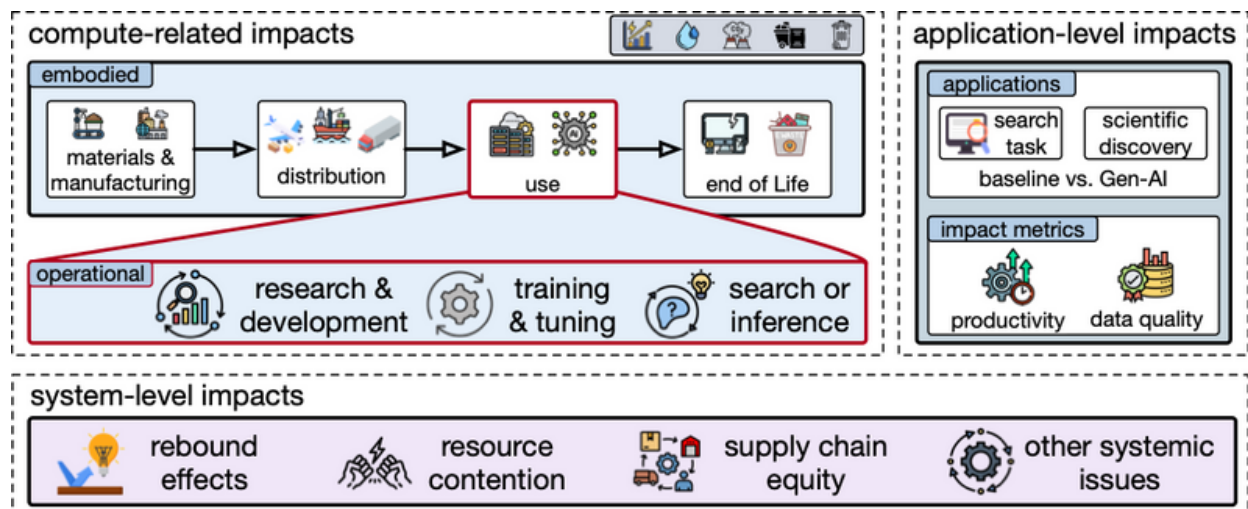


Figure 1

A preliminary example of the benefit-cost evaluation for the scientific discovery and search application.

2.4 Initial Framework Applications

We offer two brief examples of applying this framework to offer insights into what a benefit-cost analysis would contain for a specific application in sustainability, scientific discovery, and a broader application, search. Figure 1 shows a high-level overview of the proposed benefit-cost evaluation framework for the scientific discovery and search application.

One application of Gen-AI in the sustainability domain is on discovery related to chemicals, materials, and processes ([Bran et al. 2023](#); [Bran and Schwaller 2023](#)). For example, innovations are critical to advancement in chemicals for catalysis, more recyclable or biodegradable materials ([Luu et al. 2024](#)), materials for improved battery storage, or increasing the durability of metal alloys. Within scientific discovery, generative models and large language models have been used to predict the stability or reactivity of chemical compounds, discover new materials, identify pathways to synthesize these materials, and structure information contained within text, tables, or images ([Jablonka et al. 2023](#)). There are a few key elements of this application that can be leveraged conceptually to adopt the elements outlined at the beginning of this section to determine benefits (boundaries, baselines, and thresholds). The materials discovery process can be bounded in terms of life cycle stages from design to synthesis to manufacture and use. Also, the ways in which Gen-AI accelerates the determination of process pathways, reaction yields, synthesis byproducts, device performance, and manufacturing process, for example, can be incorporated into scenario and baseline development ([Subramanian et al. 2024](#)). Costs associated with increased consumption, emissions-intensive technology lock-in and model misuse should be a part of these scenarios. Second, typically, Gen-AI advances within materials discovery correspond to specific workflows for incorporating AI-enabled prediction into the scientific inquiry process, so eventually, benchmarks could be established along each step of this workflow ([MacLeod et al. 2022](#)). And finally, the application space for a particular material can be used to place a threshold on global sustainability relevance.

In analyzing the impact of using an ultra, next-generation GPT model called SearchAId for search tasks, as shown in Figure 1, we consider the unit of analysis to be the returned search query results from executing a single search. This involves comparing a baseline scenario, in which a user performs a standard Google search on CPU-based servers, and a Gen-AI scenario, in which the user prompts SearchAId, running on a GPU cluster. The evaluation covers compute-related costs across three phases: inference, development and training, and supply chain and end-of-life. In the inference phase, the focus is on quantifying the energy, carbon, and water costs of executing a search query on the required computing resources. The development and training phase accounts for the environmental impacts and computing resources necessary for training, tuning, or indexing, with costs amortized over the model's or search engine's lifecycle. The supply chain and end-of-life phase consider the resources and emissions of manufacturing the required hardware and building relevant infrastructure, such as buildings.

To assess application-related impacts and benefits, two approaches can be adopted: either fixing the number of queries or prompts to compare response quality in both scenarios, or quantifying the number of queries or

prompts needed to achieve a specific information quality or response. This choice hinges on the application's needs and the tools available for evaluating search task quality within the domain. Baselines would incorporate the time, quality, and extent of the search-query result, which could be assessed through expert input, measures of similarity, or a series of benchmark tests of what steps users take subsequent to receiving search results. Scenarios could look across Gen-AI model types, including ChatGPT 3.5, ChatGPT 4 ([Brown et al. 2020](#)), and Bard search ([Manyika and Sissie 2023](#)). Moreover, the analysis extends to system-level impacts, including the socioeconomic effects of Gen-AI applications across various domains, such as job loss, competition for electricity, AI hardware acquisition, and land use implications.

3. Stakeholder Engagement for Responsible Development of Gen-AI

The unfettered growth of Gen-AI poses significant economic, social, and environmental challenges. Advocating for an outright halt to Gen-AI's development is impractical ([Future of Life Institute 2023](#)). As the previous section outlined, even conducting a comprehensive benefit–cost analysis is far from straightforward. Although a perfect benefit–cost evaluation framework for Gen-AI may never emerge, we can work iteratively toward impact evaluation criteria that integrate knowledge from many different perspectives through active stakeholder engagement. Gen-AI's sustainability implications' complex and interdependent nature necessitates a collaborative, multistakeholder approach. Development strategies that are shaped by integrating many different stakeholder perspectives will be more robust to both sector-specific pitfalls and challenges at the interfaces between sectors. Implementing Gen-AI development strategies in a sustainability-cognizant and coordinated manner will foster adoption and help move toward relational footprinting that is geographically grounded and where regional differences across global geographies are properly accounted. Here, we identify action items for various stakeholders to build our proposed benefit–cost evaluation capability.

3.1 The Role of Leadership Teams

Leadership teams, encompassing CEOs, executive teams, and board members at organizations steering Gen-AI technologies, play a pivotal role in shaping the future trajectory of Gen-AI development. Their strategic decisions extend beyond the technological and operational domains, touching upon the ethical, social, and environmental implications of Gen-AI. A commitment to responsible AI practices by these leadership figures is essential for embedding ethical considerations into the DNA of Gen-AI initiatives. Through their governance, leadership teams can play a role in spearheading the creation of industry standards and best practices that prioritize sustainability, equity, and transparency. Moreover, by actively engaging in dialogues with policymakers, academia, and civil society, they can ensure that the evolution of Gen-AI technologies is aligned with societal values and global sustainability goals. While not precluding the need for decisive governmental action and regulation, this leadership is crucial for developing Gen-AI in a way that is responsible, beneficial to society, and mindful of environmental impacts.

3.2 The Role of Policy Makers and Legal Experts

Several factors, including lack of access to industry data, bottom-up versus top-down assessments, system boundaries, geographic averaging, functional units, and differential energy efficiency assumptions, make precise estimation of Gen-AI costs complex or impossible. For instance, even datacenters' historical energy consumption estimates vary widely ([Andrae 2019](#); [Andrae and Edler 2015](#); [Belkhir and Elmeligi 2018](#); [Ferrebœuf 2019](#); [Hintemann 2018b](#); [Masanet et al. 2020b](#); [Andrae 2020](#)); with estimates for 2018 lying in the range of 200 TWh ([Masanet et al. 2020b](#)) to more than 900 TWh ([Hintemann 2018b](#)). A set of policies that ensure transparency across the aforementioned factors would significantly aid a precise estimation of costs, enabling the development of a functional benefit–cost evaluation framework. Furthermore, even if the measurement and accounting issues are resolved, creating effective incentives for environmental and social sustainability alongside efficiency remains challenging. Poorly designed incentives can lead to unintended behaviors or even exacerbate the problem they were intended to ameliorate ([Hallegatte and Engle 2019](#)). Additionally, in a growth-driven economy, social issues, such as data privacy and job automation, are generally sidelined for economic benefits. The lack of holistic metrics skews our understanding of Gen-AI's impact, highlighting the need for policy-driven approaches that include environmental and social metrics in Gen-AI assessments. They must set clear reporting standards and foster transparent, accountable AI practices. The ethical issues surrounding Gen-AI require policymakers to work with legal experts, human rights activists, and civil society members. Legal experts can help Gen-AI practitioners ensure compliance with legislation regarding ethical boundaries and improve risk resiliency by anticipating future regulations.

3.3 The Role of AI Practitioners and Engineers

Those who develop and apply Gen-AI technologies are well positioned to describe the capabilities of the tools and predict forthcoming innovations to other stakeholders. Targeted outreach and collaborative initiatives can lower barriers that inhibit the language of technology from translating into, for example, the language of policy ([Krafft et al. 2020](#)). Developing a comprehensive impact assessment framework and standardized sustainability metrics requires insight from computer scientists to properly account for technical considerations such as the spatiotemporal variations in the environmental costs of a distributed cloud environment. AI practitioners already contribute to the discourse around ethical and sustainable AI regulations ([Schiff et al. 2020](#)), which is crucial especially given the outsized impact of leadership at hubs of AI innovation.

