Schneider Electric™
Sustainability Research Institute

Digital Series

Artificial Intelligence and Electricity

A System Dynamics Approach

Rémi Paccou and Fons Wijnhoven

December 2024

Life Is On | Schneider Electric

Welcome and Key Insights

Rémi Paccou
Director of Sustainability Research,
Schneider Electric™ Sustainability Research Institute

Dear Reader,

I'm pleased to welcome you to our research on AI and Electricity: A System Dynamics Approach. As a Research Institute focusing on the links between energy and sustainability, we are keen to better understand the relationships between Artificial Intelligence and the dynamics of energy transitions. We hope to offer an analysis that opens up potential paths for thinking and discussing AI development for the future.

We believe this study comes at a critical time, as AI's explicit and growing influence intersects with pressing environmental concerns. Its exponential growth has raised important questions about its energy requirements and potential impacts on energy systems and climate change. In response, we have undertaken this research to address these issues through a careful, multi-faceted examination of possible futures.

Central to our work is the acknowledgment of emerging AI **Schools of Thought**. These schools, though nascent, are already shaping the collective unconscious of AI development and directing the archetypal patterns emerging in AI evolution.

As Carl Jung observed, "Thinking is difficult, that's why most people judge". In a world often dominated by hype, critical thinking about AI is essential. The waves on the surface of the saturated public debate reveal fundamentally powerful undercurrents of AI thought, shaping ideologies and new courses of action. By qualifying these emerging schools, we aim to bring them into the spotlight and foster debates on their implications for climate and energy.

To enrich these Schools of Thought with quantitative material, we have employed system dynamics to construct four scenarios of AI development and their associated impacts on electricity consumption. These scenarios are not predictions but rather tools to understand the complex factors shaping our future. As you will discover, they span a range of possibilities: from **Sustainable AI** development to **Limits To Growth,** including more radical scenarios such as **Abundance Without Boundaries** and even the possibility of **Energy Crises** caused by AI.

Our methodology combines bottom-up and top-down approaches, leveraging the strengths of each to mitigate their respective weaknesses. We have intentionally drawn from a diverse range of sources - including industry data, academic theories and studies, and expert knowledge - to construct the most comprehensive view of possible AI electricity futures.

Our research reveals key global insights shaping the future of AI and electricity. These insights underscore the critical paths needed to either converge towards a sustainable future or to mitigate risks inherent in undesirable scenarios.

Sustainable AI should essentially be the result of efficiency, frugality, and demonstrable impact. Conversely, unrestricted abundance can disrupt mutiple systems, hinder decarbonization, and lead to waste. Furthermore, mismatches between energy demand and infrastructure can cause local shortages with global ripple effects.

We also provide scenario-specific insights. In the **Sustainable AI** scenario, we highlight the emerging dominance of generative AI inferencing in electricity consumption, while noting the continued importance of traditional AI in decarbonization efforts. In the **Limits To Growth** scenario, we examine the constraints facing generative AI training and deployment. As part of the **Abundance** scenario, we identify the risks associated with entropic abundance, such as the questionable legacy of building an oversized AI infrastructure and issues related to AI access inequality. We also address the challenges of insufficient grid planning and the potential for localized **Energy Crises.**

As one of the architects of this research, I invite you to critically consider these potential futures as you read. This research is not-meant to be prescriptive; instead, we hope it serves as a starting point for informed discussion and decision-making. We present our findings with the understanding that AI is a rapidly evolving field and that our knowledge is constantly growing. Our hope is that this research will contribute meaningfully to ongoing conversations about sustainable AI development, energy policy, human prosperity balanced with frugality, and technological innovation.

By exploring these potential futures, we aim to equip stakeholders with the knowledge needed to navigate the challenges and opportunities that lie ahead.

Thank you for joining us in this exploration of AI and electricity futures. We look forward to the discussions and further research that this work may inspire.

Rémi Paccou
Director of Sustainability Research,
Schneider Electric™ Sustainability Research Institute

Industry Perspectives

Powering a Sustainable AI Future

Jason Oxman
President and Chief Executive Officer of the Information Technology Industry Council (ITI)

The meteoric rise of artificial intelligence technologies has provided humanity with an incredibly potent tool that can tackle challenges we once thought unsolvable. AI is helping doctors identify otherwise undetectable cancerous growths, streamlining agricultural processes to decrease cost and increase harvest yield, and providing accurate flood predictions to prepare emergency responders. The potential benefits of harnessing AI technology are hard to overstate. AI, like electricity or the automobile, has the capability to completely revolutionize the way we live our lives. But like with any other technology, there are challenges we must address to ensure that AI can be harnessed safely and efficiently.

The innovation ecosystem must become more sustainable so that emerging technologies, such as AI, can continue to grow responsibly. Existing data center infrastructure, which underpins most current AI technology, requires significant energy to function, and will need additional resources and space to support the anticipated growth in AI use.

This raises concerns about the potential strain on power grids and the long-term environmental impacts if the demand for energy to power AI continues to rise at its current rate. While broader societal trends, including the electrification of manufacturing, transportation, and infrastructure, also play a substantial role in this increased demand for energy, it's important that industry ensures that the benefits of AI technology outweigh any potential harm inflicted on the fragile planet.

Fortunately, a significant body of evidence suggests that it is possible to implement AI technology at scale in ways that minimize environmental impacts and maximize societal benefits. At this very moment, industry leaders around the world are hard at work designing the next generation of technology with necessary sustainability considerations in mind from the start to help overcome these challenges. Improvements in computing capabilities, including hardware and software, and optimized facility design are helping increase the energy efficiency of data centers, and the private sector is working alongside policymakers to continue to improve its overall environmental impact.

In its analysis, Schneider Electric's Sustainability Research Institute Report on AI and electricity consumption explores the multifaceted dynamics of this challenge. Importantly, the report offers a tangible way to build a reliable, resilient, and modern energy infrastructure that supports sustainable innovation and ensures sufficient energy access for everyone.

Time and time again, humanity has encountered trials that seem beyond our current knowledge and resources, only for innovators to devise new means to overcome these obstacles and push forward. Historically, the tech industry has played a key part in this process, providing the funding, resources, and workforce necessary to work through these barriers to technological progress. The technological and societal advancements we enjoy as the result of these efforts, inventions like the airplane, penicillin, and the internet, now form the very bedrock of our way of life.

By viewing AI sustainability as an opportunity for growth and development, we have the opportunity to spark a wave of innovation that could reshape industries and inspire innovations in responsible technology. This shift would not only advance AI technology but could also lead to breakthroughs in renewable energy production and distribution, efficiency in computing, and eco-conscious data infrastructure, reinforcing the tech industry's role as a force for progress. In doing so, AI might not only fulfill its promise to revolutionize healthcare, agriculture, and other sectors but also become a cornerstone in building a world that is smarter, more resilient, and more resourceful.

As the digital partner for sustainability and efficiency across industries, Schneider Electric's analysis can help global policymakers, business leaders, and other stakeholders realize this vision.

Jason Oxman

President and Chief Executive Officer,
Information Technology Industry Council (ITI)

Research Perspectives

Beyond the Streetlight: A Systems Analysis of AI's Energy Future

Dr. Vlad C. Coroamă, PhD in Ubiquitous Computing, ETH Zurich
Founder of the Roegen Centre for Sustainability (RC4S)
Affiliated researcher with the TU Berlin, Germany

"*Hey, Jamie, what are you looking for?*", asked Alex, seeing her friend intently scrutinizing the sidewalk. "*Alex, hi! I lost my keys earlier*", Jamie replied, only briefly looking up. "*It's freezing cold tonight, let me join in; but are you sure they must be around here?*" "*Oh, no, almost certainly not! I think they fell somewhere around the corner. But it's so dark back there; I prefer looking for them here, under the streetlight.*"

Dan Schien (University of Bristol) was recounting this parable known as the "streetlight effect" just a few days ago. In a fundamental discussion on assessment principles, both of us agreed that it is an increasing challenge to address the energy and environmental effects of digitalization – and in particular of AI – via bottom-up methods.

This is not to dismiss them: Bottom-up methods are often based on broad and detailed primary data. By highlighting influence pathways, they offer strong explanatory power. But they are also limited to the evident mechanisms. The subtle and hard to grasp often elude them; deeply intertwined mechanisms puzzle them. Because, as in Jamie's search for keys, not all that is in front of our eyes is worth analyzing, and not all that is worth analyzing lies in plain sight.

For trends and future analyzes in particular, bottom-up methods quickly reach their limits. AI's current energy consumption can indeed be approached via estimates of the GPU stock, average utilization rate, and average power consumption. Future deployment, however, depends on much more: macroeconomic and geopolitical context, technological innovations and efficiency leaps, but also possible limits to growth and further limiting factors, including policies. Crucial are also the multitude of domains in which a technology can be deployed. This is particularly relevant for a general-purpose technology such as AI that is also a method of invention – and might, as a consequence, spark the next industrial revolution, as this study's author, Rémi Paccou, recently argued in another outstanding publication, „AI for Impact: A Method for Guiding AI-Energy Applications at Scale"

This multitude of complex and interwoven pathways requires top-down assessments; be they environmentally extended input-output analyzes, sustainability transformations, or – as this work does – quantitative system dynamics. And it does so brilliantly: In four archetypal exploratory scenarios, the study covers dozens of factors. They are modelled in a core system dynamics model, further sub-models focusing on training and inference of specific AI flavors, as well as two "envelopes", which are not system dynamics orthodoxy, but crucial to the far-reaching analysis.

Results are eloquently presented and extremely well-suited to spur discussion and further analysis. The entire study is refreshingly non-ideological. It certainly acknowledges the chances of AI; while reading the "Limits to growth" and "Energy Crisis" scenarios, one can feel the author's sorrow. But it is also deeply concerned with the environment, and as far as possibly imaginable from a greenwashing exercise by an industry representative. It is neither utopian in its technology assessment nor demonizing it. In a world of increasingly shrill tones, among conflicting alarmism and efforts to play down any criticism, this study brings much-needed calmness and objectivity to the discussion.

A brilliant read

The study is not only timely and valuable; it is also such an enjoyable read! The flow is so logically organized, with increasing levels of detail, that it reads like a good symphony, where one first hears the basic theme, which is then developed, with more and more instruments joining in, increasingly revealing its full complexity. One in which those few beginning accords can always be perceived, coming back with more and more strength and emphasis.

And as in any good symphony, it is not only the main theme being repeated and developed. Side stories emerge and grow in their own right, at first independently of the main theme, to then subtly contribute to it, until they become fully integrated. This is particularly evident when the 4 scenarios are presented in detail; each brings a new story that has not been told before, at first discreetly, but then progressively becoming part of the main harmonious polyphony.

My former PhD advisor, the outstanding scientist and humanist Prof. Friedemann Mattern, argues that science is increasingly difficult to read and enjoy, as most papers nowadays are aimed at reviewers, and not at their later readers. Not so this study. It was clearly meant for you, dear reader. And like any good writing, it is addictive. Once starting it one late evening, I had to go deep into the night to finish it. You have been warned.

Dr. Vlad C. Coroamă
Founder of the Roegen Centre for Sustainability (RC4S)

Science Perspectives

System Thinking for the Future

Pr. Fons Wijnhoven, PhD in Information Systems
Associate Professor at the University of Twente in the Netherlands

Artificial Intelligence (AI) is increasingly shaping our society, and this article explores a crucial aspect: its direct impact on data center electricity consumption. As we look towards the future, understanding the potential implications of AI on data center energy demands becomes increasingly important. In the coming years, we may see data center electricity needs competing with other essential societal functions, from mobility and heating to medical device electricity operations.

By attempting to extend our view beyond current knowledge and data, we hope to shed light on potential futures, revealing both opportunities and challenges. This foresight might assist stakeholders in making informed decisions that could influence the development of AI and data centers. These stakeholders include policymakers, electronic component manufacturers, data center managers, and AI users.

While AI offers significant potential, it's important to consider its growth carefully to avoid straining critical resources. To this end, we've developed four scenarios of data center energy consumption:

The Symbiotic-Sustainable Scenario

In a world where technology and nature coexist harmoniously, data centers could evolve into sustainable powerhouses. By integrating renewable energy sources, optimizing energy efficiency, and fostering symbiotic relationships with local ecosystems, these digital hubs could contribute to a greener future.

Advanced AI, powered by sustainable energy, could then be harnessed to address global challenges like climate change, healthcare, and poverty. This symbiotic-sustainable scenario envisions a future where AI growth is not only possible but beneficial to society as a whole.

Three Alternative Paths

However, the realization of this ideal future is contingent on several factors. One potential scenario is where external influences, such as regulatory hurdles or economic downturns, could hinder the adoption of sustainable technologies and limit the growth of AI. Another possibility is an unbounded growth model, where continuous innovation and technological breakthroughs might overcome current constraints, leading to rapid AI advancement. Yet, this rapid growth could also trigger an AI-induced energy crisis, as the increasing computational demands of AI systems could outpace the development of sustainable energy infrastructure.

This research acknowledges the uncertainty of the future while suggesting the importance of preparedness. By incorporating insights from literature and industry professionals, we have attempted to develop a dynamic systems approach to AI impacts. This perspective considers long-term effects as potentially evolving through reinforcements, balances, and rebounds of current trends.

Our system dynamics models aim to serve as tools for stakeholders to engage in discussions about the future. We have attempted to simulate potential outcomes of AI's impact on data center electricity consumption for the coming decade. However, we recognize that the end of this period will likely bring technological and social changes that would need to be considered for understanding long-term impacts of AI.

While this paper focuses on direct impacts, it also suggests the need for future studies exploring the indirect effects of increased data center usage on the economy, society, and ecology. As we consider the future of AI and data centers, we hope this work contributes to a thoughtful and responsible approach to development, aiming for a path that balances progress with sustainability.

Pr. Fons Wijnhoven
Co-author of this study

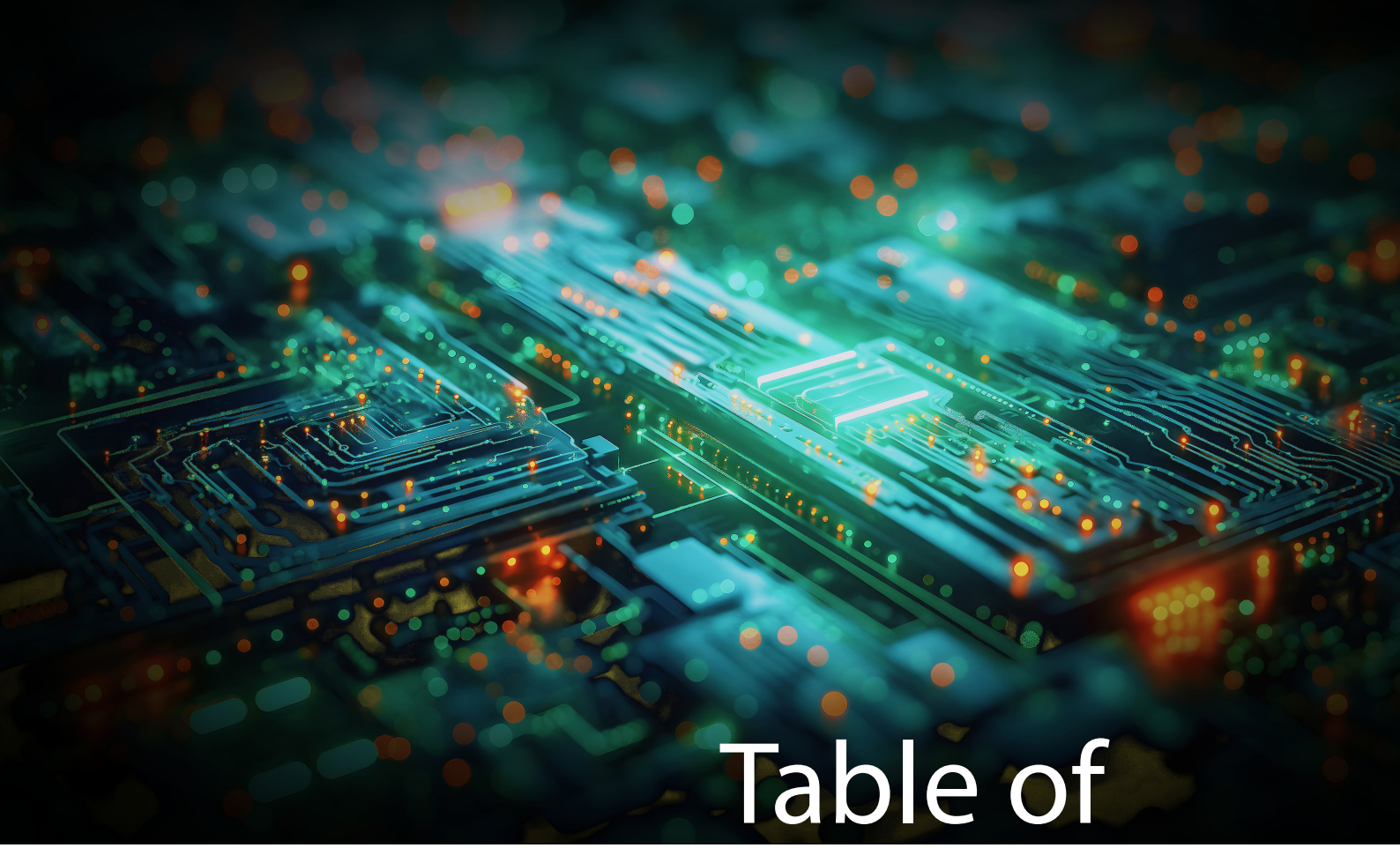


Table of Contents

List of Exhibits

Problem Statement

The escalating electricity consumption of Artificial Intelligence (AI) has become a focal point of concern for media outlets, policymakers, investors, and the public. This heightened interest is driven by the rapid expansion of AI computing by big tech companies, resulting in substantial increases in their electricity usage[1, 2, 3]. Recent announcements regarding planned AI data center expansions[4], AI hardware shipments[2], and significant investments in AI processor manufacturing[5] have further fueled these concerns. This rapid growth represents only the tip of the iceberg, with far-reaching implications that extend beyond mere electricity consumption forecasts[1, 3, 6].

The breadth of AI's projected electricity demand encompasses a wide array of interconnected issues. These issues range from infrastructure challenges to supply chain disruptions and socio-economic concerns[7, 8]. The unprecedented demand for AI has surged electricity consumption, straining global energy grids and semiconductor manufacturing capabilities[5, 9], raising concerns about energy security and environmental impacts. The reliance on specialized hardware, such as GPUs (Graphics Processing Units) and TPUs (Tensor Processing Units), has created stress in the AI supply chain, constraining the scalability of AI development and deployment. Moreover, the challenges of training generative AI models have become increasingly apparent, with explicit links to power availability, chip manufacturing capacities, data scarcity for AI training, and increasing costs[2, 10], foreshadowing future issues for generative AI inferencing in the near future.

At the global level, AI data center electricity use currently accounts for less than 0.3% of worldwide electricity demand[2]. However, rapid growth is expected across multiple geographies in the coming years. The International Energy Agency (IEA) projects global data center electricity demand to more than double between 2022 and 2026, reaching over 1,000 terawatt-hours (TWh) in 2026 (equivalent to Japan's electricity consumption). The IEA further states that AI energy demand will multiply by at least 10 in 2023-2026, highlighting AI's significant contribution to this surge[3]. This represents a significant change from the estimated 460 TWh consumed in 2022[3]. Goldman Sachs Research forecasts data center power demand to grow 160% by 2030, potentially rising from 1-2% of overall power consumption to 3-4% by decade's end[6].