In addition, AI practitioners and application domain architects can tip the benefit–cost trade-off for various Gen-AI applications by reducing the associated costs or enhancing the realized benefits. AI practitioners and engineers should keep making strides toward greater algorithmic efficiency with techniques including transfer learning, fine-tuning, and mixed-expert models. Engineers can compound algorithmic energy efficiency gains by continuing to improve the hardware's efficiency, especially during the operational life cycle stage, and by optimizing datacenters' resource and power usage effectiveness. Furthermore, computer application developers and data structure developers can help maximize the benefits of Gen-AI, e.g., by efficiently and rapidly

processing data to ensure the reliability and resilience of complex systems ([Ilic 2016](#)) and by coordinating Internet of Things (IoT) devices for demand response ([Ilic and Jaddivada 2019](#)).

3.4 The Role of Energy and Supply Chain Sectors

Addressing Gen-AI's impact involves managing supply chain accuracy and evaluating electric grid factors, such as clean energy use, benefits of demand response programs, and grid cleanliness. Proper baselines are necessary for comparison. Reviewing policies for their impact on long-term, nonfinancial, or secondary goals is important, as it may offer opportunities to solve multiple problems simultaneously. For instance, expanding the computing infrastructure can be done alongside addressing urgent challenges such as strategically improving the power infrastructure ([Kirchhoff et al. 2016](#)).

3.5 The Role of Economists

Continual capital growth is crucial for a healthy economy, yet historically, periods of rapid expansion in technologies, such as “dot com” and “crypto” have led to boom-and-bust cycles. Economists will play a key role in identifying the lessons from the prior unsustained booms and analyzing the warning signs of rapid growth. Because the mechanisms to ameliorate negative consequences in the labor market and supply chains depend on whether AI growth causes a macroeconomic disruption, a series of sector-specific disturbances, or some combination of the two ([Furman and Seamans 2019](#)), it is crucial for economists to monitor the impact of AI on economies. They can identify key indicators of economic bubbles versus sustainable capital expansion. Gen-AI has an enormous potential to change the labor market ([Chui et al. 2023](#)), and economists must study the consequences of unchecked expansion of Gen-AI deployments. These studies will also help illuminate how market speculation alone cannot sustain long-term economic health.

3.6 The Role of Social Scientists

Social scientists provide insights into the societal benefits and costs of Gen-AI. Quantitative and qualitative trade-off analyses of Gen-AI growth strategies would help illuminate a broad spectrum of potential social impacts, from user experience and societal acceptance to broader ethical considerations. This work would ensure that the evaluation framework is sensitive to the sociocultural milieu in which Gen-AI operates.

3.7 The Role of Civil Society

Civil society plays a crucial role in the impact assessment space. NGOs have shown they are well positioned to collect and synthesize diverse perspectives on AI ([Schiff et al. 2020](#)) and serve as a crucial link between those developing AI tools and communities that are the first to experience economic and social consequences of AI's growth. In conducting third-party evaluations of the potential benefits and costs of AI applications, advocacy groups and think tanks can elevate the priorities of those stakeholders who are affected by developments in the technology field but are not in a position to shape the priorities and progress directly. Because much of the existing literature on AI ethics and policies comes from wealthy countries, the field is vulnerable to

disproportionately catering to the needs and economies of wealthy countries as opposed to the priorities and challenges unique to the Global South ([Schiff et al. 2020](#)). As governmental organizations invest in AI and develop AI regulations ([Ulnicane 2022](#)), civil society can leverage existing organizational structures to convey reactions and recommendations.

4. Conclusion

In the rapidly advancing field of Gen-AI, the practical adoption of a structured and comprehensive evaluation framework is essential for fostering responsible growth. The proposed benefit–cost evaluation framework introduces a specific evaluative approach, emphasizing the need for immediate implementation ([Pasek, Vaughan, and Starosielski 2023](#)). Practitioners must be sensitive to the turbulence and inherent limitations in implementing an actively evolving framework. In concluding our proposal, we explore various ways such a framework can ensure responsible and sustainable development in the Gen-AI sector. Gen-AI’s rapid growth and substantial energy consumption suggest its emergence as a distinct sector, warranting dedicated monitoring and analysis ([de Vries 2023](#)). At present, the understanding of both the direct and indirect impacts of digitalization on energy use, carbon emissions, and potential mitigation is limited ([IPCC 2022](#)), directly influencing our understanding of Gen-AI’s implications.

To facilitate such an understanding, the framework can adopt methods used in traditional sector analysis, such as tracking economic indicators, regulatory changes, environmental impacts, and technological developments ([Parmesan, Morecroft, and Trisurat 2022](#)). By incorporating these aspects, the framework would facilitate a comprehensive understanding of Gen-AI’s evolution. The collected data aids in developing integrated assessment models and emissions projections, enabling a more accurate description of Gen-AI’s impact to aid policy and strategic decisions.

A crucial aspect of the framework could be in identifying and maximizing the benefit–cost ratio of Gen-AI initiatives. Policies governing Gen-AI should be rooted in scientific evidence and sustainable growth strategies rather than being driven solely by economic ambitions. Research in climate mitigation and adaptation enumerates the reasons that incremental changes are easier to implement than transformational changes, including the actual or perceived cost and the requisite individual or institutional behavior change ([Kates, Travis, and Wilbanks 2012](#)). The more that Gen-AI can grow both using and encouraging responsible and sustainable practices up front, the less likely a costly and difficult transformational change will be needed to address an entrenched and problematic vulnerability down the line. Therefore, the framework seeks to identify opportunities where incremental advancements can yield substantial benefits, both economically and environmentally. The framework also facilitates exploring the concept of eco-economic decoupling, emphasizing the importance of balancing technological advancement with environmental sustainability. Drawing inspiration from Bayo Akomolafe’s advocacy for “slow urgency” ([Akomolafe 2023](#)), it suggests a more measured approach to Gen-AI development. Although recognizing the value of Gen-AI products, the framework cautions against the frenetic pace of development akin to a Klondike gold rush. Instead, it promotes

a more deliberate, thoughtful approach that considers long-term environmental impacts and societal needs. This approach encourages stakeholders to reevaluate the notion of growth, advocating for a model that aligns technological progress with environmental sustainability and social well-being. By doing so, the framework aims to ensure that Gen-AI contributes positively to society without exacerbating environmental challenges.

Acknowledgments

The authors would like to acknowledge the thought-provoking conversations and support from Aneil Tripathy, Evan Coleman, and Laura Frye-Levine. We would also acknowledge Anantha Chandrakasan for the inspiration and framing to develop this contribution.

Bibliography

- Acun, Bilge, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Maeng, Udit Gupta, Manoj Chakkaravarthy, David Brooks, and Carole-Jean Wu. 2023. “Carbon explorer: A holistic framework for designing carbon aware datacenters.” In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 118–32. ASPLOS 2023. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3575693.3575754>.
- Akomolafe, Bayo. 2023. “A Slower Urgency.” <https://www.bayoakomolafe.net/post/a-slower-urgency>.
- Andrae, Anders SG. 2020. “New perspectives on internet electricity use in 2030.” *Engineering and Applied Science Letter* 3, no. 2 (2020): 19–31.
- Andrae, Anders SG. 2019. “Projecting the chiaroscuro of the electricity use of communication and computing from 2018 to 2030.” Preprint, submitted February 2019. DOI:[10.13140/RG.2.2.25103.02724](https://doi.org/10.13140/RG.2.2.25103.02724).
- Andrae, Anders S G, and Tomas Edler. 2015. “On global electricity usage of communication technology: trends to 2030.” *Challenges* 6, no.1 (2015): 117–157.
- Andrae, Anders S.G., and Otto Andersen. 2010. “Life cycle assessments of consumer electronics— are they consistent?” *International Journal of Life Cycle Assessment* 15 (2010): 827–836. <https://doi.org/10.1007/s11367-010-0206-1>.
- Artificial Intelligence Act, 2023. Committee on the Internal Market and Consumer Protection; Committee on Civil Liberties, Justice and Home Affairs. European Parliament. <https://www.europarl.europa.eu/committees/en/indexsearch?query=Artificial+Intelligence+Act%2C+2023>.
- Bashir, Noman, Yasra Chandio, David E Irwin, Fatima M. Anwar, Jeremy Gummeson, and Prashant J Shenoy. 2023. “Jointly Managing Electrical and Thermal Energy in Solar- and Battery-Powered Systems.” In *Proceedings of the 14th ACM International Conference on Future Energy Systems*, 132–43. ACM. 2023. <https://doi.org/10.1145/3575813.3595191>.

Bashir, Noman, Tian Guo, Mohammad Hajiesmaili, David Irwin, Prashant Shenoy, Ramesh Sitaraman, Abel Souza, and Adam Wierman. 2021. “Enabling sustainable clouds: The case for virtualizing the energy system.” In *Proceedings of the ACM Symposium on Cloud Computing*, 2021. <https://doi.org/10.1145/3472883.3487009>.

Becker, Gerrit, Luca Bennici, Anamika Bhargava, Andrea Del Miglio, Jeffrey Lewis, and Pankaj Sachdeva. 2022. “The green IT revolution: A blueprint for CIOs to combat climate change.” McKinsey Technology (Sept. 15, 2022). <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-green-it-revolution-a-blueprint-for-cios-to-combat-climate-change>.

Kari Beets and Michael Hartnett. 2024. North America Data Center Report | H2 2023. Technical Report. JLL Research. <https://www.us.jll.com/en/trends-and-insights/research/na-data-center-report>.

Belkhir, Lotfi, and Ahmed Elmeligi. 2018. “Assessing ICT global emissions footprint: Trends to 2040 & recommendations.” *Journal of Cleaner Production* 177 (2018): 448–463.

Bell, Allison. 2023. “Bending the ICT curve: Evaluating options to achieve 2030 sector-wide climate goals and projecting new technology impacts.” Thesis. Massachusetts Institute of Technology, 2023.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. “On the dangers of stochastic parrots: Can language models be too big?” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>.

Berton, Pierre. 2011. *Klondike: The Last Great Gold Rush, 1896-1899*. Woodbridge, Canada: Anchor Canada, 2011.

Birhane, Abeba, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. “The values encoded in machine learning research.” In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 173–184.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 2021. “On the opportunities and risks of foundation models.” Preprint, submitted August 16, 2021. <http://arxiv.org/abs/2108.07258>.

Börjesson Rivera, Miriam, Cecilia Håkansson, Åsa Svenfelt, and Göran Finnveden. 2014. “Including second order effects in environmental assessments of ICT.” *Environmental Modelling & Software* 56 (2014): 105–115. <https://doi.org/10.1016/j.envsoft.2014.02.005>.

Bran, Andres M, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. “ChemCrow: Augmenting large-language models with chemistry tools.” Preprint, submitted April 11, 2023. <http://arxiv.org/abs/2304.05376>.

Bran, Andres M, and Philippe Schwaller. 2023. “Transformers and large language models for chemistry and drug discovery.” Preprint, submitted October 9, 2023. <http://arxiv.org/abs/2310.06083>.

Chakrabarty, Abel Souza, Shruti Jasoria, Basundhara Chakrabarty, Alexander Bridgwater, Axel Lundberg, Filip Skogh, Ahmed Ali-Eldin, David Irwin, and Prashant Shenoy. 2023. “CASPER: carbon-aware scheduling and provisioning for distributed web services.”

Chips and Science Act, H.R.4346, 117th Congress (2021).

Chui, Michael, Lareina Yee, Bryce Hall, and Alex Singla. 2023. “The State of AI in 2023: Generative AI’s Breakout Year.” <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year>

Cihon, Peter, Matthijs M Maas, and Luke Kemp. 2020. “Fragmentation and the future: investigating architectures for international AI governance.” *Global Policy* 11, no 5 (2020): 545–556.

Ciroth, Andreas, and Rickard Arvidsson, eds. 2021. *Life Cycle Inventory Analysis. LCA Compendium – The Complete World of Life Cycle Assessment*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-62270-1>.

Clarke, Harrison. 2023. “Big Data’s Impact: Optimizing AI with Vast Datasets.” <https://www.harrisonclarke.com/blog/big-datas-impact-optimizing-ai-with-vast-datasets>.

Clément, Louis-Philippe P.-V.P., Quentin ES Jacquemotte, and Lorenz M Hilty. 2020. “Sources of variation in life cycle assessments of smartphones and tablet computers.” *Environmental Impact Assessment Review* 84 (2020): 106416. <https://doi.org/10.1016/j.eiar.2020.106416>.

Coeckelbergh, Mark. 2021. “AI for climate: freedom, justice, and other ethical and political challenges.” *AI and Ethics* 1, no. 1(2021): 67–72.

Council, E U. 2023. “Artificial intelligence act: council and parliament strike a deal on the first rules for AI in the world.” December 9, 2023. <https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/>.

Dally, Bill. 2023. “Wide Horizons: NVIDIA Keynote Points Way to Further AI Advances.” <https://blogs.nvidia.com/blog/hot-chips-dally-research/>.

Das, Sujit, and Elizabeth Mao. 2020. “The global energy footprint of information and communication technology electronics in connected internet-of-things devices.” *Sustainable Energy, Grids and Networks*, 24 (2020) 100408. <https://doi.org/10.1016/j.segan.2020.100408>.

- Dauvergne, Peter. 2022. “Is artificial intelligence greening global supply chains? exposing the political economy of environmental costs.” *Review of International Political Economy* 29, no.3 (2022): 696–718.
- de Vries, Alex. 2023. “The growing energy footprint of artificial intelligence.” *Joule* 7, no. 10 (2023): 2191–2194.
- Dean, Jeffrey, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc aurelio Ranzato, et al. 2012. “Large Scale Distributed Deep Networks.” *Advances in Neural Information Processing Systems* 25 (2012). https://proceedings.neurips.cc/paper_files/paper/2012/file/6aca97005c68f1206823815f66102863-Paper.pdf.
- Debnath, Ramit, Felix Creutzig, Benjamin K. Sovacool, and Emily Shuckburgh. 2023. “Harnessing human and machine intelligence for planetary-level climate action.” *NPJ Climate Action* 2, no 1 (2023): 20. <https://doi.org/10.1038/s44168-023-00056-3>.
- Debus, Charlotte, Marie Piraud, Achim Streit, Fabian Theis, and Markus Götz. 2023. “Reporting electricity consumption is essential for sustainable AI.” *Nature Machine Intelligence* 5, no. 11 (2023): 1176–1178. <https://doi.org/10.1038/s42256-023-00750-1>.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. “Bert: Pre-training of deep bidirectional transformers for language understanding.” Preprint, submitted October 11, 2018. <http://arxiv.org/abs/1810.04805>.
- Du, Nan, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, et al. 2021. “Glam: Efficient scaling of language models with mixture-of-experts.” In *International Conference on Machine Learning*, 5547-5569. PMLR, 2022
- Entefy. 2023. “2 AI Winters and 1 Hot AI Summer.” <https://www.entefy.com/blog/2-ai-winters-and-1-hot-ai-summer/>.
- Executive Order 14030 of May 20, 2021. Executive Order on Climate-Related Financial Risk.
- Fauré, Eléonore, Yevgeniya Arushanyan, Elisabeth Ekener, Sofiia Miliutenko, and Göran Finnveden. 2017. “Methods for assessing future scenarios from a sustainability perspective.” *European Journal of Futures Research* 5, no 1 (2017): 1-20. <https://doi.org/10.1007/s40309-017-0121-9>.
- Ferreboeuf, Hugues. 2019. “LEAN ICT- Towards Digital Sobriety.” The Shift Project. https://theshiftproject.org/wp-content/uploads/2019/03/Lean-ICT-Report_The-Shift-Project_2019.pdf.
- Finkbeiner, Matthias, Atsushi Inaba, Reginald B.H. Tan, Kim Christiansen, and Hans Jürgen Klüppel. 2006. “The new international standards for life cycle assessment: ISO 14040 and ISO 14044.” *The International*

Journal of Life Cycle Assessment 11(2006): 80-85. <https://doi.org/10.1065/lca2006.02.002>.

Fung, Victor, Jiaxin Zhang, Eric Juarez, and Bobby G. Sumpter. 2021. “Benchmarking graph neural networks for materials chemistry.” *NPJ Computational Materials* 7, no. 1 (2021): 84. <https://doi.org/10.1038/s41524-021-00554-0>.

Furberg, Anna, Rickard Arvidsson, and Sverker Molander. 2022. “A Practice-based framework for defining functional units in comparative life cycle assessments of materials.” *Journal of Industrial Ecology* 26, no. 3 (2022): 718–730. <https://doi.org/10.1111/jiec.13218>.

Furman, Jason, and Robert Seamans. 2019. “AI and the Economy.” *Innovation Policy and the Economy* 19, no. 1 (2019): 161–191.

Future of Life Institute. 2023. “Pause giant AI experiments: an open letter.” March 22, 2023. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.

Giampietro, Mario, and Kozo Mayumi. 2018. “Unraveling the complexity of the jevons paradox: the link between innovation, efficiency, and sustainability.” *Frontiers in Energy Research* 6 (2018): 26.

glass.ai. 2023. “Code Red: The AI Armies Of The Tech Giants.” *Medium*. March 27, 2023. <https://glassai.medium.com/code-red-the-ai-armies-of-the-tech-giants-cc8594982adb>

The Global Economy. 2023. Electricity production capacity - Country rankings. https://www.theglobaleconomy.com/rankings/electricity_production_capacity/.

Hallegatte, Stephane, and Nathan L Engle. 2019. “The search for the perfect indicator: Reflections on monitoring and evaluation of resilience for improved climate risk management.” *Climate Risk Management* 23 (2019): 1–6.

Hanafy, Walid A, Qianlin Liang, Noman Bashir, David E Irwin, and Prashant J Shenoy. 2023. “CarbonScaler: Leveraging Cloud Workload Elasticity for Optimizing-Efficiency.” *Proc. ACM Meas. Anal. Comput. Syst.* 7, 3, Article 57 (December 2023), 28 pages. <https://doi.org/10.1145/3626788>.

Hintemann, Ralph. 2018a. *Efficiency Gains Are Not Enough: Data Center Energy Consumption Continues to Rise Significantly*. Borderstep Inst. für Innovation und Nachhaltigkeit gGmbH.

———. 2018b. *Efficiency Gains Are Not Enough: Data Center Energy Consumption Continues to Rise Significantly*. Borderstep Inst. für Innovation und Nachhaltigkeit gGmbH.

Ilic, M D. 2016. “Toward a unified modeling and control for sustainable and resilient electric energy systems.” *Foundations and Trends® in Electric Energy Systems* 1, no. 1 (2016): 1–141.

Ilic, Marija, and Rupamathi Jaddivada. 2019. “Introducing dymonds-as-a-service (dymaas) for internet of things.” In *Proceedings of IEEE High Performance Extreme Computing Conference (HPEC)*, 1–9. IEEE, 2019.

Ingwersen, Wesley W., and Vairavan Subramanian. 2014. “Guidance for product category rule development: process, outcome, and next steps.” *International Journal of Life Cycle Assessment* 19 (2024): 532–537.
<https://doi.org/10.1007/s11367-013-0659-0>.

Ioffe, Sergey, and Christian Szegedy. 2015. “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” In *Proceedings of the 32nd International Conference on Machine Learning*, 448–456. PMLR, 2015. <https://proceedings.mlr.press/v37/ioffe15.html>.

IPCC. 2022. “Climate change 2022: Mitigation of climate change.” *Contribution of working group III to the sixth assessment report of the Intergovernmental Panel on Climate Change* (2022).
<https://www.ipcc.ch/report/ar6/wg3/>.

Ipsen, Kikki Lambrecht, Regitze Kjær Zimmermann, Per Sieverts Nielsen, and Morten Birkved. 2019. “Environmental assessment of Smart City Solutions using a coupled urban metabolism—life cycle impact assessment approach.” *The International Journal of Life Cycle Assessment* 24 (2019): 1239–1253.