At local levels, AI data center developments are raising significant concerns about regional grid stress and power capacity. The rapid growth is particularly acute in the United States. Only 15 states host 80% of the country's data centers[11]. In Northern Virginia's "Data Center Alley" data centers already consume 25% of the region's electricity, potentially rising to nearly 50% of the state's total in a high-growth scenario[4], prompting utilities to propose new fossil fuel infrastructures[4]. In Ireland, data centers could consume up to 32% of the country's electricity by 2026[12] [13], exceeding the combined energy usage of all urban homes in 2023, which accounted for 21% of the total electricity consumption[12]. The Netherlands is also grappling with similar issues, with Amsterdam implementing new power usage rules for data centers, including fines for failing to implement power management protocols[14].

Addressing these challenges requires decision-makers to utilize robust, data-driven modeling for more accurate forecasting[4]. Indeed, understanding AI's energy consumption presents a multifaceted challenge. It reflects its systemic interplay with technology and human development trajectories. This complexity is further amplified by AI's inherent nature of being a candidate for transformative change across industries[1, 15]. Beyond the hype, AI's integration is becoming profound, with experts like recent Nobel laureate Geoffrey Hinton viewing it as a key driver of a fourth industrial revolution, as predicted and theoretically framed by Nicholas Crafts in 2020[16]. As a General Purpose Technology (GPT) and an Invention of a Method of Invention (IMI), AI's impact can span literally every sector and play a role in climate change mitigation and adaptation, for better or worse[8].

AI's systemic nature could create feedback loops that amplify its energy consumption. While AI optimization can lead to efficiency gains and cost savings, this may paradoxically encourage wider AI adoption. Consequently, overall energy demand may increase. AI and energy systems interdependence means that disruptions or bottlenecks in one area can have widespread consequences, as exemplified by Hinton's recent observations on data scarcity hindering scientific progress in AI. Hence, modeling the system dynamics of AI development is crucial for informed decision-making[17].

To address these issues, this research examines four distinct scenarios for AI electricity consumption through 2035. We explore a **Limits To Growth** scenario, which shows that limiting AI electricity consumption may not guarantee sustainable development; a **Sustainable AI** scenario, suggesting a measured approach; an **Abundance Without Boundaries** scenario, demonstrating risks of excessive energy consumption and infrastructure strain; and a **Energy Crisis** scenario, illustrating potential energy instability when AI development outpaces grid capacity. Based on these explorations, we provide key recommendations for a balanced approach to AI development that considers resource-conscious innovation, socio-ecological responsibility, and robust governance to ensure sustainable and equitable progress for all.

Existing Forecasts of Electricity Use

The emerging field of AI electricity footprint estimation faces complex challenges, notably lacking comprehensive scenarios and forecasts. Since 2023, a surge in research has revealed divergent modeling methodologies. Additionally, there are varying perceptions of AI's future[18]. These differing perspectives, driven by methodological disagreements and data gaps, are shaping the discipline's trajectory. To address these issues, we first examine the strengths and weaknesses of bottom-up and top-down approaches. Second, recognizing the potential unreliability of relying on a single projection for long-term future impacts, we approach insights into the future by simulating multiple scenarios based on fundamental AI **Schools of Thought**. Understanding and addressing these emerging schools of thought is crucial for reliably assessing and managing AI's future electricity use[2].

Assessing the strengths and weaknesses of bottom-up and top-down approaches

In a recent article, Masanet et al. propose a modern bottom-up approach to estimating AI electricity use, arguing that despite higher power requirements, AI data centers can be modeled similarly to conventional ones[19]. This method enables scenario analysis of future electricity demand based on various factors. The authors identify several "traps" that often lead to overestimations in AI data center electricity consumption, such as multiplying AI server rated power by sales data, using advertised power capacities and assumed Power Usage Effectiveness (PUE) values, and basing projections on utility permits or assumed growth rates[19]. To implement this approach, various data points are crucial, including AI hardware power profiles, data center reporting, cooling technology performance data, and market evolution[19]. Masanet et al. identify shortcuts that fail to account for discrepancies between rated and actual power use, unclear definitions, non-representative cooling approaches, and overprovisioning. While this bottom-up approach is appealing, it requires significant alignment across stakeholders to provide critical data, positioning it as a structural and long-term method for forecasting AI electricity use[19].

In contrast, De Vries' study employs a top-down approach, utilizing industry-level assumptions and trends to project future AI energy consumption[2]. This methodology relies on broad statistics and simplified parameters, and using market leader NVIDIA data for information on AI server energy consumption[2]. While this approach allows for high-level estimates and scenario-based analysis, it has limitations due to the lack of detailed data from individual AI systems or companies[20]. Additionally, the assumption of servers operating continuously at full capacity may overestimate energy consumption[2]. However, despite these shortcomings, the study serves as an important starting point for discussing and addressing the potential environmental impact of AI's rapid growth[21].

Both approaches highlight the need for more granular data and transparency from AI developers to improve the accuracy of energy consumption projections[22]. Hence, as the field of AI continues to evolve rapidly, a combination of bottom-up and top-down approaches, along with increased data sharing and collaboration among stakeholders, may provide the most comprehensive understanding of AI's energy footprint[23].

Understanding representations of the future is essential for accurate modeling of electricity use projections

Most existing projections, grounded in data, often align with a future-specific representation, which can be shaped by researcher's positionality statements or stakeholder priorities, which significantly influence data interpretation and conclusions[24]. These projections typically blend elements of rationalism (bottom-up or top-down data, assumptions, and models to form a logical framework) and subjectivism (cultural biases, litanies, worldviews, and metaphors) that shape interpretations and projections[25]. As Ekchajzer et al.[26] caution, we must be wary of overly simplistic, narrow, and static future projections. This is particularly relevant as we witness the emergence of organized thought in the landscape of AI projections, with contributions from a diverse range of entities, including energy analysts, banks, consulting firms, foundations, IT associations, federations, and public institutions[27].

At the most fundamental level, these projections are driven by individuals whose ideas are disseminated exponentially through journals, articles, books, and social networks. Over time, these intellectual efforts are amplified by organizations, giving rise to distinct practices, movements, and beliefs that reflect collective perspectives. As these practices coalesce, they contribute to deeper intellectual currents that have started to shape the debates on AI trajectory, especially in terms of infrastructural footprint. Quite naturally, these converging forces are yet forming - explicitly or not - distinct schools of thought, still taking shape within an information-saturated debate, each offering unique perspectives on the future of AI. Ultimately, these emerging schools of thought will profoundly shape the development of AI. To foster debate and analysis, a plurality of perspectives is essential. While each school of thought has its limitations, their collective insights offer a broader spectrum for thinking AI electricity future[28]. These diverse viewpoints can inform more robust scenario planning by discussing assumptions, identifying blind spots in forecasting, and stimulating constructive debates on the multifaceted factors influencing AI electricity use[29].

Our approach utilizes system dynamics modelling[33] to explore four scenarios derived from distinct **Schools of Thought**. While many existing studies adhere to a School of Thought and forecasting approach (either bottom-up or top-down)[32], our research suggests that a combination of these methods can forge a more nuanced and comprehensive vision of potential futures[30]. Our approach not only stimulates critical thinking but also equips stakeholders to navigate a broader spectrum of scenarios[31]. We start by examining the **Sustainable AI** School of Thought which advocates for AI development that prioritizes sustainable practices within planetary boundaries, envisioning a future where AI advances harmoniously with environmental stewardship. The second, the **Limits to Growth** School, emphasizes the constraints on AI development, including power constraints, supply chain tensions, and material scarcity. The third school is the **Abundance Without Boundaries** School, which believes in technology's ability to overcome challenges but risks leading to unchecked AI growth, potentially concentrating power and exacerbating inequalities. Finally, the **Energy Crisis** School warns of the potential negative impacts of ungoverned AI development, particularly due to electricity supply limitations. While the first two scenarios are potentially closer to the most likely future evolution, the third and fourth scenarios push the boundaries and highlight potential risks that need to be addressed.

Designing Scenarios for the Future

Following a systems dynamics approach of growth and decline with potential reinforcements and balances, we construct four scenarios aligned with existing Schools of Thought and translated into system dynamics models. The appendices provide a detailed characterization and parametrization of different Schools, as identified by their emergent proponents, organizations, editorial stances, and thought leaders.

Scenario 1: Sustainable AI

In this scenario, Sustainable AI Advocates trust that AI-driven advancements in energy efficiency and resource optimization result in substantial improvements in data center operations. Advocates of Sustainable AI successfully promote sustainable practices, resulting in widespread adoption of energy-efficient algorithms, hardware, and data center designs[36, 58]. A symbiotic cycle emerges between AI and the new energy system, where AI enhances system efficiency through renewable supply, demand-side electrification, and grid management, which in turn powers more sustainable AI development. This symbiosis positions AI as a solution to data center energy challenges rather than a problem[34, 60]. The emphasis shifts towards application-driven AI and traditional computing methods that require fewer resources. Sustainable AI evolves to balance efficiency, frugality, and sustainability, addressing climate change and human prosperity simultaneously[36, 62]. It optimizes infrastructure to support both environmental and societal goals. While Sustainable AI's electricity demand is projected to grow, this increase can be managed within planetary boundaries through renewable energy sources and improved efficiency[37, 40]. The key lies in meeting this increased demand through renewable and low-carbon energy sources, continual improvements in efficiency and responsible behaviors[35, 39]. This approach allows for the expansion of AI capabilities while staying within sustainable limits, aligning with the principles of Sustainable AI and contributing to a more environmentally responsible technological future[54, 56, 63]. This scenario is characterized by an efficiency balancing mechanism, where data center electricity consumption is effectively equilibrated. It is primarily driven by factors such as hardware and software efficiencies, frugal design and operation of Generative AI (Gen AI), and a symbiotic relationship between AI infrastructure and AI demand.

Scenario 2: Limits to Growth

In this scenario, shaped by demand dynamics analysts, AI capabilities expand but encounter natural or human-related limits. These constraints include, for instance, power availability, data scarcity, material and mineral shortages, computational resources, regulatory restrictions, competencies, and social questioning[35, 39, 51]. The result is a more restricted growth trajectory for AI development. The flow between demand and infrastructure is not fully optimized, creating discrepancies and disruptions in value chains[44, 55]. Technocratic control is being fueled by a combination of regulatory frameworks, economic growth and stagnation, cultural hype bubbles, environmental and social concerns, and defensive pressures[40, 52]. This approach attempts to align AI development with broader sustainability goals, but faces challenges in reconciling rapid technological advancement with resource limitations and societal concerns[61, 66, 67]. The scenario highlights the complex interplay between technological progress, environmental sustainability, and societal needs in shaping the future of AI[35, 65]. It features a constraint balancing mechanism, where the system dynamics are characterized by a constrained trajectory due to feedback loops that hinder AI's prosperous development.

Key drivers in this scenario include endogenous and exogenous constraints for both Gen AI training and inferencing. These constraints encompass local power availability, chip manufacturing capacity, data scarcity, network latency, the cost of training Large Language Models (LLMs), infrastructure limitations, and deployment challenges. Additionally, challenges in scaling Gen AI inferencing adoption, such as power limitations and the lack of proven Return on Investment (ROI), further limit this scenario.

Scenario 3: Abundance without boundaries

This scenario, supported by Techno-Efficiency Optimists embodies the Jevons Paradox, where improvements in AI efficiency paradoxically lead to increased overall energy consumption[37, 42, 64]. Technoutopian cheerleaders push for unlimited AI deployment across all sectors, believing that technological advancements will solve any resource constraints. Elevated to a totem, increased AI efficiency lowers computational costs, yet fuels uncontrolled rebounds in AI demand[45, 49]. This unbounded growth struggles to cope with planetary boundaries, resulting in increased concentration of power, land use appropriation, e-waste and battles for resources, especially water[43, 53]. While individual AI systems become more energy-efficient, the total energy consumption of the AI sector grows dramatically due to oversized AI supply[47], [57], and does not always serve human prosperity or climate protection[59]. The scenario system dynamics mechanism is characterized by a rebound effect, wherein the pursuit of AI performance intensifies planetary pressures, necessitating robust governance frameworks to mitigate hyperconcentration.

Scenario 4: AI Energy Crisis

In this scenario, Alarmists anticipate and react to potential "black swan" events in AI energy consumption. The rapid growth of AI leads to an unforeseen energy crisis, where AI's electricity demand begins to conflict with other critical sectors of the economy[46, 51]. This triggers a cascade of negative consequences, including economic downturns and severe operational challenges for AI-dependent industries[48, 55]. Regulators scramble to implement strict controls on AI development and deployment, while researchers grapple with a "data crunch" as they try to balance the need for massive datasets with energy constraints[50, 52]. This scenario highlights the potential risks of unchecked AI growth and the need for proactive risk management in AI development[63, 67]. It is characterized by a crunch reinforcement mechanism that ultimately leads to a crisis. This scenario is driven by a confluence of factors, including insufficient grid planning, inaccurate AI demand forecasting, uncoordinated AI governance, and the increasing reliance on computationally intensive techniques like synthetic data and multimodal learning. These factors can lead to underestimated future electricity consumption, fragmented policy responses, localized energy shortages, and intensified energy demands for AI training.

Methodology

Model design

To transform scenarios into quantifiable forecasts, we developed a system dynamics approach utilizing causal diagrams to feed behavioral simulations. This methodology allows us to define and analyze the intricate relationships between various factors within each scenario. The model architecture consists of three primary components, the details of which are provided in the appendices.

1. Core System Dynamics Model

At the heart of the forecasting system lies the core model, which incorporates the four fundamental schools of thought and related scenarios—Efficiency, Constraint, Rebound, and Crunch[66, 67].

2. Sub Models

To account for the specific characteristics of AI, five sub-models are integrated: Gen AI Training, Gen AI Inferencing, Traditional AI training, Traditional AI inferencing, Traditional non-AI Data Center. These models are dynamically fed by empirical and statistical databases, ensuring that the forecasts are grounded in real-world data. As the accuracy and reliability of AI forecasts are heavily dependent on the quality of the data used to train Gen AI models, for Gen AI Training and Inferencing, we propose a novel approach that leverages published data directly from leading LLM providers. This approach ensures that our forecasts are grounded in the latest trends in the industry[68, 69].

3. Envelopes

• *Endogenous envelope:* This envelope contains 6 macro factors and 30 microfactors that evolve 'within the data center', focusing on internal dynamics of AI data centers.

• *Exogenous envelope:* This envelope accounts for external influences 'outside the data center,' categorized into four main categories, following the IIASA insight: Energy and Material, Economy and Industry, Governance and Markets, and Society and Behavior. It is derived from 10 macro factors and 51 microfactors, providing a comprehensive view of external influences. Gen AI models have their own endogenous envelopes for training factors such as power availability, data scarcity, chip manufacturing, and network latency.

Model Scope and Methodological Framework

This research focuses on the quantitative assessment of electricity consumption in terawatt-hours (TWh), using "AI electricity use per year" as the functional unit for standardized analysis across sectors and regions. Our scope deliberately excludes greenhouse gas emissions from electricity generation or consumption, broader environmental impacts of electricity production or usage, cryptocurrency-related energy consumption, and edge computing applications. We hypothesize that renewable energy impacts AI's grid demand differently based on its source: off-site renewables contribute to overall decarbonization without directly reducing AI energy use, while on-site renewables decrease grid demand without necessarily lowering AI consumption. To accurately reflect the impact of on-site renewables on AI growth, our electricity consumption graphs present grid-supplied TWh.

Triangulation

This study employs a comprehensive triangulation approach to enhance the validity and reliability of its findings. The methodology incorporates four distinct triangulation methods:

• *Theory triangulation:* The research integrates structural model validation and theory integration by combining demand-driven (top-down) and supplier-driven (bottom-up) approaches to estimating energy needs. This integration is not merely for comparative purposes but is viewed as representing two complementary aspects of AI electricity consumption dynamics, providing a more holistic understanding. This approach aligns with the concept of theory triangulation as described by Denzin (1978) in his seminal work on triangulation in research[70].

• *Investigator triangulation:* The study leverages expert knowledge from various fields related to AI and data centers. These experts contribute to both structural causal model validation (e.g., assessing variable inclusion/exclusion and relational aspects) and behavioral model validation (e.g., evaluating outcome plausibility. This method is supported by the work of Patton (1999), who emphasized the importance of using multiple analysts to review findings[71].

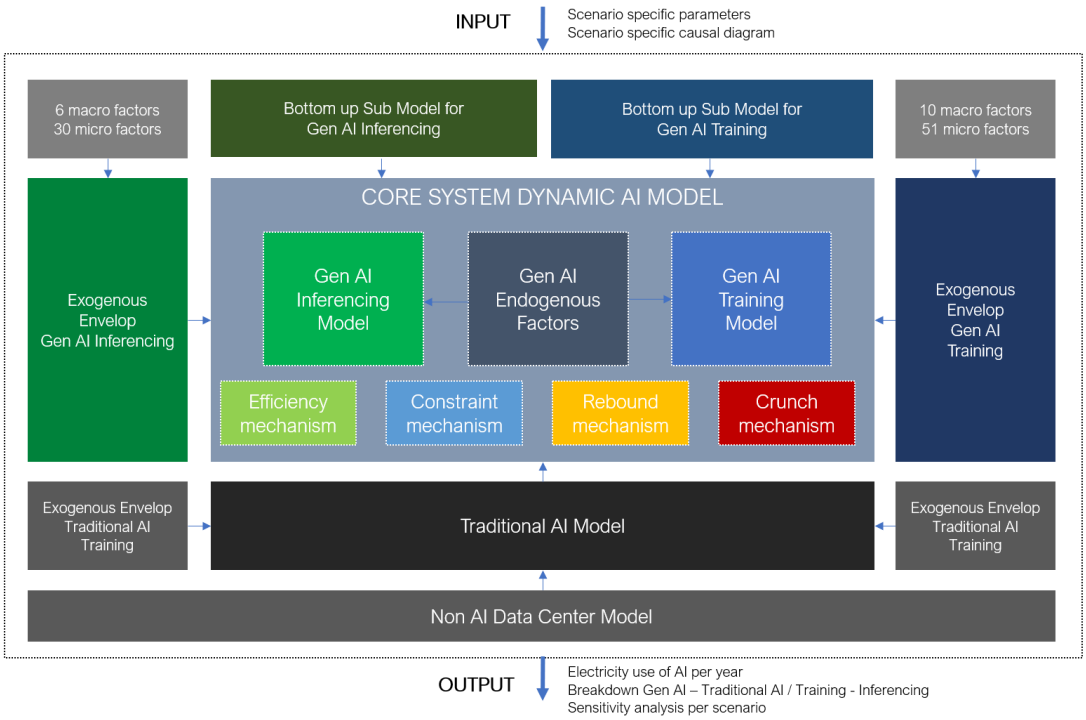
• *Data triangulation:* Recognizing the potential limitations of historical data in disruptive economic contexts, the study incorporates data as potential patterns rather than definitive truths. It relies on experience-based estimations from professionals, academics, independent researchers, and industries to supplement and contextualize quantitative data. This approach is consistent with the data triangulation method outlined by Denzin (1978) and further elaborated by Flick (2004)[70, 72].

• *Method triangulation:* The insights for modeling and parameter selection are grounded in both academic literature and professional publications. Given the scarcity of academic publications in this rapidly evolving field, the study employs a dual approach to literature review. In addition to a systematic literature search, which yields a limited but intensively used set of documents, a non-systematic literature search is conducted. This latter approach leverages the network of researchers at Schneider Electric and the company's academic and professional ecosystem[73], allowing for the incorporation of cutting-edge insights that may not yet be reflected in formal academic publications. This dual approach to literature review aligns with the concept of methodological triangulation as described by Morse (1991), which involves using multiple methods to study a single problem[74].