Itten, René, Roland Hischier, Anders S G Andrae, Jan C T Bieser, Livia Cabernard, Annemarie Falke, Hugues Ferreboeuf, et al. 2020. “Digital transformation-life cycle assessment of digital services, multifunctional devices and cloud computing.” *The International Journal of Life Cycle Assessment* 25 (2020): 2093-2098.
<https://doi.org/10.1007/s11367-020-01801-0>.

ITU. 2023. “Population of global offline continues steady decline to 2.6 billion people in 2023.” Press release, September 12, 2023. International Telecommunication Union (ITU).
<https://www.itu.int/en/mediacentre/Pages/PR-2023-09-12-universal-and-meaningful-connectivity-by-2030.aspx>.

Jablonka, Kevin Maik, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D. Bocarsly, Andres M. Bran, Stefan Bringuier, et al. 2023. “14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon.” *Digital Discovery* 2, no. 5 (2023): 1233–1250.
<https://doi.org/10.1039/d3dd00113j>.

Jelinek, Thorsten, Wendell Wallach, and Danil Kerimi. 2021. “Policy brief: the creation of a G20 coordinating committee for the governance of artificial intelligence.” *AI and Ethics* 1, no. 2 (2021): 141–50.

Jouppi, Norman P, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, et al. 2017. “In-datacenter performance analysis of a tensor processing unit.” In *Proceedings of The 44th Annual International Symposium On Computer Architecture*, 1–12. 2017.
<https://doi.org/10.1145/3140659.3080246>.

- 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, no. 6 (2022): 518–527.
- Kak, Amba, and Sarah Myers West. 2023. “2023 landscape: confronting tech power.” *AI Now*. <https://ainowinstitute.org/wp-content/uploads/2023/04/AI-Now-2023-Landscape-Report-FINAL.pdf>.
- Katal, Avita, Susheela Dahiya, and Tanupriya Choudhury. 2023. “Energy efficiency in cloud computing data centers: a survey on software technologies.” *Cluster Computing* 26, no. 3 (2023): 1845–1875.
- Kates, Robert W, William R Travis, and Thomas J Wilbanks. 2012. “Transformational adaptation when incremental adaptations to climate change are insufficient.” In *Proceedings of the National Academy of Sciences* 109, no. 19 (2012): 7156–7161.
- Kingma, Diederik P, and Jimmy Ba. 2014. “Adam: A method for stochastic optimization.” Preprint, submitted December 22, 2014. <https://arxiv.org/abs/1412.6980>.
- Kirchhoff, Hannes, Noara Kebir, Kirsten Neumann, Peter W Heller, and Kai Strunz. 2016. “Developing mutual success factors and their application to swarm electrification: microgrids with 100% renewable energies in the Global South and Germany.” *Journal of Cleaner Production* 128 (2016): 190–200.
- Klaaßen, Lena, and Christian Stoll. 2021. “Harmonizing corporate carbon footprints.” *Nature Communications* 12, no. 1 (2021): 1–13.
- Knight, Will. 2023. “Google’s Gemini is the real start of the generative AI boom.” *Wired*, December 7, 2023.
- Knox-Hayes, Janelle. 2016. *The Cultures of Markets: The Political Economy of Climate Governance*. 2016: Oxford, Oxford University Press.
- Krafft, P M, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. 2020. “Defining AI in policy versus practice.” In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 72–78. 2020.
- Kremer, Andreas, Angela Luget, Daniel Mikkelsen, Henning Soller, Malin Strandell-Jansson, and Sheila Zingg. 2023. “As gen AI advances, regulators – and risk functions – rush to keep pace.” McKinsey’s Risk & Company. <https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/as-gen-ai-advances-regulators-and-risk-functions-rush-to-keep-pace>.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. “Imagenet classification with deep convolutional neural networks.” *Advances in Neural Information Processing Systems* 25 (2012). https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Krogh, Bruce, Marija Ilic, and S Sastry. 2007. “National workshop on beyond SCADA: networked embedded control for cyber-physical systems (NEC4CPS): research strategies and roadmap.” *Technical Report, Team for Research in Ubiquitous Secure Technology (TRUST)* (2007).

Lechowicz, Adam, Nicolas Christianson, Jinhang Zuo, Noman Bashir, Mohammad Hajiesmaili, Adam Wierman, and Prashant Shenoy. 2024. “The online pause and resume problem: optimal algorithms and an application to carbon-aware load shifting.” *Proc. ACM Meas. Anal. Comput. Syst.* 7, 3, Article 45 (December 2023), 32 pages. <https://doi.org/10.1145/3626776>.

Li, Baolin, Siddharth Samsi, Vijay Gadepally, and Devesh Tiwari. 2023. “Clover: Toward Sustainable AI with Carbon-Aware Machine Learning Inference Service.” In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM. <https://doi.org/10.1145/3581784.3607034>.

Ligozat, Anne-Laure, Julien Lefevre, Aurélie Bugeau, and Jacques Combaz. 2022. “Unraveling the hidden environmental impacts of AI solutions for environment life cycle assessment of AI solutions.” *Sustainability* 14, no. 9 (2022): 5172. <https://doi.org/10.3390/su14095172>.

Luccioni, Alexandra Sasha, Yacine Jernite, and Emma Strubell. 2023. “Power hungry processing: Watts driving the cost of AI deployment?” Preprint, submitted November 28, 2023. <https://arxiv.org/abs/2311.16863>.

Luccioni, Alexandra Sasha, Sylvain Viguiet, and Anne-Laure Ligozat. 2022. “Estimating the carbon footprint of bloom, a 176b parameter language model,” *Journal of Machine Learning Research* 24, no. 253 (2023): 1–15.

Luccioni, Sasha. 2023. “Generative AI Models: History, Costs and Risks.” https://docs.google.com/presentation/d/1FRoyzdodKQ7-5rK--gZFFzK_-kvfhzxJQDYxpnA-6jE/.

Luu, Rachel, Sofia Arevalo, Wei Lu, Bo Ni, Zhenze Yang, Sabrina Shen, Jaime Berkovich, Yu-Chuan Hsu, Stone Zan, and Markus Buehler. 2024. “Learning from Nature to Achieve Material Sustainability: Generative AI for Rigorous Bio-Inspired Materials Design.” An MIT Exploration of Generative AI. <https://mit-genai.pubpub.org/pub/jpkwte2n>.

MacKenzie, D. 2009. “Making things the same: gases, emission rights and the politics of carbon markets.” *Accounting, Organizations and Society* 34, no. 3–4 (2009): 440–55.

MacLeod, Benjamin P., Fraser G.L. Parlane, Amanda K. Brown, Jason E. Hein, and Curtis P. Berlinguette. 2022. “Flexible automation accelerates materials discovery.” *Nature Materials* 21, no. 7 (2022): 772–726. <https://doi.org/10.1038/s41563-021-01156-3>.

Mammen, Priyanka Mary, Noman Bashir, Ramachandra Rao Kolluri, Eun Kung Lee, and Prashant J Shenoy. 2023. “CUFF: A configurable uncertainty-driven forecasting framework for green ai clusters.” In *Proceedings*

of the 14th ACM International Conference on Future Energy Systems, 266–70.

<https://doi.org/10.1145/3575813.3595203>.

Manyika, James, and Hsiao Sissie. 2023. “An overview of Bard: an early experiment with generative AI.” *AI. Google Static Documents 2* (2023). <https://Ai.Google/Static/Documents/Google-about-Bard.Pdf.2023>.

Masanet, Eric, Arman Shehabi, Nuo Lei, Sarah Smith, and Jonathan Koomey. 2020a. “Recalibrating Global Data Center Energy-Use Estimates.” *Science* 367, no. 6481 (2020): 984–986.

———. 2020b. “Recalibrating Global Data Center Energy-Use Estimates.” *Science* 367, no. 6481 (2020): 984–86.

Matisoff, Daniel C., Douglas Noonan, and John J. O’Brien. 2013. “Convergence in environmental reporting: assessing the carbon disclosure project.” *Business Strategy and the Environment* 22, no. 5 (2013): 285–305.

Mytton, David. 2020. “Hiding greenhouse gas emissions in the cloud.” *Nature Climate Change* 10, no. 8 (2020): 701.

Nature editorial authors. 2024. “There are holes in Europe’s AI Act – and researchers can help to fill them.” *Nature* 625 (2024): 216. <https://doi.org/10.1038/d41586-024-00029-4>.

Norem, Josh. 2023. “Nvidia reportedly sold 500,000 H100 AI GPUs in Q3 alone.” *ExtremeTech*, November 28, 2023.

Norris, Gregory A., Jasmina Burek, Elizabeth A. Moore, Randolph E. Kirchain, and Jeremy Gregory. 2021. “Sustainability health initiative for NetPositive enterprise handprint methodological framework.” *International Journal of Life Cycle Assessment* 26 (2021): 528–542. <https://doi.org/10.1007/s11367-021-01874-5>.

OpenAI. 2023. “GPT-4 Technical Report.” <https://cdn.openai.com/papers/gpt-4.pdf>.

Parmesan, Camille, Mike D Morecroft, and Yongyut Trisurat. 2022. “Climate change 2022: impacts, adaptation and vulnerability.” [Research Report] GIEC (2022): hal-03774939. <https://hal.science/hal-03774939/document>.

Pasek, Anne, Hunter Vaughan, and Nicole Starosielski. 2023. “The world wide web of carbon: toward a relational footprinting of information and communications technology’s climate impacts.” *Big Data & Society* 10, no. 1 (2023): 20539517231158990. <https://doi.org/10.1177/20539517231158994>.

Patterson, David, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David R So, Maud Texier, and Jeff Dean. 2022. “The carbon footprint of machine learning training will plateau, then shrink.” *Computer* 55, no. 7 (2022): 18–28. <https://doi.org/10.1109/MC.2022.3148714>.

Perry, Tekla S. 2018. “Move Over, Moore’s Law: Make Way for Huang’s Law.” *IEEE Spectrum*, April 2, 2018.