The use of these triangulation methods enhances the robustness of the research, as supported by the comprehensive review of triangulation in qualitative research by Carter et al. (2014)[75].

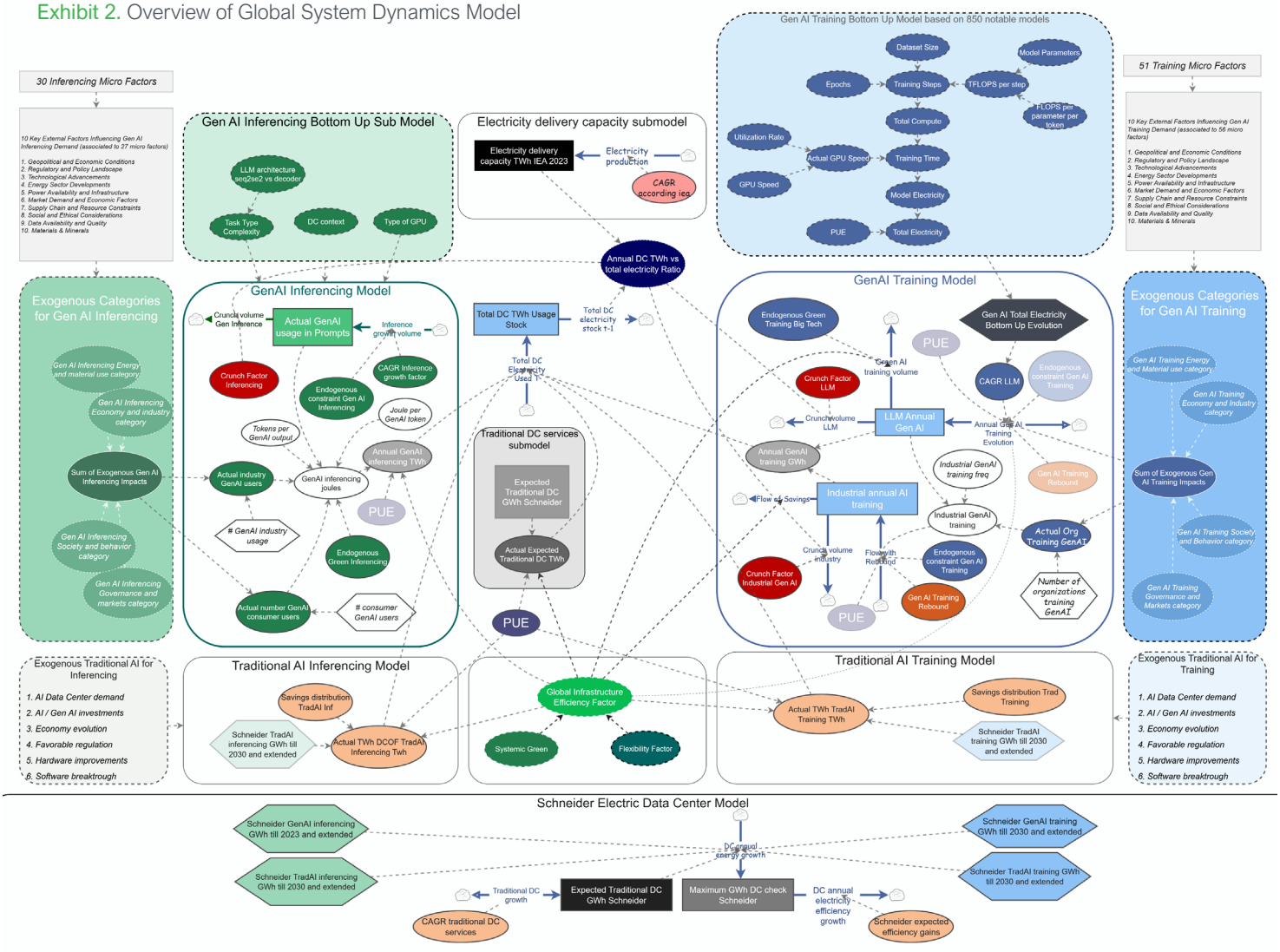
A detailed functional model and a snapshot of the global model are provided on the subsequent page.

Exhibit 1. System Dynamics Functionnal Model



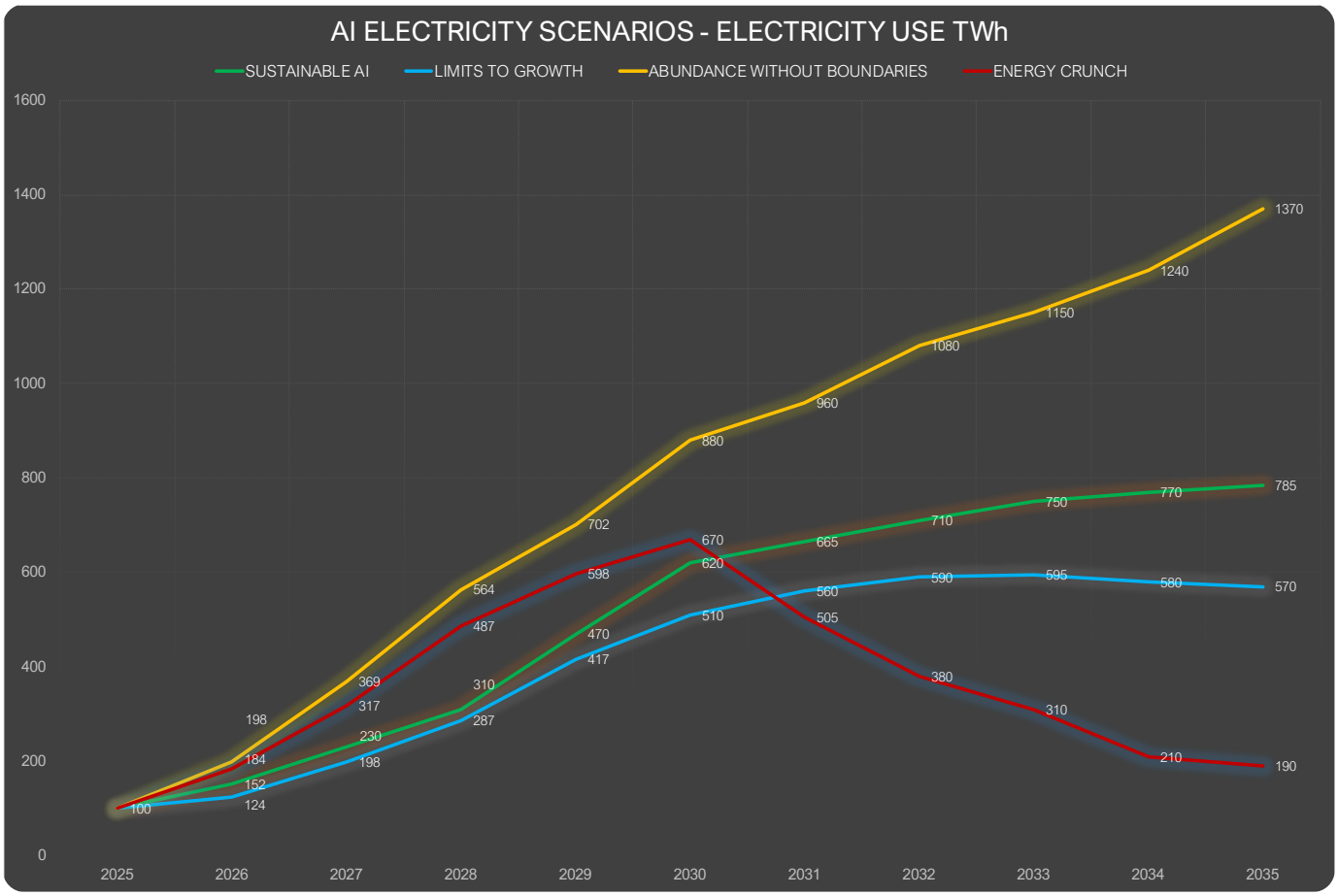
From our scenarios, we derived a functional model that was subsequently translated into a coherent system dynamics model, representing a snapshot of the global model that serves as the foundation for our simulations.

Exhibit 2. Overview of Global System Dynamics Model



Global Results (1/3)

Exhibit 3. Global AI electricity consumption forecasts from 2025 to 2035, in TWh



The decisions we make today about AI infrastructure will have lasting implications for future energy demand

The analysis of AI energy consumption scenarios from 2025 to 2035, as depicted in Exhibit 3, reveals distinct patterns of evolution. From 2025 to 2030, all scenarios initially show a general upward trend. However, their underlying dynamics diverge significantly. Around 2027-2028, as new AI-ready infrastructure is deployed, scenario-specific structural shifts become evident in electricity forecasts. This shift is driven by endogenous trends such as computational performance[76], dataset sizes[77], hardware trends, algorithmic efficiency[78], training cost evolution[79], and hardware acquisition costs[80]. It is also influenced by exogenous trends such as the evolution of global energy and material use, economic and industrial factors, governance and markets, and societal behaviors[81], as well as the reinforcing and balancing mechanisms within the overall system. By 2030, these differences become more pronounced. The **Sustainable AI** scenario exhibits significant growth, increasing from 100 TWh in 2025 to 620 TWh in 2030, a sixfold increase, and further to 785 TWh in 2035, indicative of a growing commitment to Sustainable AI principles[54, 56, 63]. A plateau emerges after 2029 as efficiency improvements begin to offset infrastructure expansion.

In contrast, the **Limits To Growth** scenario illustrates a constrained increase in electricity consumption, which may initially seem positive but masks a reality: an economy that fails to expand as desired, with consumption rising only from 510 TWh in 2030 to 570 TWh by 2035.

The **Energy Crisis** scenario presents a more extreme trajectory, peaking at 670 TWh before plummeting to 190 TWh by 2035. This indicates a potential global energy crisis or a series of localized disruptions due to rapid expansion without adequate resource planning, particularly in power grids and supply chains[82]. Moreover, the negative consequences of technologies that are prematurely forced into mainstream adoption but fail to align with societal needs could exacerbate these challenges[83]. Finally, the **Abundance Without Boundaries** scenario depicts continuous, unchecked growth, reaching a peak of 1,370 TWh by 2035. However, this scenario suffers from significant side effects, such as negative environmental impacts, unchecked development, and unequal access to AI benefits[84].

Global Insight 1: Reduced AI electricity consumption is not indicative of a sustainable and resilient development trajectory

The Limits To Growth scenario outlines a constrained trajectory for AI development, hindered by both endogenous and exogenous limitations. Energy consumption rises from 510 TWh in 2030 to 570 TWh by 2035. This scenario follows a threefold annual increase in Gen AI training compute and a 50% yearly expansion of language training datasets, reaching 25 trillion tokens by 2027[85]. Hardware performance improvements are lagging, with TFLOPs performance increasing by a modest 20-30% annually, while GFLOPs/Watt efficiency targets a 25-35% annual improvement, reflecting the constrained endogenous hypothesis of AI development embedded in the assumptions of this scenario[86].

Global Results (2/3)

Algorithmic efficiency in language models doubles yearly, but scaling remains suboptimal[87]. Gen AI inferencing faces similar challenges, struggling to achieve robust, organic demand development. In this constrained scenario, we might see a slower transition to newer formats. FP32 and FP16 might remain dominant, with limited adoption of more efficient formats like FP8 or specialized representations, potentially due to compatibility issues or development costs[88]. The full industrialization of traditional AI and Machine Learning is delayed, impacting their expected positive effects on supporting the decarbonization of the end sectors. At the Data Center level, efforts to mitigate limitations include the emergence of decentralized edge computing and a shift towards "small data" and transfer learning techniques[89]. Exogenously, resource scarcity, stringent regulatory frameworks limiting data center energy consumption, and public pressure lead to voluntary restrictions on AI energy use. The forecast evolution highlights the scenarios-specific system dynamics mechanism, characterized by a constrained and limited growth trajectory[90], with multiple feedback loops impeding AI's prosperous development.

Global Insight 2: As the demand for electricity grows, Sustainable AI is evolving to prioritize efficiency, frugality, and proven impact on end uses

The Sustainable AI scenario presents a trajectory of sustainable technological progress, balancing AI advancement with environmental stewardship. From 2025 to 2035, energy consumption increases steadily from 100 TWh to 620 TWh, reflecting a harmonious integration of AI into global energy systems[91]. In this scenario, by 2028, traditional AI and Machine Learning become highly industrialized, positively supporting decarbonization efforts. As efficiency gains become more pronounced, electricity consumption is continuously optimized. Detailed results indicate that inferencing capabilities will become dominant after 2028, signaling a shift towards more practical and widely beneficial AI applications. These efficiency improvements may stem from hardware advancements, algorithm optimization, and more energy-efficient AI models[92, 93]. As Large Language Model training compute grows by 2 times annually, and language datasets expand moderately by 1.2 times yearly to an optimal target of 20 trillion tokens, hardware performance is projected to improve significantly. TFLOPs performance is expected to increase by 1.7 times annually, while GFLOPs/Watt efficiency is projected to improve by 40-50% annually[94], aligning closely with the 50% CAGR observed in recent industry data for FP16/BF16 operations[95, 96]. Additionally, algorithmic efficiency in language models is expected to improve 4 times per year[85], enabling optimal scaling of generative AI training. The adoption of more efficient formats like BF16 and mixed precision techniques[97] could contribute to the plateau in energy consumption observed after 2029, complementing hardware and algorithmic improvements. This scenario envisions widespread adoption of FP8 for large-scale models, with BF16 becoming the standard for training[98], while edge devices predominantly utilize quantized 8-bit integer representations for inference[99, 100]. Our simulation results on traditional AI forecasts indicate that its industrialization by 2028 demonstrates the potential of existing AI methods to effectively address real-world problems and drive significant[101, 102], short-term progress across various sectors when scaled and applied broadly. Finally, efficient AI-specific hardware development and widespread adoption of AI-driven energy management systems in data centers drive internal progress[103].

Exogenously, global standards for Sustainable AI certification and regulatory landscape encourage industry-wide adoption of energy-efficient practices[8]. Companies have integrated sustainability impact assessments into their AI development, developed responsible AI policies, and set AI-carbon reduction targets[101]. They invest in energy-efficient AI algorithms, collaborate on open-source sustainable AI initiatives, and provide quantified evidence of AI's positive impact on climate mitigation and adaptation[8]. The system dynamics mechanism is characterized by a balanced control of AI and data center electricity consumptions, primarily driven by intentional actions for efficiency, frugality and virtuous impact on uses. This creates a reinforcing loop where efficiency gains in AI lead to reduced environmental impact, which in turn encourages further investment in sustainable AI practices.

Global Insight 3: Unrestricted abundance could fundamentally alter existing systems, constrain decarbonization trajectories and generate waste

The AI Abundance without boundaries scenario portrays a future of dramatic AI expansion characterized by groundbreaking technological advancements and voracious demand across high-tech sectors. Energy consumption surges from 880 TWh in 2030 to a staggering 1,370 TWh by 2035. Gen AI training compute for Large Language Models grows by 5-6 times annually, while language training dataset sizes expand by 3.5 times yearly, exceeding 50 trillion tokens[85]. In the Abundance without Boundaries scenario, computational performance improves dramatically, with overall performance increasing by 1.5 times annually[104]. This translates to a 50% annual improvement in TFLOPs performance. Simultaneously, energy efficiency sees remarkable gains, with GFLOPs/Watt efficiency improving by 50% annually. In the Abundance without Boundaries scenario, these remarkable improvements reflect the theme of unrestricted technological advancement and exponential growth in AI capabilities. Algorithmic efficiency in language models improves fourfold yearly with ambitious scaling[105], pushing the boundaries of computational efficiency[106]. This could include widespread use of mixed precision training with FP8, BF16, and even lower bit-width formats, alongside the emergence of AI-specific number systems optimized for energy efficiency at scale. However, the crisis might also spur rapid innovation in ultra-low precision formats and specialized hardware. Gen AI inferencing capabilities expand exponentially, penetrating a wide range of sectors, including aerospace, defense, biotech, healthcare, and finance[107]. Traditional AI experiences widespread adoption, further fueling the demand for computational resources. At the data center level, the development of room-temperature superconductors[108] and advances in bio-computing drive progress[109] as well. Exogenously, while breakthroughs in quantum computing by 2029[110] offer substantial potential to decarbonize energy systems, full development and large-scale deployment will take time. Similarly, recent announcements have sparked renewed interest in nuclear energy as a potential solution to future AI abundance. However, while nuclear power offers a carbon-free energy source, it faces significant challenges, including high initial investment costs, ranging from $3,000 to $6,200 per kilowatt[111, 112, 113, 114] and vulnerability to rising interest rates during lengthy construction periods. Consequently, in the AI Abundance without boundaries scenario, AI's rising electricity demand may still rely on fossil fuels in the interim[115, 116]. This reliance could result in continued environmental impacts[117] —including increased land use[118], water consumption[119], resource depletion, and waste.

Global Results (3/3)

These impacts may persist until quantum technologies are widely adopted and capable of significantly reducing carbon emissions, or until nuclear power can operate at scale[120]. The system dynamics mechanism is characterized by a reinforcing rebound loop of investment-driven growth and infrastructure development, leading to rebound cycles of increasingly powerful AI and expanded data center capacity. It pushes the boundaries of societal norms and environmental limits, raising concerns about power centralization, resource depletion, and governance.

Global Insight 4: Mismatched energy demand and infrastructure can lead to localized energy shortages and potentially trigger a global domino effect

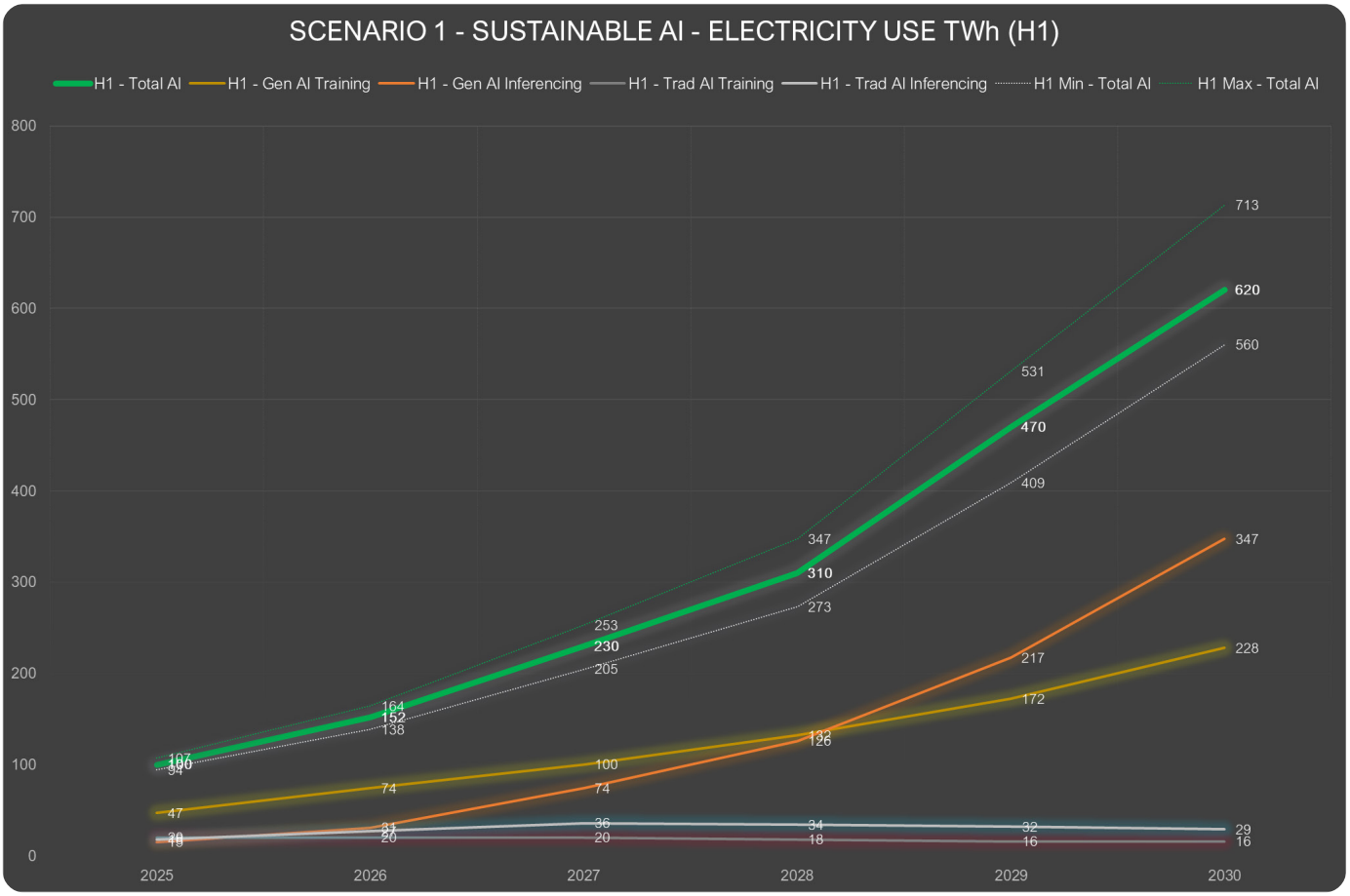
The Energy Crisis scenario depicts a tumultuous trajectory for AI development, characterized by rapid expansion followed by severe contraction. This scenario is already evident in some regions where debates emerge around the right balance between infrastructure and demand. Moreover, this highlights the fact that countries avoiding energy crunches due to AI are not only those creating additional infrastructures but also those enabling positive effects on their own energy systems' efficiency[121]. From 2025 to 2030, energy consumption surges from 100 TWh to a peak of approximately 670 TWh, driven by explosive growth in Gen AI training and inferencing, particularly in regions like the U.S., China, and the European Union. Rapid AI development outpaces infrastructure capacity, leading to localized energy crises between 2026 and 2029, as evidenced by preliminary signs already in 2024 in Ireland[122], Virginia (US)[123], and the Netherlands[124]. In response to these crises, industries and nations urgently seek to mitigate AI demand rebound effects[125]. However, as the reinforcing loop intensifies, even hasty AI efficiency measures fail to prevent the onset of the crunch[126], due to a lack of industry-wide coordination[127].

Paradoxically, the crisis may spur rapid innovation in ultra-low precision formats like FP8 or 4-bit quantization and specialized hardware as the industry scrambles for energy-efficient solutions. Traditional AI follows a similar pattern, with initial rapid adoption followed by a sharp decline as energy constraints take hold. Exogenously, increasingly erratic climate events[128], likely accelerating after 2028, may pressure power grids[129], potentially leading to local energy crises starting in 2029. These crises could result in strict rationing of electricity for AI data centers[130], and public backlash against AI's environmental impact might lead to restrictive legislation[131]. The rapid growth of AI and data centers could reach a critical point where it conflicts with other essential electricity functions in the economy, potentially triggering a series of negative consequences. The system dynamics mechanism is characterized by a crunch reinforcement loop that ultimately results in a crisis. Consequently, by 2035, energy consumption plummets to 190 TWh, highlighting the potential consequences of unchecked AI growth without adequate infrastructure planning. This scenario serves as a stark warning about the delicate balance between AI advancement and power constraints, emphasizing the risks of failing to maintain this equilibrium.

It's important to note that while traditional AI contributes modestly to overall energy consumption, it could play a crucial role in optimizing demand-side energy systems, which may significantly improve the efficiency of end uses and the broader economy. To gain a deeper understanding of AI's full impact, further system dynamics research will explore its indirect effects.

Sustainable AI Scenario (1/2)

Exhibit 4. Sustainable AI Scenario electricity consumption forecast from 2025 to 2035, in TWh



High-level factors driving Sustainable AI scenario

• *System Dynamics Mechanism:* Efficiency Balancing
• *Key Endogenous Factors:* TFLOPs performance increases by 70% annually. GFLOPs/Watt efficiency improves by 40-50% annually. BF16 for training, FP8 for large-scale models, 8-bit integer for edge device inference. Algorithmic efficiency in language models improves 4x per year. Large Language Model training compute grows 2x annually. Language datasets expand 1.2x yearly to 20 trillion tokens.
• *Key Exogenous Factors:* Traditional AI is expected to reach industrialization by 2028. Frugal design and operation of Gen AI. A symbiotic relationship between AI infrastructure and AI demand. Development of task-specific models for discriminative tasks.

General Analysis of Energy Consumption Trends (2025-2030)

From 2025 to 2030, the Sustainable AI scenario reveals a steady increase in energy consumption, reflecting the expanding demand for AI applications and the enhancement of AI capabilities. Annual generative AI training energy consumption rises from 47 TWh in 2025 to 228 TWh by 2030, a more than threefold increase which highlights the growing complexity and scale of generative models. Generative AI inferencing grows stronger starting in 2027 as deployment increases, from about 15 TWh to 310 TWh within the same period. This substantial increase underscores the growing importance of AI technologies and highlights the need for efficient energy management strategies to support this expansion sustainably[132]. Large language models (LLMs) and industrial generative AI training also show significant growth. Consumer LLMs training increases from 40 TWh in 2025 to 167 TWh by 2030, reflecting their expanding applications in sectors such as customer service and content generation. Industrial generative AI training rises from 7 TWh to 61 TWh, indicating the rapid integration of generative AI into industrial processes for enhanced automation and operational efficiency. Traditional AI electricity use maintains a relatively stable energy consumption pattern until 2028, then observes a slight decrease as it becomes more industrialized and as open-source and low-code model communities[133] gain momentum. This trend suggests a shift towards more efficient traditional AI models. Inferencing and training for traditional AI evolve from 18 TWh to 20 TWh in 2025 to 16 TWh and 29 TWh by 2030, respectively, showing modest growth in training requirements while inferencing remains stable.

Key Insight 1: Generative AI inferencing is emerging as the dominant electricity consumer

As shown in Exhibit 4, scenario projections indicate that Gen AI inference will become the primary driver of electricity consumption within the AI sector by 2027-2028. This trend could lead to electricity consumption exceeding 200 TWh within three years, highlighting the intensifying computational requirements of generative AI models. Despite this, the Sustainable AI scenario is evolving positively, largely due to three factors. First, the 2023 seminal work by Luccioni et al., "Power Hungry Processing: Watts Driving the Cost of AI Deployment?"[134], offers stronger evidence for both public perception and industry practices.

Sustainable AI Scenario (2/2)

Their comprehensive study of 88 diverse AI models revealed that generative AI models can consume significantly more energy for inference tasks than conventional algorithms, with some models requiring up to 30 times more energy[135] than traditional search engines for real-time processing. A key insight from this research - that using multi-purpose models for discriminative tasks can be more energy-intensive than employing task-specific models - is being seriously incorporated into industry sustainable AI roadmaps. This awareness is prompting a positive reevaluation of AI model deployment strategies, emphasizing the need for more energy-efficient approaches in AI application design. Second, significant hardware efficiency breakthroughs are materializing, offering substantial benefits to companies embracing Sustainable AI school of thought. Advancements in AI hardware and cooling technologies are driving significant improvements in performance and energy efficiency. The NVIDIA GB200 NVL72 system[139], utilizing multiple GB200 superchips, exemplifies this progress with substantial performance gains over previous-generation systems based on H100 GPUs[140]. System-level comparisons demonstrate up to a 30-times performance increase for LLM inference workloads[141], attributed to the new Blackwell architecture[142], improved cooling solutions, and system-level optimizations. At the chip level, the B200 GPU offers a 127% improvement over the H100, delivering 2,250 TFLOPS of FP16/BF16 compute compared to the H100's 989 TFLOPS[143]. These advancements are complemented by server rack densification[137] and widespread adoption of liquid cooling systems[138], supporting up to 132kW rack power density[136]. As awareness of AI's energy implications grows[136], these combined innovations are facilitating substantial chip evolutions and driving significant changes in datacenter design and efficiency. Hence, in the Sustainable AI scenario, while direct chip-to-chip comparisons may initially suggest performance gains, companies are prioritizing system-level optimizations through energy-efficient technologies to achieve a competitive advantage[144], maintain sustainability, and mitigate potential performance and cost-related challenges in the AI landscape. Third, the Sustainable AI scenario is characterized by a symbiotic relationship between AI infrastructure and demand, where efficiency and resource conservation are mutually reinforced. This synergy extends from user-centric applications to broader systemic impacts. Research from the 2020s is now translating into real operational gains. While AI inference may initially increase direct energy consumption, its potential to optimize energy usage across multiple sectors can lead to net positive energy outcomes. Applications in HVAC systems, microgrids, electric vehicle-to-grid integration, and waste management demonstrate this potential[145]. This symbiosis provides real-life effects, emphasizing the importance of considering AI's indirect effects on energy consumption alongside its direct impacts[146].

Key Insight 2: Traditional AI will continue to play a crucial role in decarbonization efforts across various end-use applications

As a key influencer of the efficiency balancing mechanism in the model, Traditional AI plays a crucial role in the Sustainable AI scenario. It is confirmed as an effective tool for decarbonization efforts across various sectors[147]. Its steady growth demonstrates that machine learning models such as decision trees, random forests, and support vector machines are more energy-efficient than generative models[148]. This provides a crucial counterbalance to the rising energy demands of generative AI. In this scenario, Traditional AI training energy consumption is projected to increase modestly from 38 TWh in 2025 to 45 TWh by 2030.

These models are currently deployed at scale across numerous industries, actively contributing to decarbonization efforts in ways that generative AI has yet to match in impact and widespread adoption. In this scenario, Traditional AI is widely used in energy grid optimization, smart building management, industrial process optimization, transportation and logistics, and precision agriculture, among other areas. These applications have been refined and optimized over years of deployment, making them highly effective in reducing carbon emissions across various sectors of the economy. As highlighted by Rolnick et al.'s comprehensive study[145], AI, particularly traditional models, holds significant potential for addressing climate change challenges across various domains. While the growth of generative AI inferencing is inevitable (projected to exceed 200 TWh by 2027-2028) and may offer new possibilities for decarbonization, traditional AI remains the primary tool for immediate and large-scale impact due to its efficiency, widespread adoption, and proven effectiveness. However, ongoing optimization and innovation in traditional AI models are crucial to keep pace with evolving demands, with the development of Small Language Models (SLM)[149] and Tiny AI[150] representing promising directions that combine the strengths of traditional AI and machine learning.

Key Insight 3: Resource-conscious generative AI training approaches intensify their focus on less energy-intensive models

Gen AI training strikes a delicate balance between technological and infrastructure progress and frugality. This balance is achieved through advancements in three key areas: infrastructure improvements[151], hardware and software innovations[151], and frugality strategies[153]. On the infrastructure side, multi-datacenter training may revolutionize Gen AI infrastructure[154], making it systemically more efficient and environmentally friendly. This approach to sustainable AI development encompasses several key aspects, integrating hardware, software, and infrastructure optimizations. It leverages distributed computational load[155] and network optimization[157] through load balancing and time zone optimization[158], while utilizing emerging memory technologies like STT-RAM and Memristors to enhance server performance and energy efficiency. Training strategies prioritize renewable energy utilization and adapt to seasonal variations, aligning with efforts to improve data center PUE[159]. Advanced cooling technologies[161, 162], exemplified by NVIDIA's Blackwell GB200 family[163], complement these efforts. Hardware efficiency[156] is boosted through flexible scaling across data centers[164], supported by Moore's Law[165] and projections of increasing transistor counts, with NVIDIA targeting over 200 billion transistors by 2024[166] and TSMC aiming for 1nm processes with over a trillion transistors by 2030. Advanced network technologies optimize communication, reducing latency for real-time processing[167]. On the software front, distributed learning algorithms[168], hyperparameter tuning[169], transfer learning[169], and AutoML[170] are employed to reduce model sizes and improve resource utilization. This system approach combines hardware advancements, energy-efficient infrastructure, and software optimizations to create a more sustainable AI ecosystem. Moreover, in this scenario, initiatives such as Luccioni's work using CodeCarbon to measure carbon footprints[171, 172] and Schwartz et al.'s proposal to emphasize computational cost reporting and prioritize efficient hardware and algorithms[174] are integrated. Frugality strategies, drawing on the first specification standard about frugal AI (AFNOR)[175, 176], offer actionable levers to reduce AI's impact.

Limits To Growth Scenario (1/2)

Exhibit 5. Limits To Growth Scenario electricity consumption forecast from 2025 to 2035, in TWh

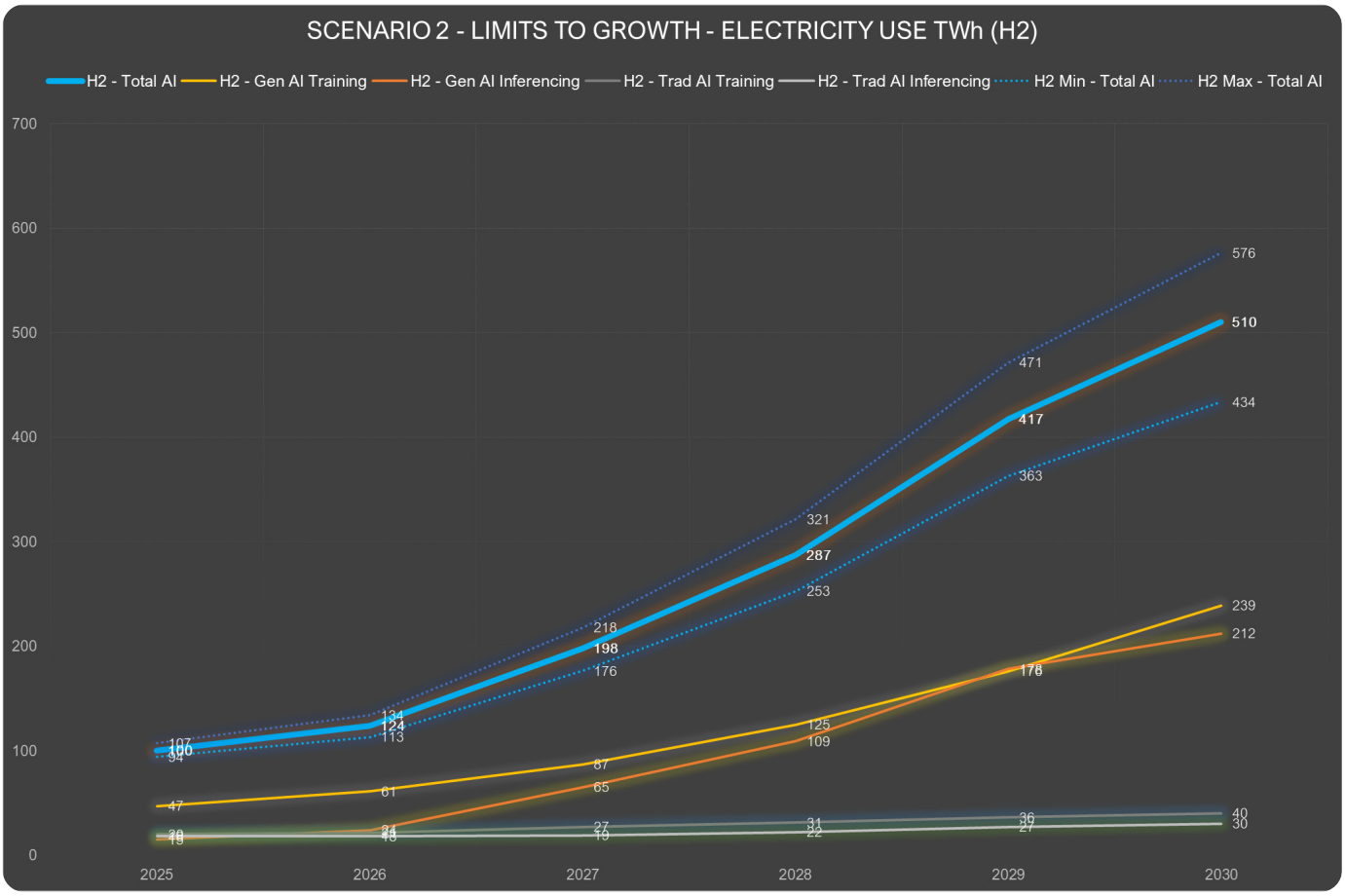


Figure 5. Limits To Growth Scenario electricity consumption forecast from 2025 to 2035, in TWh.

High-level factors driving Limits To Growth scenario

• *System Dynamics Mechanism:* Constraint Balancing
• *Key Endogenous Factors:* TFLOPs performance: Increase by 20-30% annually. GFLOPs/Watt efficiency: Improve by 25-35% annually. Formats: FP32 and FP16 remain dominant, limited adoption of more efficient formats. Algorithmic efficiency in language models: Double yearly, but with suboptimal scaling. Gen AI training compute: Increase threefold annually. Language training dataset sizes: Expand by 50% yearly, reaching 25 trillion tokens.
• *Key Exogenous Factors:* Training Constraints: Global power availability, chip manufacturing capacity, data scarcity, network latency, and cost of training LLMs. Inference Constraints: Local power availability, infrastructure limitations, and deployment challenges. Challenges in scaling AI adoption due to barriers and the lack of proven Return On Investment.

General Analysis of Energy Consumption Trends (2025-2030)

Traditional AI demonstrates divergent trends in energy consumption between 2025 and 2030. Training energy consumption doubles from 20 TWh to 40 TWh, indicating increased computational demands possibly due to more complex models or larger datasets. In contrast, inferencing energy consumption shows a more modest increase from 18 TWh to 30 TWh over the same period. This smaller growth in inferencing energy use suggests some efficiency gains in deployment and execution of AI models, potentially through improved hardware or optimized algorithms.

Overall, total AI energy use is projected to grow significantly from 100 TWh in 2025 to 510 TWh by 2030, underscoring several fundamental constraints in the AI industry. In the training phase, Gen AI faces challenges such as grid power availability in key data center hubs, manufacturing bottlenecks for specialized AI chips, and data scarcity for large language models. The inferencing phase, grapples with operational deployment issues, potential network latency problems, and the prospect of reduced consumer and industry adoption following the expected peak of hype around 2026.

Key Insight 4: Generative AI training is likely to face constraints due to limitations in power availability, chip manufacturing, data scarcity, and cost challenges

In the Limits To Growth scenario, Gen AI development faces a three-pronged challenge: limited power[177], chip shortages[178], and data constraints[179]. These factors may collectively limit AI advancement, making it increasingly difficult and concentrated among a select few industry players. The exponential growth in computational requirements for training large language models has led to unprecedented energy demands. Projections suggest that by 2030, data center campuses may require 1 to 5 GW[180] to support training runs of 1e28 (e stands for "times 10 to the power of") to 3e29 FLOP[181]. This represents a staggering increase from GPT-4's estimated 2e25 FLOP, underscoring the escalating power needs of advanced AI models. Such energy-intensive processes raise concerns about sustainability and the feasibility of continued AI model scaling.