Pierce, Matthew. 2020. “High-Performance Networking to Support Critical Workloads for AI and ML.” *redapt*.
<https://www.redapt.com/blog/high-performance-networking-to-support-critical-workloads-for-ai-and-ml>.

Radovanović, Ana, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyué Xiao, et al. 2023. “Carbon-aware computing for datacenters.” *IEEE Transactions on Power Systems* 38, no. 2 (2022): 1270–80.

Rao, Lei, Xue Liu, Marija D Ilic, and Jie Liu. 2011. “Distributed Coordination of Internet Data Centers under Multiregional Electricity Markets.” *Proceedings of the IEEE* 100, no. 1 (2011): 269–82.

Rasoldier, Aina, Jacques Combaz, Alain Girault, Kevin Marquet, and Sophie Quinton. 2022. “How realistic are claims about the benefits of using digital technologies for ghg emissions mitigation?” In *LIMITS 2022-Eighth Workshop on Computing within Limits*. 2022. <https://inria.hal.science/hal-03949261/document>

Riahi, Keywan, Detlef P. van Vuuren, Elmar Kriegler, Jae Edmonds, Brian C. O’Neill, Shinichiro Fujimori, Nico Bauer, et al. 2017. “The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: an overview.” *Global Environmental Change* 42 (2017), 153–168.
<https://doi.org/10.1016/j.gloenvcha.2016.05.009>.

Richards, Catherine E., Asaf Tzachor, Shahar Avin, and Richard Fenner. 2023. “Rewards, risks and responsible deployment of artificial intelligence in water systems.” *Nature Water* 1 (2023): 422–432.
<https://doi.org/10.1038/s44221-023-00069-6>.

Rolnick, David, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavlin Ross, et al. 2022. “Tackling climate change with machine learning.” *ACM Computing Surveys (CSUR)* 55, no. 2 (2022): 1–96. <https://doi.org/10.1145/3485128>.

Roussilhe, Gauthier, Anne Laure Ligozat, and Sophie Quinton. 2023. “A long road ahead: a review of the state of knowledge of the environmental effects of digitization.” *Current Opinion in Environmental Sustainability* 62 (2023): 101296. <https://doi.org/10.1016/j.cosust.2023.101296>.

Ruan, Zhenyuan, Malte Schwarzkopf, Marcos K Aguilera, and Adam Belay. 2020. “{AIFM}:{High-Performance},{Application-Integrated} far memory.” In *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*, 315–332. 2020.

S.3732 - 118th Congress (2023-2024): Artificial Intelligence Environmental Impacts Act of 2024. (2024, February 1). <https://www.congress.gov/bill/118th-congress/senate-bill/3732>.

Science Based Targets. 2019. “Foundations of science-based target setting.” Version 1.0, April 2019.

<https://sciencebasedtargets.org/resources/files/foundations-of-SBT-setting.pdf>.

Scherer, Matthew U. 2015. “Regulating artificial intelligence systems: risks, challenges, competencies, and strategies.” *Harv. JL & Tech.* 29 (2015): 353.

Schiff, Daniel, Justin Biddle, Jason Borenstein, and Kelly Laas. 2020. “What’s next for AI ethics, policy, and governance? A global overview.” In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 153–58, 2020.

“The Enhancement and Standardization of Climate-Related Disclosures for Investors.” SECURITIES AND EXCHANGE COMMISSION, 2024. <https://www.sec.gov/files/rules/final/2024/33-11275.pdf>.

Shazeer, Noam, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. “Outrageously large neural networks: the sparsely-gated mixture-of-experts layer.” Preprint, submitted January 23, 2017. <https://arxiv.org/abs/1701.06538>.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. “Dropout: a simple way to prevent neural networks from overfitting.” *The Journal of Machine Learning Research* 15, no. 1 (2014): 1929–1958.

Steffen, Will, Katherine Richardson, Johan Rockström, Sarah E. Cornell, Ingo Fetzer, Elena M. Bennett, Reinette Biggs, et al. 2015. “Planetary boundaries: guiding human development on a changing planet.” *Science* 347, no. 6223 (2015): 1259855. <https://doi.org/10.1126/science.1259855>.

Subramanian, Akshay, Wenhao Gao, Regina Barzilay, Jeffrey C. Grossman, Tommi Jaakkola, Stefanie Jegelka, Mingda Li, et al. 2024 “Closing the Execution Gap in Generative AI for Chemicals and Materials: Freeways or Safeguards.” An MIT Exploration of Generative AI.

Sukprasert, Thanathorn, Abel Souza, Noman Bashir, David Irwin, and Prashant Shenoy. 2024. “On the Limitations of Carbon-Aware Temporal and Spatial Workload Shifting in the Cloud.” Zenodo. <https://doi.org/10.5281/ZENODO.10790855>.

Switzer, Jennifer, Gabriel Marcano, Ryan Kastner, and Pat Pannuto. 2023. “Junkyard computing: repurposing discarded smartphones to minimize carbon.” In *Proceedings of the ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 400–412. <https://doi.org/10.1145/3575693.3575710>.

Thiede, John, Noman Bashir, David Irwin, and Prashant Shenoy. 2023. “Carbon containers: a system-level facility for managing application-level carbon emissions.” In *Proceedings of the ACM 2023 Symposium on Cloud Computing (SoCC '23)*, 17–31. <https://doi.org/10.1145/3620678.362464417-31>.

Ulnicane, Inga. 2022. “Artificial intelligence in the European Union: policy, ethics and regulation.” In *The Routledge Handbook of European Integrations*. Taylor & Francis, 2022.

Uptime Institute. 2024. Five Data Center Predictions for 2024. Technical Report. Uptime Institute.
<https://uptimeinstitute.com/resources/research-and-reports/five-data-center-predictions-for-2024>.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is all you need.” In *Advances in Neural Information Processing Systems*, 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vinitsky, Eugene, Aboudy Kreidieh, Luc Le Flem, Nishant Kheterpal, Kathy Jang, Cathy Wu, Fangyu Wu, Richard Liaw, Eric Liang, and Alexandre M Bayen. 2018. “Benchmarks for reinforcement learning in mixed-autonomy traffic.” In *Conference on Robot Learning*, 399–409. PMLR, 2018. <https://github.com/flow-project/flow>.

Ward, Isabella, and Natalie Lung. 2023. “Big Tech’s Year of Partnering Up With AI Startups.” *Bloomberg*, December 18, 2023.

Weidema, Bo Pedersen. 2003. *Market Information in Life Cycle Assessment*. Vol. 863. København, Denmark: Miljøstyrelsen, 2003.

West, Sarah Myers. 2023. “Competition authorities need to move fast and break up AI.” *Financial Times*, April 17, 2023.

Wiesner, Philipp, Ilja Behnke, Dominik Scheinert, Kordian Gontarska, and Lauritz Thamsen. 2021. “Let’s wait awhile: how temporal workload shifting can reduce carbon emissions in the cloud.” In *Proceedings of the 22nd International Middleware Conference*, 260–272. 2021.

Wijewardane, Nuwan K., Yufeng Ge, Skye Wills, and Terry Loecke. 2016. “Prediction of soil carbon in the conterminous United States: visible and near infrared reflectance spectroscopy analysis of the rapid carbon assessment project.” *Soil Science Society of America Journal* 80, no. 4 (2016): 973–982.
<https://doi.org/10.2136/sssaj2016.02.0052>.








Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, and Philip E Bourne. 2016. “The FAIR Guiding Principles for scientific data management and stewardship.” *Scientific Data* 3, no. 1 (2016): 1–9.

Wright, Dustin, Christian Igel, Gabrielle Samuel, and Raghavendra Selvan. 2023. “Efficiency Is Not Enough: A Critical Perspective of Environmentally Sustainable AI.” Preprint, submitted September 5, 2023.
<https://arxiv.org/abs/2309.02065>.

Wu, Carole-Jean, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, et al. 2022. “Sustainable AI: environmental implications, challenges and opportunities.” *Proceedings of Machine Learning and Systems* 4 (2022): 795–813.

Zhang, Yi, Marija D Ilić, and Ozan K Tonguz. 2010. “Mitigating blackouts via smart relays: a machine learning approach.” *Proceedings of the IEEE* 99, no. 1 (2010): 94–118.

References

- 
- “The Enhancement and Standardization of Climate-Related Disclosures for Investors.” *SECURITIES AND EXCHANGE COMMISSION*, 2024. <https://www.sec.gov/files/rules/final/2024/33-11275.pdf>.
- 
- Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." *arXiv preprint arXiv:1603.04467* (2016).
- 
- Acun, Bilge, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Maeng, Udit Gupta, Manoj Chakkaravarthy, David Brooks, and Carole-Jean Wu. 2023. “Carbon explorer: A holistic framework for designing carbon aware datacenters.” In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 118–32. ASPLOS 2023. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3575693.3575754>.
- 
- Akomolafe, Bayo. 2023. “A Slower Urgency.” <https://www.bayoakomolafe.net/post/a-slower-urgency>.
- 
- Andrae, Anders S G, and Tomas Edler. 2015. “On global electricity usage of communication technology: trends to 2030.” *Challenges* 6, no.1 (2015): 117–157.
- 
- Andrae, Anders S.G., and Otto Andersen. 2010. “Life cycle assessments of consumer electronics—are they consistent?” *International Journal of Life Cycle Assessment* 15 (2010): 827–836. <https://doi.org/10.1007/s11367-010-0206-1>.
- 
- Andrae, Anders SG. 2019. “Projecting the chiaroscuro of the electricity use of communication and computing from 2018 to 2030.” Preprint, submitted February 2019.

↑

- Andrae, Anders SG. 2020. “New perspectives on internet electricity use in 2030.” *Engineering and Applied Science Letter* 3, no. 2 (2020): 19–31.

↑

- Artificial Intelligence Act, 2023. Committee on the Internal Market and Consumer Protection; Committee on Civil Liberties, Justice and Home Affairs. European Parliament.
<https://www.europarl.europa.eu/committees/en/indexsearch?query=Artificial+Intelligence+Act%2C+2023>.