Limits To Growth Scenario (2/2)

Chip manufacturing bottlenecks[182] further complicate this scenario, as the production of advanced AI chips is constrained by packaging and high-bandwidth memory capacities. While current estimates suggest a capacity for 100 million H100-equivalent GPUs, potentially supporting a 9e29 FLOP training run, projections vary widely[183]. Estimates range from 20 million to 400 million H100 equivalents, corresponding to 1e29 to 5e30 FLOP[184]. This uncertainty in chip production capabilities might add another layer of complexity to future AI development[185]. Also, in this scenario, data scarcity emerges as another significant hurdle. By 2030, available training data could range from 400 trillion to 20 quadrillion tokens, potentially enabling training runs of 6e28 to 2e32 FLOP[184]. This estimate factors in the projected 50% growth of the indexed web by 2030 and the potential tripling of available data through multimodal learning incorporating image, video, and audio inputs[186]. However, as models grow larger, finding high-quality, diverse data becomes increasingly challenging, potentially limiting further improvements in model performance[187, 188]. These constraints, combined with escalating training costs projected to exceed a billion dollars by 2027, may create substantial barriers to entry in the AI industry. As a result, AI development is likely to become concentrated among a few key players with the resources to overcome these challenges. This scenario, with its prohibitively high cost of generative AI training, means that very few companies will be capable of developing their own models[189, 190, 191]. Consequently, the AI landscape may evolve into an oligopoly, where a handful of tech giants and well-funded organizations can afford to push the boundaries of AI technology. It could lead to reduced diversity in AI development and applications, increased concentration of AI technologies, and a potential slowdown in AI innovation due to limited competition. However, this concentration of AI capacities might also shift the focus from pure scale to efficiency and optimization of existing models, as well as drive greater emphasis on specialized, task-specific AI models that require fewer computational resources[192].

Key Insight 5: Generative AI inferencing growth is susceptible to potential constraints from power and infrastructure

While our findings suggest that global generative AI inference could reach 212 TWh by 2030, its development is constrained by power availability and infrastructure limitations[193]. This scenario paints a picture of an AI landscape grappling with the challenges of scaling inference capabilities to meet growing demand[194]. The sheer volume of compute required for inference workloads presents significant hurdles, even though these workloads can be more distributed than training[195]. The evolution of AI inference is limited by aggregate capacity in various regions and the rapid advancement of AI models. These limitations are underscored by ClearML surveys[194], which revealed that 52% of organizations are actively exploring alternatives to GPUs for inference in 2024[196], and found that only 25% of organizations believe their GPU infrastructure achieves 85% utilization. Indeed, despite efforts to optimize GPU utilization, many data centers report underutilization during peak times[197]. In this scenario, AI infrastructure efficiency could become a critical bottleneck. Underutilized GPUs may exacerbate energy consumption, with Gen AI queries potentially consuming four to five times more power than typical internet searches[199]. Without significant efficiency breakthroughs, AI growth could be limited by energy availability and costs, potentially slowing adoption and development, and creating significant barriers to the widespread scaling of generative AI inference capabilities[200].

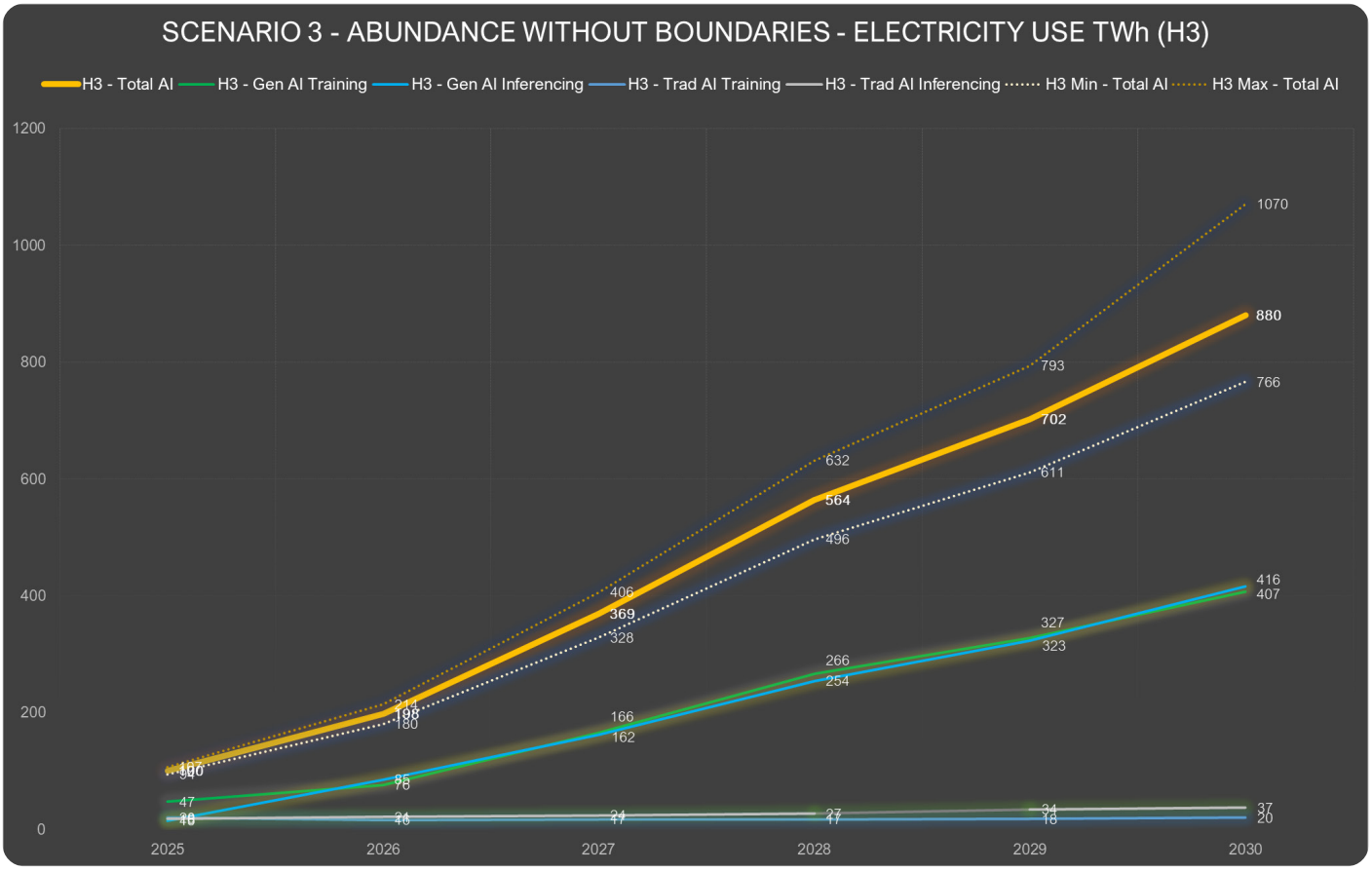
These challenges collectively contribute to a scenario where the growth of AI inference capabilities is constrained by physical and infrastructural limitations. Organizations are increasingly hindered by external constraints such as limited energy resources and infrastructure limitations. This may lead to a more reserved approach to AI deployment, with companies focusing on optimizing existing infrastructure and exploring more energy-efficient alternatives[201]. The Limits To Growth scenario also implies potential regional disparities in AI inference capabilities. Areas with more robust power infrastructure and cooler climates may have advantages in scaling their AI operations, potentially leading to geographical concentrations of AI inference capabilities.

Key Insight 6: Generative AI deployment may be constrained in scaling due to adoption barriers and lack of proven Return On Investment

In the context of the Limits To Growth scenario, the generative AI market is approaching a critical inflection point. Despite the initial rapid adoption of Gen AI, particularly on the consumer side, the technology is encountering systemic barriers reminiscent of resource limitations in traditional growth models[202]. The potential stagnation or decline in adoption rates reflects predictions of diminishing returns as generative AI technology reaches certain thresholds. Industrial companies face challenges in understanding and integrating generative AI into their performance and productivity processes. A Goldman Sachs report[203] notes that high costs associated with generative AI adoption may lead to diminishing returns if organizations cannot effectively integrate these solutions. This trend is further confirmed by Gartner's forecast[204] that 30% of generative AI projects might be abandoned by 2025 due to a lack of ROI, highlighting a growing emphasis on demonstrable value in an increasingly cautious market. After 2026, these trends may become structural, marking the official end of the generative AI hype. While generative AI holds immense promise, with the potential to contribute up to $4.4 trillion annually to the global economy[205], many enterprises are struggling to move beyond the experimentation phase and demonstrate clear returns on investment. The Limits To Growth scenario aligns with Gartner's 2024 prediction due to factors such as poor data quality, inadequate risk controls, escalating costs, or unclear business value[206]. As the industry potentially enters Gartner's "Trough of Disillusionment," companies are being forced to reevaluate their AI strategies, focusing on more targeted, value-driven applications. As described in the forecast results, generative AI's electricity use begins to plateau around 2029, stabilizing or even declining until 2035. This plateau may be contingent on the emergence of more efficient AI models or the next AI generation such as Meta's world model[207]. In this scenario, initial barriers like power availability and infrastructure limitations are early signs of a broader shift in the AI landscape. As these constraints become more apparent, the industry is moving away from inflated expectations and toward a focus on demonstrable ROI. This shift is forcing organizations to adopt a more measured approach, prioritizing efficient AI deployment over speculative investment. This aligns with the Limits To Growth scenario, where the initial optimism surrounding AI is being tempered by the realities of resource constraints and the need for practical, application-driven AI.

Abundance Scenario (1/2)

Exhibit 6. Abundance Without Boundaries Scenario electricity consumption forecast from 2025 to 2035, in TWh



High-level factors driving Abundance Without Boundaries scenario

• *System Dynamics Mechanism:* Rebound Reinforcement
• *Key Endogenous Factors:* TFLOPs performance: Improve by 50% annually. GFLOPs/Watt efficiency: Improve by 50% annually. Formats: Rapid development and adoption of novel number formats, including mixed precision training with FP8, BF16, and lower bit-width formats. Algorithmic efficiency in language models: Improve fourfold yearly with ambitious scaling. Gen AI training compute for Large Language Models: Grow by 5-6x annually. Language training dataset sizes: Expand by 3.5x yearly, exceeding 50 trillion tokens.
• *Key Exogenous Factors:* Planetary boundaries. Risks associated with oversized AI infrastructure. Insufficient AI governance. E-AI-Waste Dilemma.

General Analysis of Energy Consumption Trends (2025-2030)

In this scenario, total AI energy consumption is projected to rise substantially from 100 TWh in 2025 to 880 TWh by 2030, ultimately reaching a staggering 1,370 TWh in 2035. This reflects a robust expansion in AI capabilities and deployment, driven primarily by Generative AI. Gen AI training is expected to experience substantial growth, increasing from 47 TWh in 2025 to 407 TWh in 2030. This surge highlights the intensive computational demands of training advanced generative models, necessitating significant infrastructure and resource investments.

Energy consumption for generative AI inferencing is also set to grow, from 15 TWh in 2025 to 416 TWh by 2030. This indicates a broadening application of generative models across industries, as they move from research to practical deployment. Traditional AI training remains stable at 20 TWh over the period, while inferencing grows from 18 TWh to 37 TWh. These figures suggest incremental improvements and optimizations rather than major shifts in traditional AI practices. Taking a broader perspective, these electricity evolutions imply, directly or indirectly, potential consequences. These risks include the possibility of excessive infrastructure investment leading to stranded assets, the emergence of a trilemma of power concentration[208], supply chain vulnerabilities[209], and competition for critical resources, material and minerals[210], as well as a potential increase in AI-related electronic waste.

Key Insight 7: Oversized AI infrastructure is prone to risks of unsustainable operational costs and inefficient resource utilization

The AI Abundance without Boundaries scenario observes that the rapid and unrestrained development of AI systems can pose a risk of a constant race towards bigger and more powerful infrastructure, often outpacing the capacity for sustainable maintenance. A study by OpenAI estimated that the compute used in the largest AI training runs has been doubling every 3.4 months since 2012[211], far outpacing Moore's Law. This relentless growth raises concerns about the construction of increasingly massive data centers, many of which may risk becoming obsolete.

Abundance Scenario (2/2)

This is exemplified by Meta's recent decision to demolish an outdated data center design[212]. This issue, embedded in our modeling as a trigger for a Jevons Paradox risk, is confirmed by simulation results projecting a significant energy consumption increase from 100 TWh in 2025 to 880 TWh by 2030, which could accelerate the cycle of increased demand and resource depletion. This potential rebound is highlighted in recent nuclear hype announcements that indicate that, while nuclear energy might be a possible solution to support the electricity needs of AI Abundance without Boundaries, current debates often dismiss the fact that these projects require substantial initial investments, which are estimated to range from $3,000 to $6,200 per kilowatt for new plants in the United States[213]. At a discount rate of 10%, the median cost of nuclear energy can exceed that of natural gas and coal plants[213]. This is largely due to the capital-intensive nature of nuclear power, where capital costs account for at least 60% of the levelized cost of electricity (LCOE). Moreover, a risk remains that high interest rates can significantly inflate these costs, particularly over lengthy construction periods during which no revenue is generated, leading to compounded interest expenses that can jeopardize project viability. In contrast, countries like China and India have managed to achieve more competitive nuclear economics through ongoing construction experience and lower labor costs[213]. This scenario highlights the risk of building a Gen AI-standard asset legacy[214] (which may be outdated with 2030 models such as World Model, potentially trapping Gen AI) due to the rapid buildup of oversized infrastructure. This infrastructure, lacking long-term design or refurbishment considerations, may become unsustainable and burden future generations.

Key Insight 8: Insufficient governance and concentrated power could exacerbate AI access inequality

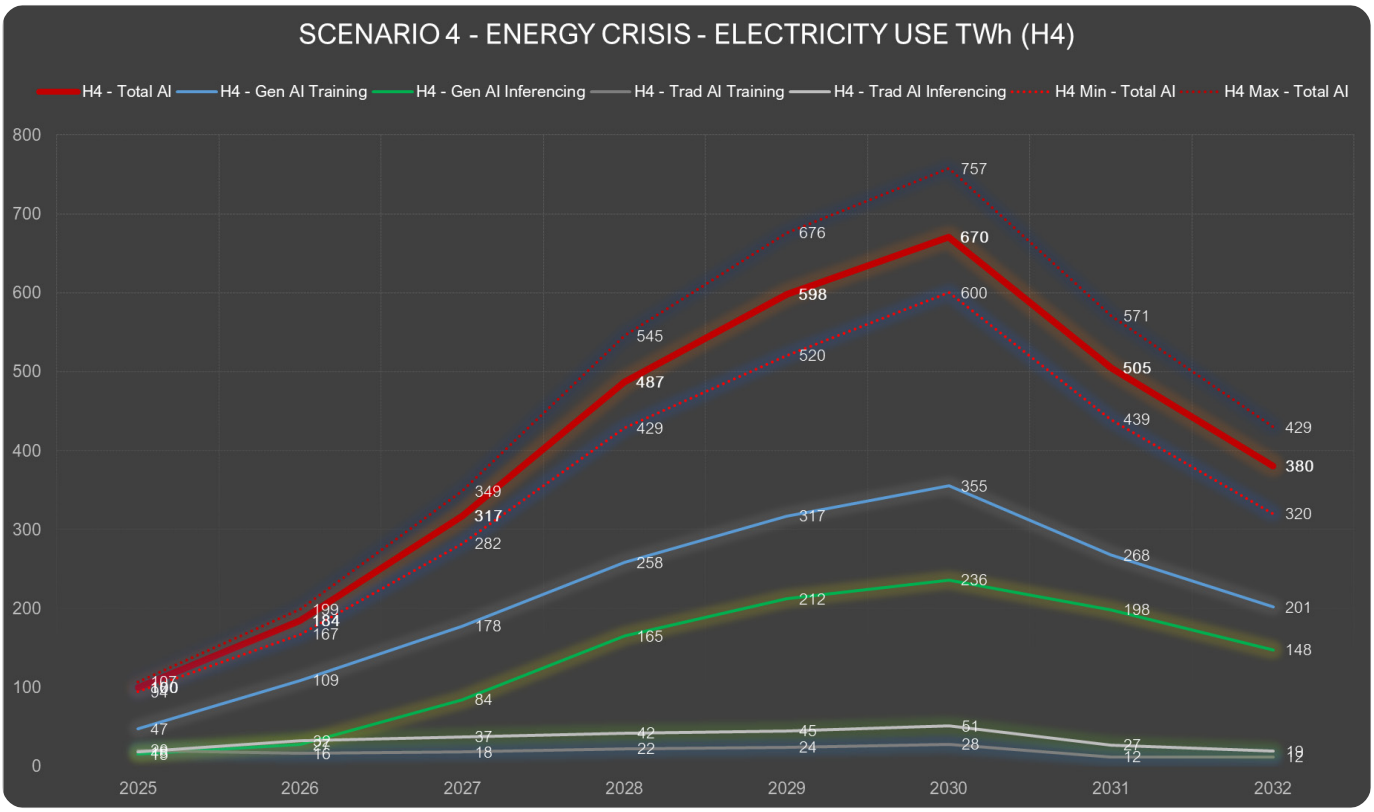
This scenario, driven primarily by extreme demand growth and exogenous factors like concentrated power and access to resources, could lead to increased AI-access inequality and requires robust governance structures to mitigate these risks. Over time, the concentration of power could transform a small number of tech giants and nations from pioneers into entrenched gatekeepers of the AI landscape. In 2024, the top tech companies - Apple, Microsoft, NVIDIA, Alphabet, Amazon, and Meta - expressed concern as their combined market value exceeded the GDP of most countries, reaching a staggering $15.2 trillion, surpassing the GDP of any single country except for the United States and China[215]. Such economic power, particularly in the semiconductor industry—with key players like TSMC (market capitalization (market cap of $1.07 trillion on October 17, 2024[216]), Broadcom (market cap of $857.70 billion as of November 8, 2024)[217], and Samsung (market cap of $276 billion USD as of November 2024)[218] consolidating their power in advanced chip manufacturing may lead to potential issues over time, including reduced competition, higher prices, and uneven innovation. A potential risk to address in maintaining a resource-symbiotic AI development is the supply of critical minerals. China's dominance in producing 68% of the world's rare earth minerals[219] and 77% of all graphite production and 97% of global anode output[220] could intensify the trilemma of power concentration, supply chain pressure, and competition for these critical resources. This trilemma could widen the gap in AI access and capabilities, where organizations or nations with greater resources benefit disproportionate access to AI development[221]. This disparity is evident in the distribution of generative AI resources, with Meta deploying at least 16,000 A100 GPUs in its Research Super Cluster[222], Tesla claiming 7,360 A100 GPUs[223], and the EU's Leonardo supercomputer, utilizing 13,824 A100 GPUs[224]. This concentration of computational resources in a few large corporations and national facilities highlights the potential for widening gaps in AI capabilities. Companies or states with access to cutting-edge AI tools might experience productivity gains of 30-40% in specific knowledge work tasks[225], while those without such access struggle to compete, creating a new form of digital divide[226] that exacerbates economic and social disparities on a global scale[227]. Consequently, this scenario raises the importance of robust governance structures to manage potential unchecked development that could further entrench these inequalities. Importantly, in this environment of unbounded AI abundance, certain key tipping points could trigger shifts between scenarios. These triggers, which involve feedback loops that confront practical boundaries -societal, physical, material, environmental, etc. - can introduce constraints that ultimately lead to the Limits to Growth scenario or, more extremely, the Energy Crisis scenario.