↑

- Bashir, Noman, Tian Guo, Mohammad Hajiesmaili, David Irwin, Prashant Shenoy, Ramesh Sitaraman, Abel Souza, and Adam Wierman. 2021. “Enabling sustainable clouds: The case for virtualizing the energy system.” In *Proceedings of the ACM Symposium on Cloud Computing*, 2021.
<https://doi.org/10.1145/3472883.3487009>.

↑

- Bashir, Noman, Yasra Chandio, David E Irwin, Fatima M. Anwar, Jeremy Gummesson, and Prashant J Shenoy. 2023. “Jointly Managing Electrical and Thermal Energy in Solar- and Battery-Powered Systems.” In *Proceedings of the 14th ACM International Conference on Future Energy Systems*, 132–43. ACM. 2023.
<https://doi.org/10.1145/3575813.3595191>.

↑

- Becker, Gerrit, Luca Bennici, Anamika Bhargava, Andrea Del Miglio, Jeffrey Lewis, and Pankaj Sachdeva. 2022. “The green IT revolution: A blueprint for CIOs to combat climate change.” McKinsey Technology (Sept. 15, 2022). <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-green-it-revolution-a-blueprint-for-cios-to-combat-climate-change>.

↑

- Belkhir, Lotfi, and Ahmed Elmeligi. 2018. “Assessing ICT global emissions footprint: Trends to 2040 & recommendations.” *Journal of Cleaner Production* 177 (2018): 448–463.

↑

- Bell, Allison. 2023. “Bending the ICT curve: Evaluating options to achieve 2030 sector-wide climate goals and projecting new technology impacts.” Thesis. Massachusetts Institute of Technology, 2023.

↑

- Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. “On the dangers of stochastic parrots: Can language models be too big?” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
<https://doi.org/10.1145/3442188.3445922>.

↑

- Berton, Pierre. 2011. *Klondike: The Last Great Gold Rush, 1896-1899*. Woodbridge, Canada: Anchor Canada, 2011.

↑

- Birhane, Abeba, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. “The values encoded in machine learning research.” In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 173–184.

↑

- Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 2021. “On the opportunities and risks of foundation models.” Preprint, submitted August 16, 2021. <http://arxiv.org/abs/2108.07258>.

↑

- Börjesson Rivera, Miriam, Cecilia Håkansson, Åsa Svenfelt, and Göran Finnveden. 2014. “Including second order effects in environmental assessments of ICT.” *Environmental Modelling & Software* 56 (2014): 105–115. <https://doi.org/10.1016/j.envsoft.2014.02.005>.

↑

- Bran, Andres M, and Philippe Schwaller. 2023. “Transformers and large language models for chemistry and drug discovery.” Preprint, submitted October 9, 2023. <http://arxiv.org/abs/2310.06083>.

↑

- Bran, Andres M, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. “ChemCrow: Augmenting large-language models with chemistry tools.” Preprint, submitted April 11, 2023. <http://arxiv.org/abs/2304.05376>.

↑

- Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. “Language models are few-shot learners.” *Advances in neural information processing systems* 33 (2020): 1877-1901.

↑

- Chakrabarty, Abel Souza, Shruti Jasoria, Basundhara Chakrabarty, Alexander Bridgwater, Axel Lundberg, Filip Skogh, Ahmed Ali-Eldin, David Irwin, and Prashant Shenoy. 2023. “CASPER: carbon-aware scheduling and provisioning for distributed web services.”

↑

- Chips and Science Act, H.R.4346, 117th Congress (2021).

↑

- Chui, Michael, Lareina Yee, Bryce Hall, and Alex Singla. 2023. “The State of AI in 2023: Generative AI’s Breakout Year.”

↑

- Cihon, Peter, Matthijs M Maas, and Luke Kemp. 2020. “Fragmentation and the future: investigating architectures for international AI governance.” *Global Policy* 11, no 5 (2020): 545–556.

↑

- Ciroth, Andreas, and Rickard Arvidsson, eds. 2021. *Life Cycle Inventory Analysis. LCA Compendium – The Complete World of Life Cycle Assessment*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-62270-1>. ↵
- Clarke, Harrison. 2023. “Big Data’s Impact: Optimizing AI with Vast Datasets.” <https://www.harrisonclarke.com/blog/big-datas-impact-optimizing-ai-with-vast-datasets>.

↑

- Clément, Louis-Philippe P.-V.P., Quentin ES Jacquemotte, and Lorenz M Hilty. 2020. “Sources of variation in life cycle assessments of smartphones and tablet computers.” *Environmental Impact Assessment Review* 84 (2020): 106416. <https://doi.org/10.1016/j.eiar.2020.106416>.

↑

- Coeckelbergh, Mark. 2021. “AI for climate: freedom, justice, and other ethical and political challenges.” *AI and Ethics* 1, no. 1(2021): 67–72.

↑

- Council, E U. 2023. “Artificial intelligence act: council and parliament strike a deal on the first rules for AI in the world.” December 9, 2023. <https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/>.

↑

- Dally, Bill. 2023. “Wide Horizons: NVIDIA Keynote Points Way to Further AI Advances.” <https://blogs.nvidia.com/blog/hot-chips-dally-research/>.

↑

- Das, Sujit, and Elizabeth Mao. 2020. “The global energy footprint of information and communication technology electronics in connected internet-of-things devices.” *Sustainable Energy, Grids and Networks*, 24 (2020) 100408. <https://doi.org/10.1016/j.segan.2020.100408>.

↑

- Dauvergne, Peter. 2022. “Is artificial intelligence greening global supply chains? exposing the political economy of environmental costs.” *Review of International Political Economy* 29, no.3 (2022): 696–718.

↵

- de Vries, Alex. 2023. “The growing energy footprint of artificial intelligence.” *Joule* 7, no. 10 (2023): 2191–2194.

↵

- Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M.A., Senior, A., Tucker, P., Yang, K. and Le, Q., 2012. Large scale distributed deep networks. *Advances in neural information processing systems*, 25.

↵

- Dean, Jeffrey, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc aurelio Ranzato, et al. 2012. “Large Scale Distributed Deep Networks.” *Advances in Neural Information Processing Systems* 25 (2012). https://proceedings.neurips.cc/paper_files/paper/2012/file/6aca97005c68f1206823815f66102863-Paper.pdf.

↵

- Debnath, Ramit, Felix Creutzig, Benjamin K. Sovacool, and Emily Shuckburgh. 2023. “Harnessing human and machine intelligence for planetary-level climate action.” *NPJ Climate Action* 2, no 1 (2023): 20. <https://doi.org/10.1038/s44168-023-00056-3>.

↵

- Debus, Charlotte, Marie Piraud, Achim Streit, Fabian Theis, and Markus Götz. 2023. “Reporting electricity consumption is essential for sustainable AI.” *Nature Machine Intelligence* 5, no. 11 (2023): 1176–1178. <https://doi.org/10.1038/s42256-023-00750-1>.

↵

- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. “Bert: Pre-training of deep bidirectional transformers for language understanding.” Preprint, submitted October 11, 2018. <http://arxiv.org/abs/1810.04805>.

↵

- Du, Nan, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, et al. 2021. “Glam: Efficient scaling of language models with mixture-of-experts.” In *International Conference on Machine Learning*, 5547-5569. PMLR, 2022

↵

-

[↵](#)

- Executive Order 14030 of May 20, 2021. Executive Order on Climate-Related Financial Risk.

[↵](#)

- Fauré, Eléonore, Yevgeniya Arushanyan, Elisabeth Ekener, Sofiia Miliutenko, and Göran Finnveden. 2017. “Methods for assessing future scenarios from a sustainability perspective.” *European Journal of Futures Research* 5, no 1 (2017): 1-20. <https://doi.org/10.1007/s40309-017-0121-9>.

[↵](#)

- Ferreboeuf, Hugues. 2019. “LEAN ICT- Towards Digital Sobriety.” The Shift Project. https://theshiftproject.org/wp-content/uploads/2019/03/Lean-ICT-Report_The-Shift-Project_2019.pdf.

[↵](#)

- Finkbeiner, Matthias, Atsushi Inaba, Reginald B.H. Tan, Kim Christiansen, and Hans Jürgen Klüppel. 2006. “The new international standards for life cycle assessment: ISO 14040 and ISO 14044.” *The International Journal of Life Cycle Assessment* 11(2006): 80-85. <https://doi.org/10.1065/lca2006.02.002>.

[↵](#)

- Fung, Victor, Jiaxin Zhang, Eric Juarez, and Bobby G. Sumpter. 2021. “Benchmarking graph neural networks for materials chemistry.” *NPJ Computational Materials* 7, no. 1 (2021): 84. <https://doi.org/10.1038/s41524-021-00554-0>.

[↵](#)

- Furberg, Anna, Rickard Arvidsson, and Sverker Molander. 2022. “A Practice-based framework for defining functional units in comparative life cycle assessments of materials.” *Journal of Industrial Ecology* 26, no. 3 (2022): 718–730. <https://doi.org/10.1111/jiec.13218>.

[↵](#)

- Furman, Jason, and Robert Seamans. 2019. “AI and the Economy.” *Innovation Policy and the Economy* 19, no. 1 (2019): 161–191.

[↵](#)

- Future of Life Institute. 2023. “Pause giant AI experiments: an open letter.” March 22, 2023. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.

[↵](#)

- Giampietro, Mario, and Kozo Mayumi. 2018. “Unraveling the complexity of the jevons paradox: the link between innovation, efficiency, and sustainability.” *Frontiers in Energy Research* 6 (2018): 26.

[↵](#)

- [glass.ai](https://www.glass.ai). 2023. “Code Red: The AI Armies of The Tech Giants.” *Medium*. March 27, 2023.

↑

- Hallegatte, Stephane, and Nathan L Engle. 2019. “The search for the perfect indicator: Reflections on monitoring and evaluation of resilience for improved climate risk management.” *Climate Risk Management* 23 (2019): 1–6.