Key Insight 9: Unrestrained AI development can worsen the e-AI-Waste Dilemma by prioritizing performance over practicality

In the AI Abundance Without Boundaries scenario, the rapid development of Generative AI, if not effectively managed, might pose a significant risk to the global e-waste crisis. By 2030, Generative AI could potentially generate between 1.2 million and 5 million metric tons of e-waste annually[228], primarily from high-performance computing hardware in data centers, including servers, GPUs, CPUs, memory modules, and storage devices[229]. This could represent a potential 1000-fold increase from current AI-related e-waste levels, exacerbating the existing global e-waste problem, which already exceeds 60 million metric tons annually[228]. The primary contributors to this potential surge could be the short lifespans of hardware - typically two to five years - and the rapid turnover of technology as companies might strive to keep up with advancements in AI capabilities[230]. Within this scenario, as the compute requirements for training large language models (LLMs) could increase by 5-6 times annually, with language training datasets potentially expanding 3.5 times yearly[85], a risk might occur of a relentless waste of discarded Gen AI-specific electronics. These electronics often contain hazardous materials such as lead, mercury, and chromium[231], which could pose serious environmental and health risks if not disposed of properly. In this scenario, while circular economy strategies - such as extending hardware lifespan and recycling components - might reduce e-waste generation by up to 86%[230], challenges would likely persist. While we didn't quantify in our modeling the potential e-waste output as part of indirect effects, we consider that pure AI efficiency will not naturally save AI waste, which is a specific topic to address, notably by designing AI-ready equipment for refurbishment.

Energy Crisis Scenario (1/2)

Exhibit 7. Energy Crisis Scenario electricity consumption forecast from 2025 to 2035, in TWh



High-level factors driving the Energy Crisis scenario

• *System Dynamics Mechanism:* Crunch Reinforcement
• *Key Endogenous Factors:* Same as Abundance without limits, with rushed adoption of lower precision formats like FP8 or 4-bit quantization.
• *Key Exogenous Factors:* Insufficient grid planning and inaccurate AI demand projections. Uncoordinated AI governance. Synthetic data and multimodal learning.

General Analysis of Energy Consumption Trends (2025-2030)

The Energy Crisis scenario graph illustrates a peak in AI energy consumption around 2029, reaching approximately 670 TWh, followed by a decline to about 380 TWh by 2032 and a further reduction to 190 TWh in 2035. This indicates potential constraints such as resource limitations or economic pressures impacting growth. Generative AI training and inferencing show significant increases, peaking alongside total AI consumption, which reflects the intensive demands of advanced model development and application. However, the subsequent decline suggests challenges in sustaining such growth, possibly due to energy availability or cost issues. Traditional AI training and inferencing remain relatively stable, with minor fluctuations, indicating consistent but limited growth focused on incremental improvements rather than major shifts. The scenario underscores the need to manage the risks of unsustainable practices and inefficient resource management. Hence, policymakers and industry leaders must strategically plan to balance growth with environmental and economic sustainability, investing in energy-efficient technologies and infrastructure to mitigate the risks associated with an energy crunch while continuing to advance AI capabilities responsibly.

Key Insight 10: Insufficient grid planning and inaccurate AI demand projections are likely to underestimate future electricity needs

In this scenario, very high exogenous constraints are put on local grid capacities as the compound annual growth rate (CAGR) of data center energy consumption might reach staggering levels, potentially surpassing 25% in some regions[232]. A typical example is Ireland's EirGrid, which had initially projected AI to boost data center CAGR from 10.6% to 16.7%[232], but might find itself grappling with growth rates exceeding 20% by 2028. EirGrid's projected AI-driven demand could exceed the reliable grid limit as early as 2026, significantly earlier than anticipated without AI growth. This surge in demand could outpace even the most aggressive grid capacity planning, potentially leading to severe infrastructure shortfalls. While this scenario is fundamentally the sum of potential local risks, we can anticipate a limitation after a certain time driven by the inadequacy of grid planning, which might become glaringly evident across several key markets[234]. A relevant example is the situation in Dominion's grid in the US[235], which might be even more dire, where relaxed reliability standards for new data centers could fail to meet the skyrocketing demand. By 2029, new facilities might face power availability as low as 85%, potentially impacting their operations and reliability. In this scenario, the mismatch between AI-driven demand and grid capacity might be further exacerbated by the disparity in development timelines. In a scenario where mitigation strategies for AI power demands fail to materialize, the energy landscape for AI training could face severe challenges by 2030[236].

Energy Crisis Scenario (2/2)

Without geographically distributed training, AI workloads could remain concentrated in specific regions like Northern Virginia[237], where data center power capacity is projected to grow from 5 GW to 10 GW by 2030. Concentration could intensify competition for limited grid capacity, potentially causing power shortages and service disruptions. The absence of on-site generation projects, such as Meta's 350MW and 300MW solar farms[238] or Amazon's 960 MW nuclear power contract[239], would leave AI companies heavily reliant on strained local grids[240]. The lack of significant grid upgrades would mean existing infrastructure would struggle to meet the projected 6 GW power demand for frontier AI training runs by 2030, a 200-fold increase from current levels[236]. Without substantial improvements in energy efficiency beyond the expected 24x increase, power consumption for AI training would grow exponentially, potentially necessitating power rationing or rolling blackouts[241]. This scenario could severely constrain AI development, with companies potentially forced to relocate, scale back training runs, or face extended downtime, ultimately slowing the pace of AI advancement and limiting access to high-performance AI capabilities.

Key Insight 11: Uncoordinated AI governance can lead to fragmented policies and localized energy crunches

In the Energy Crisis scenario, the period from 2025 to 2028 initially resembles an "AI abundance without boundaries" hypothesis, characterized by rapid data center construction often completed within 8-12 months[242]. However, uncoordinated decision-making in AI governance exacerbates the system dynamic crunch mechanism, which collides with power infrastructure limitations, creating a severe mismatch between AI energy demands and grid capacity. This lag between data center expansion and grid capacity growth creates a persistent capacity gap, potentially leading to grid instability and localized blackouts in high-demand areas. Signals from 2024, with countries like the United States[243], Ireland[244], and the Netherlands[245] already experiencing tensions due to high data center energy demands, are confirmed over time. The scale of this challenge becomes even more apparent when considering current and projected AI power demands. For instance, the recent Llama 3.1 405B model, with its 4e25 FLOP training run, required 27MW of total installed capacity using a cluster of 16,000 H100 GPUs[246]. By 2030, LLM training runs are projected to be 5,000x larger, reaching 2e29 FLOP and potentially requiring up to 6 GW of power - a 200-fold increase from current levels[246]. The lag between data center expansion and grid capacity growth, typically requiring 3-4 years for new power additions[247], creates a persistent capacity gap. The 2024 example of Northern Virginia, which currently houses nearly 300 data centers connected to 5 GW of power in peak capacity, is projected to grow to around 10 GW by 2030[248]. In the Energy Crunch scenario, many other nations - such as Germany, Japan, and parts of Southeast Asia[249] (which could face similar challenges before 2030. Without coordinated decision-making, short-term solutions may prevail, resulting in extreme measures such as introducing AI usage quotas. In addition to grid constraints, economic pressures like rising electricity prices and potential carbon taxes could further strain AI operations[250]. Companies attempt geographically distributed generative training, allowing for training runs of 2e28 to 2e30 FLOP, but fail to address local constraints[85]. Projected advancements in inter-data center bandwidths, expected to reach 4 to 20 Petabits per second (Pbps) by 2030 and capable of supporting training runs of 3e29 to 2e31 FLOP[85], are not achieved.

Under these conditions, generative AI is likely to become economically unsustainable. Governments might respond by imposing restrictions on high-energy-consuming activities, such as mandating on-site renewable energy generation or limiting the scale of generative AI models[251]. These regulatory responses would likely force companies to either scale back their AI operations or invest heavily in more energy-efficient hardware and infrastructure. The post-2029 decline in AI-related energy use, as shown in the graph, may be a direct result of such economic and regulatory pressures. Future research should focus on developing territorial scenarios using the Energy Crisis insights, exploring how different regions and countries might be affected and respond to these challenges. Additionally, country-level analyses should be linked to the energy crunch issue, providing a more granular understanding.

Key Insight 12: Synthetic data and multimodal learning are likely to intensify local energy crunches for AI training

In the Energy Crisis scenario, data scarcity might start to put pressure on electricity use between 2027 and 2030 as AI models continue to grow in size and complexity. While scaling has been a key driver of AI progress, limits are emerging due to the finite stock of human-generated public text data, estimated at around 300 trillion tokens[246]. As models are trained on increasingly enormous datasets, leading to exponential growth in training compute and performance improvements, projections suggest that language models will fully utilize this stock between 2026 and 2032, or even earlier if intensely overtrained. Consequently, this data scarcity is exacerbating the energy crunch factor in AI development. In this scenario, the current data paradigm based on public human text data will not be able to continue a decade from now[252]. Hence, it is likely that alternative sources of data will be adopted before then, allowing AI systems to continue scaling, but requiring additional electricity not necessarily planned in the grid evolution. To maintain progress beyond 2030, innovations in three key areas will be crucial: synthetic data generation[253, 254], learning from other data modalities[255], and data efficiency improvements[256]. However, these solutions may come with their own energy costs. Synthetic data generation, while promising, could be computationally intensive and potentially double the energy requirements[257, 258]. The shift towards multimodal data processing (including images, video, and audio) may demand more energy-intensive processing compared to text-only data[259]. Overtraining models to improve inference efficiency might lead to increased energy consumption during training[260]. Moreover, as data becomes scarce, a trade-off between data efficiency and energy efficiency may emerge, with more data-efficient models potentially requiring greater computational resources[261]. Lastly, the geographical implications of these trends could lead to a concentration of AI development in areas with access to both vast amounts of data and energy, potentially exacerbating local energy challenges in these regions[262].

Conclusion

By employing a system dynamics approach to model four distinct scenarios - Sustainable AI, Limits To Growth, Abundance Without Boundaries, and Energy Crisis - this research offers open perspectives on possible trajectories of AI electricity use, which is fundamental as current AI infrastructure decisions will significantly shape electricity consumption well beyond 2030. While all scenarios show initial growth, their trajectories diverge after 2030 due to a combination of internal and external factors.

Importantly, reduced AI electricity consumption doesn't necessarily indicate sustainable development. The **Limits To Growth** scenario illustrates this, showing a minimal increase in consumption (510 TWh in 2030 to 570 TWh in 2035), but masking stunted economic expansion. The **Sustainable AI** scenario emerges as a promising approach, prioritizing efficiency, frugality and proven impact while steadily increasing energy consumption (100 TWh in 2025 to 785 TWh in 2035). This balances technological advancement with environmental stewardship, potentially positioning AI as a solution to energy challenges. In contrast, the **Abundance Without Boundaries** scenario highlights the potential risks of unchecked growth, projecting energy consumption to reach 1,370 TWh by 2035, which could lead to challenges like increased centralization of power and inequitable AI access. The **Energy Crisis** scenario highlights risks of mismatched energy demand and infrastructure, potentially leading to widespread energy shortages. It shows consumption peaking at 670 TWh before plummeting to 190 TWh by 2035, indicating possible global or localized energy crises.

Moreover, the four **Schools of Thought** are bringing scenario-specific insights. The Sustainable AI scenario foresees generative AI inferencing emerging as the dominant electricity consumer, while traditional AI continues to play a crucial role in advancing decarbonization efforts across various applications. Resource-conscious AI training approaches are accelerating their focus on less energy-intensive models, balancing performance and sustainability. However, the Limits To Growth scenario highlights potential constraints in AI development, including power availability, chip manufacturing, data scarcity, and adoption barriers that may hinder scaling and deployment. The Abundance Without Boundaries scenario warns of risks associated with unchecked AI growth, including unsustainable infrastructure costs, exacerbated global inequality, and increased e-waste due to prioritizing performance over practicality. The Energy Crisis scenario underscores the likelihood of underestimated future electricity needs due to insufficient grid planning and inaccurate demand projections. Uncoordinated AI governance may lead to fragmented policies and localized energy crunches, while synthetic data and multimodal learning are expected to intensify local energy demands for AI training.

When comparing our projections to existing benchmarks, we encountered challenges in obtaining a clear picture of electricity consumption forecasts in TWh. This discipline is still nascent, making direct comparisons difficult. Despite these challenges, we observe some convergence in the projections. Our AI Abundance Without Boundaries scenario, with its 880 TWh projection for 2030, seems to align with SemiAnalysis' estimates[263]. Conversely, Wells Fargo's projection of 652 TWh by 2030[264] looks similar to our Sustainable AI scenario for the same year, suggesting it aligns with a balanced growth trajectory.

It is crucial to note that these scenarios can be applied at different country levels, leading to a more nuanced global landscape. Rather than a single, uniform global scenario, we are likely to see a mosaic of different scenarios playing out across various countries and regions. This is influenced by a complex set of factors, including global trends, local conditions, policies, and decisions.

Furthermore, these scenarios should not be viewed as static or mutually exclusive constructs. Nations may transition between scenarios over time, or concurrently exhibit characteristics of multiple scenarios. This phenomenon is exemplified by the United States, where significant regional and socioeconomic disparities in energy consumption patterns are observed. Hence, as an example, a country initially facing an energy crunch might implement strategic decisions that reorient its AI electricity trajectory towards a more balanced development path.

Stepping back, we understand that a key element in shaping the future of AI's energy consumption is to avoid the negative aspects of scenarios (3) **Abundance without Boundaries** and (4) **Energy Crisis**, while enabling scenario (1) **Sustainable AI** by steering away from of scenario (2) **Limits To Growth.** As the Sustainable AI scenario appears to be the most balanced approach, it is crucial to consider the actions and decisions necessary to align with this vision.

To this end, we employed a normative scenario approach to identify the key factors that could make the Sustainable AI scenario a reality. Building on these insights, the next section presents a set of recommendations for decision-makers and policymakers to guide the development of AI electricity use towards a sustainable and resilient future.

Recommendations Towards Sustainable AI

As AI's future remains unwritten, it is crucial to guide its trajectory towards a sustainable future. To ensure AI development aligns with human prosperity and planetary boundaries, we suggest a set of nine guiding principles to position AI on a sustainable path and prevent its mutation into harmful scenarios.

AI Infrastructure And Deployment

1. Build and optimize next-generation AI data centers

We recommend implementing cutting-edge cooling technologies, high-density computing solutions, and investing in modern energy-efficient AI-specific hardware like GPUs and TPUs. Regularly assess and upgrade infrastructure, while collecting and sharing cooling system data, including site-specific technologies, temperature set points, and performance metrics. Policymakers should create incentives for companies adopting these technologies, while businesses should set ambitious targets for reducing Power Usage Effectiveness (PUE) in data centers, aiming for industry-leading standards below 1.2.

2. Expand and integrate renewable energy sources and advanced storage solutions

We suggest accelerating the deployment of on-site renewable energy generation combined with advanced energy storage solutions to ensure a stable power supply and mitigate the intermittency of renewable sources. Implement smart energy management systems that dynamically adjust energy usage based on availability and demand, and invest in cutting-edge storage technologies such as solid-state batteries or hydrogen storage or even Power Purchase Agreements (PPA).

3. Plan and implement strategic grid capacity enhancements

We recommend proactive grid capacity planning to anticipate the growing energy demands of AI. This involves collaboration between energy providers, policymakers, and AI companies to align on comprehensive strategies. Incorporating demand response strategies allows for adjusting energy usage based on grid conditions and renewable energy availability.

AI Development And Circularity

4. Optimize software efficiency and refine AI model performance

We promote techniques such as model pruning, quantization, and the use of lightweight architectures that can significantly improve efficiency. Encouraging research and development in low-power AI models not only contributes to energy savings but also makes AI more accessible by reducing hardware demands. Developing measured AI hardware power profiles is essential, including up-to-date data on actual power use and performance of different AI server configurations under various workloads. Additionally, addressing the energy implications of model versioning and updates through AI Model Lifecycle is crucial for long-term sustainability.

5. Quantify, assess, and prioritize AI impact on sustainability

We recommend implementing evidence-based AI impact quantification methodologies for understanding the direct and indirect effects of AI usage. This involves collaborating with academic institutions and industry to operationalize scientifically grounded frameworks for quantifying AI's impact, leveraging consequential approaches. AI companies should establish clear Key Performance Indicators (KPIs) for AI projects that include energy efficiency and environmental impact alongside business outcomes.

6. Implement circular economy practices for AI hardware and software

We suggest to implement circular economy principles in the lifecycle of AI hardware and software to minimize negative impact. Design hardware for longevity, ease of repair, and recyclability. Design software for modularity, updatability, and cross-platform compatibility. Establish take-back programs for obsolete equipment, ensuring responsible recycling and reuse of materials. Source components from suppliers committed to sustainable practices. Implement AI Design for Disassembly (AIDfD) principles to facilitate easier recycling and reuse of components. Develop AI-powered systems for predictive maintenance to extend hardware lifespan and reduce waste.

Governance, Standards, and Education

7. Develop and enforce Sustainable AI certification standards

We promote creating and enforcing standardized Sustainable AI certifications as a key step towards sustainable AI practices. Policymakers should develop certification schemes with clear, measurable criteria for energy efficiency and environmental impact. Companies should align their organizational goals with these certification requirements and commit to achieving them within specific timeframes. Regular audits and progress reports towards certification standards ensure accountability and drive continuous improvement. Developing tiered certification levels recognizes varying degrees of sustainability achievements and encourages ongoing advancements in Sustainable AI practices.

8. Establish and maintain robust AI governance frameworks

We suggest developing a comprehensive framework to guide responsible AI development and deployment, addressing energy consumption, data privacy, and ethical considerations. This approach could include a risk-based classification system for AI applications, encouraging regular audits, and proposing industry-wide energy consumption benchmarks. We recommend emphasizing robust data protection practices, considering the establishment of an AI ethics advisory board, and promoting transparency in AI decision-making processes. To foster fair competition and innovation, we propose exploring interoperability guidelines, considering increased public funding for smaller players and academic institutions, and examining ways to prevent market concentration.

9. Enhance AI skills and promote digital literacy

We promote developing comprehensive AI education programs that emphasize sustainable practices as crucial for building a workforce equipped to address the challenges of Sustainable AI. Policymakers should allocate funding for AI curricula in universities and vocational institutions that integrate sustainability principles. Companies should establish partnerships with educational institutions to create training programs that combine technical AI skills with environmental awareness. Setting targets for increasing the number of employees with Sustainable AI certifications promotes a culture of continuous learning in sustainable AI practices.

Future Research

While this study provides insights into potential AI electricity consumption scenarios, it also highlights several areas that warrant further investigation. Future research should aim to address the limitations of the current study and expand our understanding of the complex interplay between AI development, energy systems, and broader societal and environmental factors.

The study of lifecycle sustainable greenhouse gas emissions is crucial for understanding the full environmental impact of AI systems

One significant limitation of the current study is its focus on direct electricity consumption without considering lifecycle Sustainable-house gas emissions. Future research should aim to integrate a comprehensive lifecycle assessment of AI systems, including the environmental impact of hardware manufacturing, data center construction, and end-of-life disposal. This would provide a more holistic understanding of AI's environmental footprint and help identify opportunities for sustainability improvements across the entire AI lifecycle. Such research could involve collaborations between computer scientists, environmental engineers, and lifecycle assessment experts to develop standardized methodologies for assessing AI's full environmental impact.

Improving models to capture the dynamic nature of AI demand is essential for accurate forecasting

The system dynamics model in this study primarily focuses on direct effects, with limitations in integrating the variability of changing demand patterns. Future research should aim to develop more sophisticated models that can better capture the dynamic nature of AI demand across different sectors and applications. This could involve incorporating machine learning techniques to predict and model demand patterns based on historical data and emerging trends. Additionally, future research could explore the use of agent-based modeling to simulate the behavior of various stakeholders in the AI ecosystem and their impact on energy demand.

Enhancing system dynamics models to better integrate exogenous factors is crucial for more accurate predictions

In the current study, exogenous factors are "factored in" rather than fully integrated into causal diagrams to balance complexity with model convergence. Future research should explore methods to more fully integrate these exogenous factors into system dynamics models without compromising model stability or interpretability. This could involve developing hierarchical models or employing advanced simulation techniques that can handle increased complexity. Future research might also consider using Bayesian networks or other probabilistic modeling approaches to better represent the uncertainties associated with exogenous factors.

Advocating for greater transparency in LLM Model Architecture

The lack of transparency in large language model (LLM) architecture and training details from major tech companies poses challenges for accurately modeling Gen AI Training and Inferencing sub-models.

Future research should advocate for greater transparency in AI development and explore methods to estimate model architectures and training processes based on available information. This could involve developing reverse engineering techniques or collaborating with industry partners to gain more accurate insights into AI model development processes. Additionally, researchers could work on developing standardized reporting frameworks for AI energy consumption and model architecture to facilitate more accurate and comparable studies.

Developing country-specific and regional models aligns with localized contexts

Given that different countries may experience various scenarios simultaneously or transition between scenarios over time, future research should focus on developing country-specific models and analyses. This would involve considering local energy infrastructure, regulatory environments, economic conditions, and technological adoption rates to create more tailored and actionable insights for policymakers and industry leaders in specific regions. Such research could help identify best practices and potential pitfalls in AI energy management across different cultural and economic contexts.

Understanding the dynamics of transitions between scenarios is crucial for strategic planning

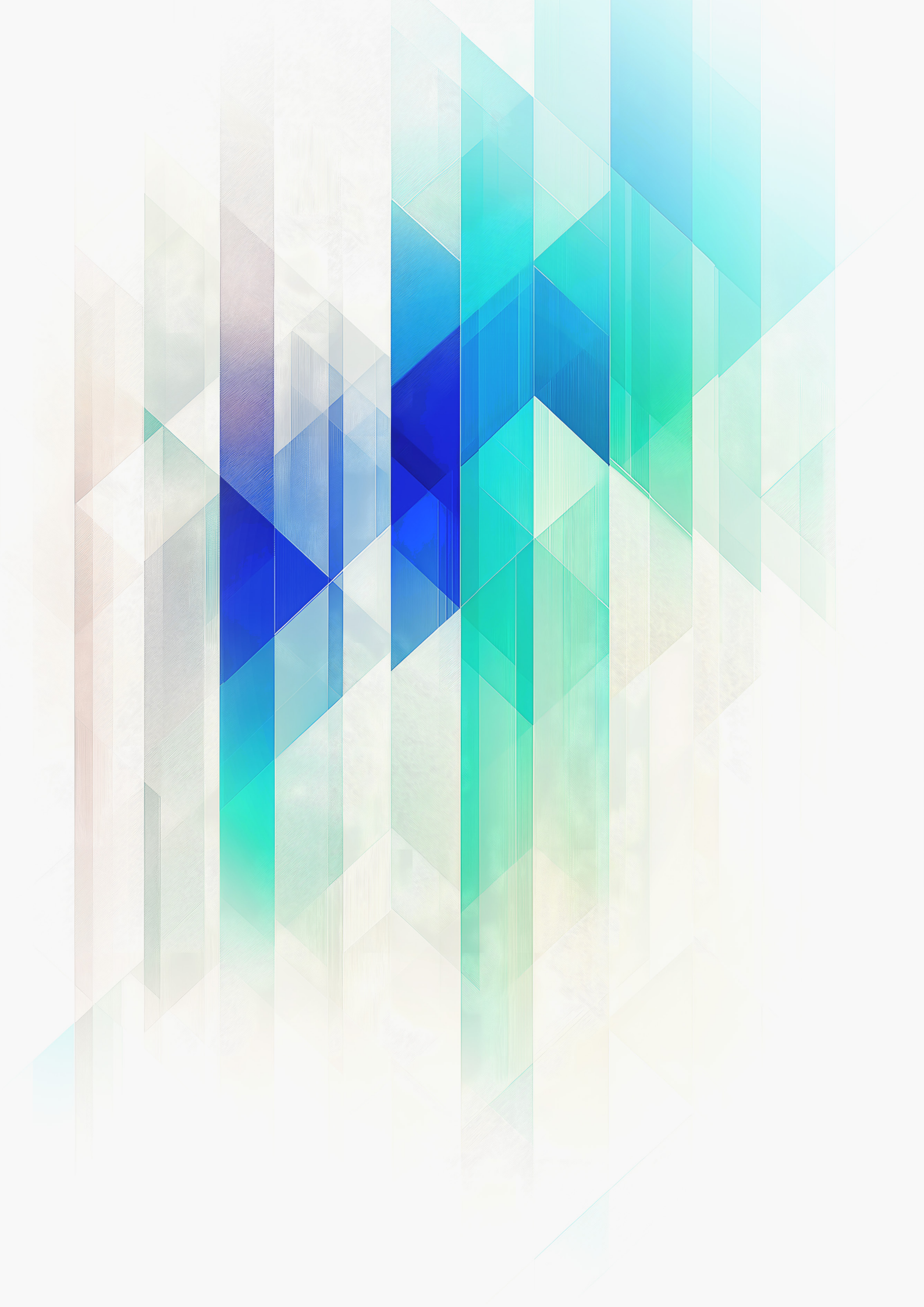
Further research is needed to understand the dynamics of transitions between scenarios. This could involve developing models that capture the tipping points and feedback loops driving shifts from one scenario to another. Such research would be valuable for identifying early warning signs of potential energy crises or opportunities for transitioning towards more sustainable AI development pathways. These AI-related dynamics can take various forms, such as scenario variants, scenario adaptations, scenario rewritings, or even scenario twists. The dynamics of these digital transitions should be studied in future research, employing system dynamics and game theory for modeling.

Quantifying and modeling indirect effects of AI on energy consumption is essential for a comprehensive understanding

Future studies should aim to quantify and model the indirect effects of AI on energy consumption in various sectors. This includes investigating potential rebound effects where efficiency gains in one area may lead to increased energy consumption elsewhere. Understanding these complex interactions is crucial for developing a more comprehensive picture of AI's overall impact on energy systems and sustainability. Researchers could employ econometric techniques and system dynamics modelling to capture these indirect effects and rebound phenomena.

Assessing the effectiveness of different policy approaches is crucial for shaping the future of AI energy consumption

As the study highlights the importance of policy interventions in shaping AI's energy future, further research is needed to assess the effectiveness of different policy approaches. This could involve comparative analyses of AI and energy policies across different countries, as well as modelling the potential impacts of proposed policy interventions.



Appendices

Thomas Edison

Appendices

References (1/12)

1.Rathi, A., & Bass, D. (2024). Microsoft's AI Push Imperils Climate Goal as Carbon Emissions Jump 30%. Bloomberg. https://www.bloomberg.com/news/articles/2024-05-15/microsoft-s-ai-investment-imperils-climate-goal-as-emissions-jump-30?embedded-checkout=true

2.de Vries, A. (2023). The growing energy footprint of artificial intelligence. Joule, 7(10), 2191–2194. https://doi.org/10.1016/j.joule.2023.09.004

3.International Energy Agency (IEA). (2023). Data Centres and Data Transmission Networks. https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks

4.Halper, E. (2024). Amid explosive demand, America is running out of power. The Washington Post. https://www.washingtonpost.com/business/2024/03/07/ai-data-centers-power/

5.Tilley, A., & Jie, Y. (2024). Apple Is Developing AI Chips for Data Centers, Seeking Edge in Arms Race. The Wall Street Journal. https://www.wsj.com/tech/ai/apple-is-developing-ai-chips-for-data-centers-seeking-edge-in-arms-race-0bedd2b2

6.Goldman Sachs Research. (2023). The Potentially Large Energy Footprint of Artificial Intelligence. https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand

7.Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. arXiv preprint arXiv:1906.02243. https://arxiv.org/abs/1906.02243

8.Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., ... Bengio, Y. (2019). Tackling Climate Change with Machine Learning. arXiv preprint arXiv:1906.05433. https://arxiv.org/abs/1906.05433

9.ABI Research. (2023). Data Center and Colocation Market Tracker. https://www.abiresearch.com/market-research/product/market-data/MD-NGDC/

10.Koomey, J., & Masanet, E. (2021). Does not compute: Avoiding pitfalls assessing the Internet's energy and carbon impacts. Joule, 5(7), 1625-1628. https://doi.org/10.1016/j.joule.2021.05.007

11.TIME. (n.d.). How AI Is Fueling a Boom in Data Centers and Energy Demand. https://time.com/6987773/ai-data-centers-energy-usage-climate-change/

12.The Guardian. (n.d.). Ireland: datacentres overtake electricity use of all homes combined, figures show. https://www.theguardian.com/world/article/2024/jul/23/ireland-datacentres-overtake-electricity-use-of-all-homes-combined-figures-show

13.Data Center Frontier. (n.d.). IEA Study Sees AI, Cryptocurrency Doubling Data Center Energy Consumption by 2026. https://www.datacenterfrontier.com/energy/article/33038469/iea-study-sees-ai-cryptocurrency-doubling-data-center-energy-consumption-by-2026

14.Data Centers. (n.d.). New Power Usage Rules Will Impact Amsterdam Data Centers. https://www.datacenters.com/news/new-power-usage-rules-will-impact-amsterdam-data-centers

15.Paccou, R. (2024). AI for Impact: A Method for Guiding AI-Energy Applications at Scale. https://www.se.com/ww/en/insights/sustainability/sustainability-research-institute/AI-for-impact-a-method-for-guiding-AI-energy-applications-at-scale/

16.Nicholas Crafts, 2021. "Artificial intelligence as a general-purpose technology: an historical perspective," Oxford Review of Economic Policy, Oxford University Press and Oxford Review of Economic Policy Limited, vol. 37(3), pages 521-536. https://ideas.repec.org/a/oup/oxford/v37y2021i3p521-536..html

17.Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. Nature Communications, 11(1), 233. https://www.nature.com/articles/s41467-019-14108-y

18.Jones, N. (2018). How to stop data centres from gobbling up the world's electricity. Nature, 561(7722), 163-166. https://doi.org/10.1038/d41586-018-06610-y

19.Masanet, E., Lei, N., & Koomey, J. (2024). How will the electricity use of AI data centers evolve? To answer this question, energy analysts need better data. https://doi.org/10.13140/RG.2.2.11203.00801

20.Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54-63. https://doi.org/10.1145/3381831

References (2/12)

21.Rahman-Jones, I. (2024). Google's greenhouse gas emissions up 48% in 5 years. BBC News. https://www.bbc.com/news/technology-12345678

22.Google. (2024). 2024 Environmental Report. https://sustainability.google/reports/2024-environmental-report

23.Hintemann, R., & Hinterholzer, S. (2022). Energy consumption of data centers worldwide. Borderstep Institute for Innovation and Sustainability. https://www.borderstep.de/en/publications/energy-consumption-data-centers

24.Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645-3650). https://doi.org/10.18653/v1/P19-1363

25.Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (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 (pp. 610-623). https://doi.org/10.1145/3442188.3445922

26.Ekchajzer, D., Hilty, L. M., Kern, E., & Penzenstadler, B. (2024). Decision-making under environmental complexity: The need for moving from avoided impacts of ICT solutions to systems thinking approaches. In ICT4S 2024: International Conference on ICT for Sustainability. https://hal-lara.archives-ouvertes.fr/LITEM/hal-04637677v1

27.Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700. https://arxiv.org/abs/1910.09700

28.Bartlett, K. (2024). Google's carbon emissions surge nearly 50% due to AI energy demand. CNBC. https://www.cnbc.com/2024/01/01/google-carbon-emissions-surge.html

29.Milmo, D. (2024). Can the climate survive the insatiable energy demands of the AI arms race? The Guardian. https://www.theguardian.com/environment/2024/jan/01/climate-ai-energy-demands

30.Petit, V. (2021). Digital economy and climate impact. https://www.se.com/ww/en/insights/electricity-4-0/digital-for-recovery/digital-economy-and-climate-impact.jsp

31.Kerr, D. (2024, July 12). AI brings soaring emissions for Google and Microsoft, a major contributor to climate change. NPR. https://www.npr.org/2024/07/12/g-s1-9545/ai-brings-soaring-emissions-for-google-and-microsoft-a-major-contributor-to-climate-change

32.Sterman, J. D. (2000). Business dynamics: Systems thinking and modeling for a complex world. McGraw-Hill Education. https://www.mheducation.com/highered/product/business-dynamics-sterman/M9780072317202.html

33.Forrester, J. W. (1994). System dynamics, systems thinking, and soft OR. System Dynamics Review, 10(203), 245-256. https://doi.org/10.1002/sdr.4260100211

34.Jones, N. (2018). How to stop data centres from gobbling up the world's electricity. Nature, 561(7722), 163-166. https://doi.org/10.1038/d41586-018-06610-y

35.Masanet, E., Lei, N., & Koomey, J. (2024). How will the electricity use of AI data centers evolve? To answer this question, energy analysts need better data. https://doi.org/10.13140/RG.2.2.11203.00801

36.Schwartz, R., Dodge, J., Smith, N.A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54-63. https://doi.org/10.1145/3381831

37.Rahman-Jones, I. (2024). Google's greenhouse gas emissions up 48% in 5 years. BBC News. https://www.bbc.com/news/technology-12345678

38.Google. (2024). 2024 Environmental Report. https://sustainability.google/reports/2024-environmental-report

39.Hintemann, R., & Hinterholzer, S. (2022). Energy consumption of data centers worldwide. Borderstep Institute for Innovation and Sustainability. https://www.borderstep.de/en/publications/energy-consumption-data-centers

40.Ekchajzer, D., Hilty, L.M., Kern, E., & Penzenstadler, B. (2024). Decision-making under environmental complexity: The need for moving from avoided impacts of ICT solutions to systems thinking approaches. In ICT4S 2024: International Conference on ICT for Sustainability. https://hal-lara.archives-ouvertes.fr/LITEM/hal-04637677v1

41.Bartlett, K. (2024, July 28). How the massive power draw of generative AI is overtaxing our grid. CNBC. https://www.cnbc.com/2024/07/28/how-the-massive-power-draw-of-generative-ai-is-overtaxing-our-grid.html

42.Milmo, D. (2024). Can the climate survive the insatiable energy demands of the AI arms race? The Guardian. https://www.theguardian.com/business/article/2024/jul/04/can-the-climate-survive-the-insatiable-energy-demands-of-the-ai-arms-race

43.BloombergNEF, Statkraft, & Eaton. (2021). Data Centers and Critical Infrastructure. BloombergNEF, Statkraft, & Eaton. (2021). Data Centers and Critical Infrastructure. https://www.eaton.com/content/dam/eaton/company/news-insights/energy-transition/documents/bnef-eaton-statkraft-data-center-study-en-us.pdf

References (3/12)

44.CNBC TV18. (2024, July 12). Google's emissions up 48% in five years due to AI. https://www.cnbctv18.com/technology/google-emissions-up-48-percent-in-5-years-due-to-ai-19437794.htm

45.The Conference Board. (2024). Report: The Rise of AI Threatens to Explode US Electricity. https://www.conference-board.org/topics/AI-for-business/press/ai-electricity-demand

46.EPRI. (2024). EPRI Study: Data Centers Could Consume up to 9% of U.S. Electricity Generation by 2030. https://www.epri.com/about/media-resources/press-release/q5vU86fr8TKxATfX8IHf1U48Vw4r1DZF

47.Carbon Credits. (2024). US Data Center Power Use Will Double by 2030 Because of AI. https://carboncredits.com/us-data-centers-power-requirement-will-double-by-2030/

48.Goldman Sachs. (2024). AI is poised to drive 160% increase in data center power demand. https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand

49.Jockims, T. L. (2024, September 5). A big tech plan for carbon storage at center of climate change debate. CNBC. https://www.cnbc.com/2024/09/05/a-big-tech-plan-for-carbon-storage-at-center-of-climate-change-debate.html

50.CNBC. (2024, October 18). Google reorganization puts AI in the spotlight. https://www.cnbc.com/video/2024/10/18/google-reorganization-puts-ai-in-the-spotlight.html

51.Lohr, S. (2024, August 26). Will A.I. Ruin the Planet or Save the Planet? The New York Times. https://www.nytimes.com/2024/08/26/climate/ai-planet-climate-change.html

52.Ballard, E. (2024). Air Conditioning and AI Are Demanding More of the World's Power—Renewables Can't Keep Up. The Wall Street Journal. https://www.wsj.com/business/energy-oil/air-conditioning-and-ai-are-demanding-more-of-the-worlds-powerrenewables-cant-keep-up-987a58f3

53.The Washington Post. (2024). Washington Post Live. https://www.washingtonpost.com/washington-post-live/

54.Financial Times. (2024). Future of AI. https://b2blandingpages.ft.com/rs/235-OIJ-339/images/2024%20_%20Future%20of%20AI%20special%20report.pdf?version=0

55.World Economic Forum. (2024, September 15). AI will accelerate sustainability — but is no silver bullet. https://www.weforum.org/stories/2024/09/ai-accelerator-sustainability-silver-bullet-sdim/

56.Environment + Energy Leader. (2024, October 3). Does AI have a sustainability dilemma? Salesforce research shows optimism despite AI energy demands. https://www.environmentenergyleader.com/stories/does-ai-have-a-sustainability-dilemma-salesforce-research-shows-optimism-despite-ai-energy-demands,44919

57.AIMultiple. (2024). Top 10 Sustainability AI Applications. https://research.aimultiple.com/sustainability-ai/

58.United Nations Environment Programme. (2024). AI has an environmental problem. Here's what the world can do about it. https://www.unep.org/news-and-stories/story/ai-has-environmental-problem-heres-what-world-can-do-about

59.World Economic Forum. (2024). AI will accelerate sustainability — but is no silver bullet. https://www.weforum.org/stories/2024/09/ai-accelerator-sustainability-silver-bullet-sdim/

60.Koomey, J., & Masanet, E. (2024, July 18). The uneven distribution of AI's environmental impacts. Harvard Business Review. https://hbr.org/2024/07/the-uneven-distribution-of-ais-environmental-impacts

61.AI for Education. (2024). AI's Impact on the Environment. https://www.aiforeducation.io/ai-resources/ais-impact-on-the-environment

62.United Nations University. (2024). Artificial Intelligence – Help or Harm for the Climate? https://unu.edu/inweh/press-release/unu-report-dont-dismiss-great-power-ai-climate-change-impact-assessment

63.Jones, N. (2023). The energy cost of artificial intelligence. Nature, 614(7947), 398-401. https://www.nature.com/articles/d41586-023-00288-7

64.Zhang, C., et al. (2022). The hidden cost of AI. Harvard Business Review. https://hbr.org/2022/09/the-hidden-cost-of-ai

65.Naturvårdsverket. (n.d.). Using systems approach to integrate causal loop diagrams modelling. https://www.naturvardsverket.se/publikationer/6900/using-systems-approach-to-integrate-causal-loop-diagrams-modelling/

66.Wang, Q. (2024). Ecological footprints, carbon emissions, and energy transitions. Nature Human Behaviour. https://www.nature.com/articles/s41599-024-03520-5