↑

- Hanafy, Walid A, Qianlin Liang, Noman Bashir, David E Irwin, and Prashant J Shenoy. 2023. “CarbonScaler: Leveraging Cloud Workload Elasticity for Optimizing-Efficiency.” Preprint, submitted February 17, 2023. <https://doi.org/10.48550/ARXIV.2302.08681>.

↑

- Hintemann, Ralph. 2018a. *Efficiency Gains Are Not Enough: Data Center Energy Consumption Continues to Rise Significantly*. Borderstep Inst. für Innovation und Nachhaltigkeit gGmbH.

↑

- Hintemann, Ralph. 2018b. *Efficiency Gains Are Not Enough: Data Center Energy Consumption Continues to Rise Significantly*. Borderstep Inst. für Innovation und Nachhaltigkeit gGmbH.

↑

- Huang, Yanping, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, and Yonghui Wu. 2019 "Gpipe: Efficient training of giant neural networks using pipeline parallelism." *Advances in neural information processing systems* 32 (2019). https://proceedings.neurips.cc/paper_files/paper/2019/file/093f65e080a295f8076b1c5722a46aa2-Paper.pdf

↑

- IEA (2021), Net Zero by 2050, IEA, Paris <https://www.iea.org/reports/net-zero-by-2050>.

↑

- IEA (2024), Electricity 2024, IEA, Paris <https://www.iea.org/reports/electricity-2024>.

↑

- Ilic, M D. 2016. “Toward a unified modeling and control for sustainable and resilient electric energy systems.” *Foundations and Trends® in Electric Energy Systems* 1, no. 1 (2016): 1–141.

↑

- Ingwersen, Wesley W., and Vairavan Subramanian. 2014. “Guidance for product category rule development: process, outcome, and next steps.” *International Journal of Life Cycle Assessment* 19 (2024): 532–537. <https://doi.org/10.1007/s11367-013-0659-0>.

↑

-

↵

- IPCC. 2022. “Climate change 2022: Mitigation of climate change.” *Contribution of working group III to the sixth assessment report of the Intergovernmental Panel on Climate Change* (2022).
<https://www.ipcc.ch/report/ar6/wg3/>.

↵

- Ipsen, Kikki Lambrecht, Regitze Kjær Zimmermann, Per Sieverts Nielsen, and Morten Birkved. 2019. “Environmental assessment of Smart City Solutions using a coupled urban metabolism—life cycle impact assessment approach.” *The International Journal of Life Cycle Assessment* 24 (2019): 1239–1253.

↵

- Itten, René, Roland Hischier, Anders S G Andrae, Jan C T Bieser, Livia Cabernard, Annemarie Falke, Hugues Ferreboeuf, et al. 2020. “Digital transformation-life cycle assessment of digital services, multifunctional devices and cloud computing.” *The International Journal of Life Cycle Assessment* 25 (2020): 2093–2098. <https://doi.org/10.1007/s11367-020-01801-0>.

↵

- ITU. 2023. “Population of global offline continues steady decline to 2.6 billion people in 2023.” Press release, September 12, 2023. International Telecommunication Union (ITU).
<https://www.itu.int/en/mediacentre/Pages/PR-2023-09-12-universal-and-meaningful-connectivity-by-2030.aspx>.

↵

- Jablonka, Kevin Maik, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D. Bocarsly, Andres M. Bran, Stefan Bringuier, et al. 2023. “14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon.” *Digital Discovery* 2, no. 5 (2023): 1233–1250. <https://doi.org/10.1039/d3dd00113j>.

↵

- Jelinek, Thorsten, Wendell Wallach, and Danil Kerimi. 2021. “Policy brief: the creation of a G20 coordinating committee for the governance of artificial intelligence.” *AI and Ethics* 1, no. 2 (2021): 141–50.

↵

- Jouppi, Norman P, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, et al. 2017. “In-datacenter performance analysis of a tensor processing unit.” In *Proceedings of The 44th Annual International Symposium On Computer Architecture*, 1–12. 2017.
<https://doi.org/10.1145/3140659.3080246>.

↵

-

↑

- Kak, Amba, and Sarah Myers West. 2023. “2023 landscape: confronting tech power.” *AI Now*. <https://ainowinstitute.org/wp-content/uploads/2023/04/AI-Now-2023-Landscape-Report-FINAL.pdf>.

↑

- Kari Beets and Michael Hartnett. 2024. North America Data Center Report | H2 2023. Technical Report. JLL Research. <https://www.us.jll.com/en/trends-and-insights/research/na-data-center-report>

↑

- Katal, Avita, Susheela Dahiya, and Tanupriya Choudhury. 2023. “Energy efficiency in cloud computing data centers: a survey on software technologies.” *Cluster Computing* 26, no. 3 (2023): 1845–1875.

↑

- Kates, Robert W, William R Travis, and Thomas J Wilbanks. 2012. “Transformational adaptation when incremental adaptations to climate change are insufficient.” In *Proceedings of the National Academy of Sciences* 109, no. 19 (2012): 7156–7161.

↑

- Kingma, Diederik P, and Jimmy Ba. 2014. “Adam: A method for stochastic optimization.” Preprint, submitted December 22, 2014. <https://arxiv.org/abs/1412.6980>.

↑

- Kirchhoff, Hannes, Noara Kebir, Kirsten Neumann, Peter W Heller, and Kai Strunz. 2016. “Developing mutual success factors and their application to swarm electrification: microgrids with 100% renewable energies in the Global South and Germany.” *Journal of Cleaner Production* 128 (2016): 190–200.

↑

- Klaaßen, Lena, and Christian Stoll. 2021. “Harmonizing corporate carbon footprints.” *Nature Communications* 12, no. 1 (2021): 1–13.

↑

- Knight, Will. 2023. “Google’s Gemini is the real start of the generative AI boom.” *Wired*, December 7, 2023.

↑

- Knox-Hayes, Janelle. 2016. *The Cultures of Markets: The Political Economy of Climate Governance*. 2016: Oxford, Oxford University Press.

↑

- Krafft, P M, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. 2020. “Defining AI in policy versus practice.” In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 72–78. 2020.

↩

- Kremer, Andreas, Angela Luget, Daniel Mikkelsen, Henning Soller, Malin Strandell-Jansson, and Sheila Zingg. 2023. “As gen AI advances, regulators – and risk functions – rush to keep pace.” McKinsey’s Risk & Company. <https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/as-gen-ai-advances-regulators-and-risk-functions-rush-to-keep-pace>.

↩

- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. “Imagenet classification with deep convolutional neural networks.” *Advances in Neural Information Processing Systems* 25 (2012). https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

↩

- Lechowicz, Adam, Nicolas Christianson, Jinhang Zuo, Noman Bashir, Mohammad Hajiesmaili, Adam Wierman, and Prashant Shenoy. 2024. “The online pause and resume problem: optimal algorithms and an application to carbon-aware load shifting.” Preprint, submitted March 30, 2023. <https://arxiv.org/abs/2303.17551>.

↩

- Li, Baolin, Siddharth Samsi, Vijay Gadepally, and Devesh Tiwari. 2023. “Clover: Toward Sustainable AI with Carbon-Aware Machine Learning Inference Service.” In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM. <https://doi.org/10.1145/3581784.3607034>.
- Ligozat, Anne-Laure, Julien Lefevre, Aurélie Bugeau, and Jacques Combaz. 2022. “Unraveling the hidden environmental impacts of AI solutions for environment life cycle assessment of AI solutions.” *Sustainability* 14, no. 9 (2022): 5172. <https://doi.org/10.3390/su14095172>.

↩

- Luccioni, Alexandra Sasha, Sylvain Viguiet, and Anne-Laure Ligozat. 2022. “Estimating the carbon footprint of bloom, a 176b parameter language model,” *Journal of Machine Learning Research* 24, no. 253 (2023): 1–15.

↩

- Luccioni, Alexandra Sasha, Yacine Jernite, and Emma Strubell. 2023. “Power hungry processing: Watts driving the cost of AI deployment?” Preprint, submitted November 28, 2023. <https://arxiv.org/abs/2311.16863>.

↩

- Luccioni, Sasha. 2023. “Generative AI Models: History, Costs and Risks.” https://docs.google.com/presentation/d/1FRoyzdodKQ7-5rK--gZFFzK_-kvfhzxJQDYxpnA-6jE/.

↑

- Luu, Rachel, Sofia Arevalo, Wei Lu, Bo Ni, Zhenze Yang, Sabrina Shen, Jaime Berkovich, Yu-Chuan Hsu, Stone Zan, and Markus Buehler. 2024. “Learning from Nature to Achieve Material Sustainability: Generative AI for Rigorous Bio-Inspired Materials Design.”

↑

- MacKenzie, D. 2009. “Making things the same: gases, emission rights and the politics of carbon markets.” *Accounting, Organizations and Society* 34, no. 3–4 (2009): 440–55.

↑

- MacLeod, Benjamin P., Fraser G.L. Parlane, Amanda K. Brown, Jason E. Hein, and Curtis P. Berlinguette. 2022. “Flexible automation accelerates materials discovery.” *Nature Materials* 21, no. 7 (2022): 772–726. <https://doi.org/10.1038/s41563-021-01156-3>.

↑

- Mammen, Priyanka Mary, Noman Bashir, Ramachandra Rao Kolluri, Eun Kung Lee, and Prashant J Shenoy. 2023. “CUFF: A configurable uncertainty-driven forecasting framework for green ai clusters.” In *Proceedings of the 14th ACM International Conference on Future Energy Systems*, 266–70. <https://doi.org/10.1145/3575813.3595203>.

↑

- Manyika, James, and Hsiao Sissie. 2023. “An overview of Bard: an early experiment with generative AI.” *AI. Google Static Documents* 2 (2023). <https://Ai.Google/Static/Documents/Google-about-Bard.Pdf.2023>.

↑

- Masanet, Eric, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey. 2020a. “Recalibrating Global Data Center Energy-Use Estimates.” *Science* 367, no. 6481 (2020): 984–986.

↑

- Masanet, Eric, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey. 2020b. “Recalibrating Global Data Center Energy-Use Estimates.” *Science* 367, no. 6481 (2020): 984–86.

↑

- Matisoff, Daniel C., Douglas Noonan, and John J. O’Brien. 2013. “Convergence in environmental reporting: assessing the carbon disclosure project.” *Business Strategy and the Environment* 22, no. 5 (2013): 285–305.

↑

- Mytton, David. 2020. “Hiding greenhouse gas emissions in the cloud.” *Nature Climate Change* 10, no. 8 (2020): 701.

↑

- Nature editorial authors. 2024. “There are holes in Europe’s AI Act – and researchers can help to fill them.” *Nature* 625 (2024): 216. <https://doi.org/10.1038/d41586-024-00029-4>.

↵

- Norem, Josh. 2023. “Nvidia reportedly sold 500,000 H100 AI GPUs in Q3 alone.” *ExtremeTech*, November 28, 2023.

↵

- Norris, Gregory A., Jasmina Burek, Elizabeth A. Moore, Randolph E. Kirchain, and Jeremy Gregory. 2021. “Sustainability health initiative for NetPositive enterprise handprint methodological framework.” *International Journal of Life Cycle Assessment* 26 (2021): 528–542. <https://doi.org/10.1007/s11367-021-01874-5>.

↵

- OpenAI. 2023. “GPT-4 Technical Report.” <https://cdn.openai.com/papers/gpt-4.pdf>.

↵

- Parmesan, Camille, Mike D Morecroft, and Yongyut Trisurat. 2022. “Climate change 2022: impacts, adaptation and vulnerability.” [Research Report] GIEC (2022): hal-03774939. <https://hal.science/hal-03774939/document>.

↵

- Pasek, Anne, Hunter Vaughan, and Nicole Starosielski. 2023. “The world wide web of carbon: toward a relational footprinting of information and communications technology’s climate impacts.” *Big Data & Society* 10, no. 1 (2023): 20539517231158990. <https://doi.org/10.1177/20539517231158994>.

↵

- Patterson, David, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David R So, Maud Texier, and Jeff Dean. 2022. “The carbon footprint of machine learning training will plateau, then shrink.” *Computer* 55, no. 7 (2022): 18–28. <https://doi.org/10.1109/MC.2022.3148714>.

↵

- Perry, Tekla S. 2018. “Move Over, Moore’s Law: Make Way for Huang’s Law.” *IEEE Spectrum*, April 2, 2018.

↵

- Pierce, Matthew. 2020. “High-Performance Networking to Support Critical Workloads for AI and ML.” *redapt*. <https://www.redapt.com/blog/high-performance-networking-to-support-critical-workloads-for-ai-and-ml>.

↵

- Radovanović, Ana, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyiue Xiao, et al. 2023. “Carbon-aware computing for datacenters.” *IEEE Transactions on Power Systems* 38, no. 2 (2022): 1270–80.

↵

- Rasoldier, Aina, Jacques Combaz, Alain Girault, Kevin Marquet, and Sophie Quinton. 2022. “How realistic are claims about the benefits of using digital technologies for ghg emissions mitigation?” In *LIMITS 2022-Eighth Workshop on Computing within Limits*. 2022. <https://doi.org/10.21428/bf6fb269.6d7bd21b>.

↵

- Riahi, Keywan, Detlef P. van Vuuren, Elmar Kriegler, Jae Edmonds, Brian C. O’Neill, Shinichiro Fujimori, Nico Bauer, et al. 2017. “The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: an overview.” *Global Environmental Change* 42 (2017), 153–168. <https://doi.org/10.1016/j.gloenvcha.2016.05.009>.

↵

- Richards, Catherine E., Asaf Tzachor, Shahar Avin, and Richard Fenner. 2023. “Rewards, risks and responsible deployment of artificial intelligence in water systems.” *Nature Water* 1 (2023): 422–432. <https://doi.org/10.1038/s44221-023-00069-6>.

↵

- Rolnick, David, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, et al. 2022. “Tackling climate change with machine learning.” *ACM Computing Surveys (CSUR)* 55, no. 2 (2022): 1–96. <https://doi.org/10.1145/3485128>.

↵

- Roussilhe, Gauthier, Anne Laure Ligozat, and Sophie Quinton. 2023. “A long road ahead: a review of the state of knowledge of the environmental effects of digitization.” *Current Opinion in Environmental Sustainability* 62 (2023): 101296. <https://doi.org/10.1016/j.cosust.2023.101296>.

↵

- S.3732 - 118th Congress (2023-2024): Artificial Intelligence Environmental Impacts Act of 2024. (2024, February 1). <https://www.congress.gov/bill/118th-congress/senate-bill/3732>

↵

- Scherer, Matthew U. 2015. “Regulating artificial intelligence systems: risks, challenges, competencies, and strategies.” *Harv. JL & Tech.* 29 (2015): 353.

↵

- Schiff, Daniel, Justin Biddle, Jason Borenstein, and Kelly Laas. 2020. “What’s next for AI ethics, policy, and governance? A global overview.” In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 153–58, 2020.



- Science Based Targets. 2019. “Foundations of science-based target setting.” Version 1.0, April 2019. <https://sciencebasedtargets.org/resources/files/foundations-of-SBT-setting.pdf>.



- Shazeer, Noam, Azalia Mirhoseini, Krzysztof Maziarczyk, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. “Outrageously large neural networks: the sparsely-gated mixture-of-experts layer.” Preprint, submitted January 23, 2017. <https://arxiv.org/abs/1701.06538>.



- Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. “Dropout: a simple way to prevent neural networks from overfitting.” *The Journal of Machine Learning Research* 15, no. 1 (2014): 1929–1958.



- Steffen, Will, Katherine Richardson, Johan Rockström, Sarah E. Cornell, Ingo Fetzer, Elena M. Bennett, Reinette Biggs, et al. 2015. “Planetary boundaries: guiding human development on a changing planet.” *Science* 347, no. 6223 (2015): 1259855. <https://doi.org/10.1126/science.1259855>.



- Subramanian, Akshay, Wenhao Gao, Regina Barzilay, Jeffrey C. Grossman, Tommi Jaakkola, Stefanie Jegelka, Mingda Li, et al. 2024 “Closing the Execution Gap in Generative AI for Chemicals and Materials: Freeways or Safeguards.” An MIT Exploration of Generative AI.



- Sukprasert, Thanathorn, Abel Souza, Noman Bashir, David Irwin, and Prashant Shenoy. 2024. “On the Limitations of Carbon-Aware Temporal and Spatial Workload Shifting in the Cloud.” Zenodo. <https://doi.org/10.5281/ZENODO.10790855>. ↵
- Switzer, Jennifer, Gabriel Marciano, Ryan Kastner, and Pat Pannuto. 2023. “Junkyard computing: repurposing discarded smartphones to minimize carbon.” In *Proceedings of the ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 400–412. <https://doi.org/10.1145/3575693.3575710>.



- The Global Economy. 2023. Electricity production capacity - Country rankings https://www.theglobaleconomy.com/rankings/electricity_production_capacity/.

↑

- Thiede, John, Noman Bashir, David Irwin, and Prashant Shenoy. 2023. “Carbon containers: a system-level facility for managing application-level carbon emissions.” In *Proceedings of 14th Symposium on Cloud Computing*, 17–31. 2023.

↑

- Ulnicane, Inga. 2022. “Artificial intelligence in the European Union: policy, ethics and regulation.” In *The Routledge Handbook of European Integrations*. Taylor & Francis, 2022.

↑

- Uptime Institute. 2024. Five Data Center Predictions for 2024. Technical Report. Uptime Institute. <https://uptimeinstitute.com/resources/research-and-reports/five-data-center-predictions-for-2024>.

↑

- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is all you need.” In *Advances in Neural Information Processing Systems*, 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

↑

- Vinitzky, Eugene, Aboudy Kreidieh, Luc Le Flem, Nishant Kheterpal, Kathy Jang, Cathy Wu, Fangyu Wu, Richard Liaw, Eric Liang, and Alexandre M Bayen. 2018. “Benchmarks for reinforcement learning in mixed-autonomy traffic.” In *Conference on Robot Learning*, 399–409. PMLR, 2018. <https://github.com/flow-project/flow>.

↑

- Ward, Isabella, and Natalie Lung. 2023. “Big Tech’s Year of Partnering Up With AI Startups.” *Bloomberg*, December 18, 2023.

↑

- Weidema, Bo Pedersen. 2003. *Market Information in Life Cycle Assessment*. Vol. 863. København, Denmark: Miljøstyrelsen, 2003.

↑

- West, Sarah Myers. 2023. “Competition authorities need to move fast and break up AI.” *Financial Times*, April 17, 2023.

↑

- Wiesner, Philipp, Ilja Behnke, Dominik Scheinert, Kordian Gontarska, and Lauritz Thamsen. 2021. “Let’s wait awhile: how temporal workload shifting can reduce carbon emissions in the cloud.” In *Proceedings of*

↵

- Wijewardane, Nuwan K., Yufeng Ge, Skye Wills, and Terry Loecke. 2016. “Prediction of soil carbon in the conterminous United States: visible and near infrared reflectance spectroscopy analysis of the rapid carbon assessment project.” *Soil Science Society of America Journal* 80, no. 4 (2016): 973–982.
<https://doi.org/10.2136/sssaj2016.02.0052>.

↵

- Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, and Philip E Bourne. 2016. “The FAIR Guiding Principles for scientific data management and stewardship.” *Scientific Data* 3, no. 1 (2016): 1–9.

↵

- Wright, Dustin, Christian Igel, Gabrielle Samuel, and Raghavendra Selvan. 2023. “Efficiency Is Not Enough: A Critical Perspective of Environmentally Sustainable AI.” Preprint, submitted September 5, 2023.
<https://arxiv.org/abs/2309.02065>.

↵

- Wu, Carole-Jean, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, et al. 2022. “Sustainable AI: environmental implications, challenges and opportunities.” *Proceedings of Machine Learning and Systems* 4 (2022): 795–813.

↵