67.Lin, G., Palopoli, M., & Dadwal, V. (2020). From Causal Loop Diagrams to System Dynamics Models in a Data-Rich Ecosystem. In L. A. Celi, M. S. Majumder, P. Ordóñez, J. S. Osorio, K. E. Paik, & M. Somai (Eds.), Leveraging Data Science for Global Health (pp. 101-127). Springer, Cham. https://doi.org/10.1007/978-3-030-47994-7_6

References (4/12)

68.Gan, W., Huang, J., Xiao, X., Xie, Y., Huang, T., & Yu, P. S. (2023). Web3: The Next Internet Revolution. arXiv. https://arxiv.org/abs/2304.06111

69.Wang, Q. (2024). Ecological footprints, carbon emissions, and energy transitions. Nature Human Behaviour. https://www.nature.com/articles/s41599-024-03520-5

70.Denzin, N. K. (1978). The research act: A theoretical introduction to sociological methods. New York: McGraw-Hill. https://www.taylorfrancis.com/books/mono/10.4324/9781315134543/research-act-norman-denzin

71.Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. Health Services Research, 34(5 Pt 2), 1189-1208. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1089059/

72.Flick, U. (2004). Triangulation in qualitative research. In U. Flick, E. von Kardorff, & I. Steinke (Eds.), A companion to qualitative research (pp. 178-183). London: Sage. https://www.researchgate.net/profile/Hubert-Knoblauch/publication/329165047_A_Companion_to_QUALITATIVE_RESEARCH_Edited_by_Uwe_Flick_Ernst_von_Kardorff_and_Ines_Steinke/links/5bf93e0c458515a69e3860d6/A-Companion-to-QUALITATIVE-RESEARCH-Edited-by-Uwe-Flick-Ernst-von-Kardorff-and-Ines-Steinke.pdf

73.Schneider Electric. (Avelar et al.). Data Centers and Networks. https://www.se.com/ww/en/work/solutions/for-business/data-centers-and-networks/#Resources

74.Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. Nursing Research, 40(2), 120-123. https://www.researchgate.net/profile/Janice-Morse/publication/231749811_Development_of_a_Scale_to_Identify_the_Fall-Prone_Patient/links/54e88a070cf27a6de10f01bc/Development-of-a-Scale-to-Identify-the-Fall-Prone-Patient.pdf

75.Carter, N., Bryant-Lukosius, D., DiCenso, A., Blythe, J., & Neville, A. J. (2014). The use of triangulation in qualitative research. Oncology Nursing Forum, 41(5), 545-547. https://doi.org/10.1188/14.ONF.545-547

76.Hobbhahn, M., Heim, L., & Aydos, G. (2023). Trends in Machine Learning Hardware. Epoch. https://epochai.org/blog/trends-in-machine-learning-hardware

77.Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., & Sifre, L. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556. https://arxiv.org/abs/2203.15556

78.Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., Atkinson, D., Thompson, N., & Sevilla, J. (2024). Algorithmic progress in language models. arXiv preprint arXiv:2403.05812. https://arxiv.org/abs/2403.05812

79.Epoch. (2023). Trends in the dollar training cost of machine learning. LessWrong. https://www.lesswrong.com/posts/REBFQF43nwcJgp8Ge/trends-in-the-dollar-training-cost-of-machine-learning-1

80.SemiAnalysis. (2023, May 29). AI Server Cost Analysis: Memory Is The New Bottleneck. https://semianalysis.com/2023/05/29/ai-server-cost-analysis-memory-is/

81.Wilson, C., & Verdolini, E. (Organizers). (2024, May 13-14). Expert Workshop on Digitalisation Narratives and Climate Change Mitigation [Workshop]. International Institute for Applied Systems Analysis, Laxenburg, Austria. https://iiasa.ac.at/blog/jun-2024/expert-workshop-on-digitalization-narratives-and-climate-change-mitigation

82.Cohen, A. (2024, May 23). AI is pushing the world towards an energy crisis. Forbes. https://www.forbes.com/sites/arielcohen/2024/05/23/ai-is-pushing-the-world-towards-an-energy-crisis/

83.Ofcom. (2023). Online Nation 2023 report. https://www.ofcom.org.uk/research-and-data/internet-and-on-demand-research/online-nation/online-nation-2023

84.Center for Economic Policy Research. (2023). The impact of artificial intelligence on the labour market. https://cepr.org/voxeu/columns/impact-artificial-intelligence-labour-market

85.Schneider Electric™ Sustainability Research Institute, based on Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). Compute Trends Across Three Eras of Machine Learning. arXiv preprint arXiv:2202.05924. https://arxiv.org/abs/2202.05924

86.MLCommons. (2023). MLPerf Inference v3.0 Results. https://mlcommons.org/en/inference-datacenter-23/

87.Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., Atkinson, D., Thompson, N., & Sevilla, J. (2024). Algorithmic progress in language models. arXiv preprint arXiv:2403.05812. https://arxiv.org/abs/2403.05812

88.Gao, Y., Shen, S., Peng, Z., Luo, Y., Bian, J., & Liu, T. Y. (2023). The Efficiency Spectrum of Large Language Models: An Algorithmic Survey. arXiv preprint arXiv:2312.00678v2. https://arxiv.org/abs/2312.00678

89.Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2020). Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence. IEEE Internet of Things Journal, 7(8), 7457-7469. https://ieeexplore.ieee.org/document/8976180

References (5/12)

90.European Parliament. (2024). AI and energy consumption | E-001977/2024. https://www.europarl.europa.eu/doceo/document/E-10-2024-001977_EN.html

91.Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felten, A., Kisser, J., Lundin, N., Nerini, F. F., Taddeo, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. Nature Communications, 11(1), 233. https://www.nature.com/articles/s41467-019-14108-y

92.Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2021). Deep learning's diminishing returns. IEEE Spectrum, 58(10), 50-55. https://spectrum.ieee.org/deep-learning-computational-cost

93.Xu, M., Qian, F., Pushp, S., Tian, Q., & Song, M. (2022). Rethinking AI computing systems for energy efficiency. Nature Communications, 13(1), 1-12. https://www.nature.com/articles/s41467-022-32354-5

94.Schneider Electric™ Sustainability Research Institute - Derived from data from Nvidia, AMD, Intel, and Google's TPUs

95.Barroso, L. A., et al. (2024). Beyond Efficiency: Scaling AI Sustainably. arXiv. https://arxiv.org/html/2406.05303v1

96.Kindig, B. (2024, June 20). AI Power Consumption: Rapidly Becoming Mission-Critical. Forbes. https://www.forbes.com/sites/bethkindig/2024/06/20/ai-power-consumption-rapidly-becoming-mission-critical/

97.Lightning AI. (2023). Doubling Neural Network Finetuning Efficiency with 16-bit Precision Techniques. Retrieved from https://lightning.ai/blog/doubling-neural-network-finetuning-efficiency-with-16-bit-precision-techniques/

98.Hugging Face. (n.d.). Methods and tools for efficient training on a single GPU. Retrieved from https://huggingface.co/docs/transformers/perf_train_gpu_one

99.Reddit. (2022). Mixed Precision Training: Difference between BF16 and FP16. Retrieved from https://www.reddit.com/r/MachineLearning/comments/vndtn8/d_mixed_precision_training_difference_between/

100.Halbiniak, K., et al. (2024). Unleashing the Potential of Mixed Precision in AI-Accelerated CFD Simulation on Intel CPU/GPU. Retrieved from https://www.iccs-meeting.org/archive/iccs2024/papers/148370197.pdf

101.Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). A definition, benchmark and database of AI for social good initiatives. Nature Machine Intelligence, 3(2), 111-115. https://doi.org/10.1038/s42256-021-00296-0

102.Kaack, L. H., Donti, P. L., Strubell, E., & Rolnick, D. (2022). Artificial intelligence and climate change: Opportunities, considerations, and policy levers to align AI with climate change goals. The Brookings Institution. https://www.brookings.edu/research/artificial-intelligence-and-climate-change-opportunities-considerations-and-policy-levers-to-align-ai-with-climate-change-goals/

103.Gupta, U., Kim, Y. G., Lee, S., Tse, J., Lee, H. H. S., Wei, G. Y., ... & Wu, C. J. (2022). Chasing carbon: The elusive environmental footprint of computing. IEEE Micro, 42(4), 37-47. https://ieeexplore.ieee.org/document/9793296

104.Tladi, T. (2024, July 12). AI and energy: Will AI reduce emissions or increase demand? World Economic Forum. https://www.weforum.org/stories/2024/07/generative-ai-energy-emissions/

105.SemiAnalysis. (2024, March 13). AI Datacenter Energy Dilemma – Race for AI Datacenter Space. https://semianalysis.com/2024/03/13/ai-datacenter-energy-dilemma-race/

106.Next Kraftwerke. (n.d.). Artificial Intelligence in the Energy Industry. https://www.next-kraftwerke.com/knowledge/artificial-intelligence

107.Bloomberg Intelligence. (2023). Generative AI to become a $1.3 trillion market by 2032. https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/

108.Elnion. (2024, May 16). Are superconducting computers poised to revolutionize data centres? https://elnion.com/2024/05/16/are-superconducting-computers-poised-to-revolutionise-data-centres/

109.Polytechnique Insights. (n.d.). Biocomputing: The promise of biological computing. https://www.polytechnique-insights.com/en/columns/science/biocomputing-the-promise-of-biological-computingbrains/

110.Singularity Hub. (2023). IBM is planning to build its first fault-tolerant quantum computer by 2029. https://singularityhub.com/2023/12/06/ibm-is-planning-to-build-its-first-fault-tolerant-quantum-computer-by-2029/

111.Jones, P. M. S., & Woite, G. (2009). Cost of nuclear and conventional baseload electricity generation. IAEA Bulletin. https://www.iaea.org/sites/default/files/publications/magazines/bulletin/bull32-3/32304781823.pdf

112.Our World of Energy. (2024). What does it cost to build a nuclear power plant? https://ourworldofenergy.org/nuclear-power-plant-cost/

113.Pielke Jr., R. (2024). How much does it cost to build a nuclear power plant? Pielke Jr.'s Substack. https://pielkejr.substack.com/p/how-much-does-it-cost-to-build-a-nuclear-power-plant

References (6/12)

114.Schlissel, D., & Biewald, B. (2008). Nuclear power plant construction costs. Synapse Energy Economics. https://www.synapse-energy.com/downloads/SynapseReport.2008-10.Nuclear-Costs.08-030.pdf

115.Orbach, B., & Orbach, E. (2024, September 12). The US is not prepared for the AI electricity demand shock. ProMarket. https://www.promarket.org/2024/09/12/the-us-is-not-prepared-for-the-ai-electricity-demand-shock/

116.DNV. (2024). Energy related emissions will peak in 2024 - DNV. https://www.dnv.com/news/eto-energy-related-emissions-will-peak-in-2024/

117.Yale E360. (2024). As use of A.I. soars, so does the energy and water it requires. https://e360.yale.edu/features/artificial-intelligence-climate-energy-emissions

118.Wired. (2024). AI's energy demands are out of control: Welcome to the internet's hyper-consumption era. https://www.wired.com/story/ai-energy-demands-water-impact-internet-hyper-consumption-era/

119.Planet Detroit. (2024). AI's environmental impact: Energy consumption and water use. https://planetdetroit.org/2024/10/ai-energy-carbon-emissions/

120.Whitehurst, A. (2024). Lowering AI's environmental impact. Business Reporter. https://www.business-reporter.com/sustainability/lowering-ais-environmental-impact

121.S&P Global. (2024). US DOE flags AI energy security risks, urges industry to tread carefully. https://www.spglobal.com/commodityinsights/en/market-insights/latest-news/electric-power/042924-us-doe-flags-ai-energy-security-risks-urges-industry-to-tread-carefully

122.Curious Earth. (2024). Energy-hungry AI challenges Ireland's data centres. https://curious.earth/blog/ireland-data-centre-ai/

123.S&P Global Market Intelligence. (2024). Power of AI: Surging datacenter load has Dominion bracing for AI's added demand. https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/power-of-ai-surging-datacenter-load-has-dominion-bracing-for-ai-s-added-demand-77792886

124.InnovationQuarter. (n.d.). Artificial intelligence accelerating energy transition in the Netherlands. https://www.investinrotterdamthehaguearea.org/news/artificial-intelligence-energy-transition-netherlands/

125.Cohen, A. (2024, May 23). AI is pushing the world towards an energy crisis. Forbes. https://www.forbes.com/sites/arielcohen/2024/05/23/ai-is-pushing-the-world-towards-an-energy-crisis/

126.SemiAnalysis. (2024, March 13). AI datacenter energy dilemma – Race for AI datacenter space. https://semianalysis.com/2024/03/13/ai-datacenter-energy-dilemma-race/

127.McKinsey & Company. (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/notes%20from%20the%20frontier%20modeling%20the%20impact%20of%20ai%20on%20the%20world%20economy/mgi-notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy-september-2018.ashx

128.The Climate Reality Project. (n.d.). Blackouts and the climate crisis. https://www.climaterealityproject.org/blog/blackouts-and-climate-crisis

129.Eurelectric. (n.d.). The coming storm: Building electricity resilience to extreme weather. https://resilience.eurelectric.org

130.Environment+Energy Leader. (2023). What the U.S. power grid needs to survive climate change. https://www.environmentenergyleader.com/stories/what-the-us-power-grid-needs-to-survive-climate-change,48311

131.Earth.org. (n.d.). The green dilemma: Can AI fulfil its potential without harming the environment? https://earth.org/the-green-dilemma-can-ai-fulfil-its-potential-without-harming-the-environment/

132.Stanford University. (n.d.). Artificial intelligence and the future of energy. https://energy.stanford.edu/news/artificial-intelligence-and-future-energy

133.Flow.ninja. (n.d.). How generative AI will change low-code/no-code development. https://www.flow.ninja/blog/how-generative-ai-will-change-low-code-no-code-development

134.Luccioni, A., Viguier, S., Ligozat, A. L., & Lefèvre, S. (2023). Power hungry processing: Watts driving the cost of AI deployment? arXiv. https://arxiv.org/abs/2311.16863

135.Markedium. (n.d.). Generative AI uses 30x more energy than search engines. https://markedium.com/generative-ai-uses-30x-more-energy-than-search-engines/

136.Lin, P., Bunger, R., & Avelar, V. (n.d.). Schneider Electric Energy Management Research Center. Schneider Electric. https://www.se.com/ww/en/download/document/SPD_WP133_EN/

References (7/12)

137.Jetcool. (n.d.). How power density is changing in data centers. https://jetcool.com/post/how-power-density-is-changing-in-data-centers/

138.Iceotope. (n.d.). Cooling the AI revolution in data centers. Data Center Frontier. https://www.datacenterfrontier.com/sponsored/article/55040910/iceotope-cooling-the-ai-revolution-in-data-centers

139.AIserver.eu. (n.d.). NVIDIA GB200 NVL72 - AI server. https://aiserver.eu/product/nvidia-gb200-nvl72/

140.SemiAnalysis. (2024, April 10). Nvidia Blackwell perf TCO analysis – B100 vs B200 vs GB200NVL72. https://semianalysis.com/2024/04/10/nvidia-blackwell-perf-tco-analysis/

141.NVIDIA Developer. (n.d.). NVIDIA GB200 NVL72 delivers trillion-parameter LLM training and real-time inference. https://developer.nvidia.com/blog/nvidia-gb200-nvl72-delivers-trillion-parameter-llm-training-and-real-time-inference/

142.NVIDIA Newsroom. (n.d.). NVIDIA Blackwell platform arrives to power a new era of computing. https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing

143.Exxact Corporation. (n.d.). Comparing Blackwell vs Hopper | B200 & B100 vs H200 & H100. https://www.exxactcorp.com/blog/hpc/comparing-nvidia-tensor-core-gpus

144.Semiengineering.com. (n.d.). Better optimization for many-core AI chips. https://semiengineering.com/better-optimization-for-manycore-ai-chips/

145.Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., ... Bengio, Y. (2022). Tackling climate change with machine learning. ACM Computing Surveys, 55(2), 1-96. https://doi.org/10.1145/3485128

146.Jouppi, N. P., Yoon, D. H., Kurian, G., Li, S., Patil, N., Laudon, J., Young, C., & Patterson, D. (2021). A domain-specific supercomputer for training deep neural networks. Communications of the ACM, 64(7), 67-76. https://doi.org/10.1145/3460225

147.Agarwal, S., Agarwal, S., Garg, A., Jain, A., Khandelwal, A., Chandra, S., ... & Saxena, S. (2024). MLPerf Power: Benchmarking the energy efficiency of machine learning systems. arXiv preprint arXiv:2410.12032. https://arxiv.org/abs/2410.12032

148.Baraniuk, C. (2024). Electricity grids creak as AI demands soar. BBC News. https://www.bbc.com/news/articles/cj5ll89dy2mo

149.Wang, L., & Wang, T. (n.d.). Small language models (SLMs): A cheaper, greener route into AI. UNESCO. https://www.unesco.org/en/articles/small-language-models-slms-cheaper-greener-route-ai

150.Infosys. (n.d.). Tiny AI for a sustainable digital future. https://www.infosys.com/iki/perspectives/tiny-ai-sustainable-digital-future.html

151.IBM. (n.d.). What is AI infrastructure? https://www.ibm.com/topics/ai-infrastructure

152.C-Suite Strategy. (n.d.). AI Hardware: Boosting Performance and Efficiency in Machine Learning Applications. https://www.c-suite-strategy.com/blog/ai-hardware-boosting-performance-and-efficiency-in-machine-learning-applications

153.Dataiku. (n.d.). Frugal AI: Value at Scale Without Breaking the Bank. https://blog.dataiku.com/frugal-ai-value-at-scale-without-breaking-the-bank

154.SemiAnalysis. (2024, September 4). Multi-Datacenter Training: OpenAI's Ambitious Plan To Beat Google. https://semianalysis.com/2024/09/04/multi-datacenter-training-openais/

155.GeeksforGeeks. (n.d.). Role of AI in Distributed Systems. https://www.geeksforgeeks.org/role-of-ai-in-distributed-systems/

156.Lark. (n.d.). Distributed Computing. https://www.larksuite.com/en_us/topics/ai-glossary/distributed-computing

157.Run:ai. (n.d.). 6 Amazing Distributed Computing Examples. https://www.run.ai/guides/distributed-computing/distributed-computing-examples

158.Janardhanan, H. (2024). AI-Driven Load Balancing for Energy-Efficient Data Centers. International Journal of Computer Trends and Technology, 72(8), 13-18. https://ijcttjournal.org/2024/Volume-72%20Issue-8/IJCTT-V72I8P103.pdf

159.Cloudzy. (n.d.). The Top Benefits of Load Balancing for Enterprises. https://cloudzy.com/blog/benefits-of-load-balancing/

160.NVIDIA. (2024). NVIDIA Blackwell Platform: Advancing AI through cooling and scaling innovations. NVIDIA Newsroom. https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-advances-ai-through-cooling-and-scaling-innovations

161.CoolIT Systems. (2024, October 15). CoolIT Systems' Growing Liquid-Cooling Product Line, Manufacturing Capacity to Support NVIDIA Blackwell Platform Ramp. https://www.coolitsystems.com/coolit-systems-growing-liquid-cooling-product-line-manufacturing-capacity-to-support-nvidia-blackwell-platform-ramp/

References (8/12)

162.ZutaCore. (2024, April 8). ZutaCore's HyperCool Liquid Cooling Technology to Support NVIDIA's Advanced GPUs for Sustainable AI. Intelligent Data Centres. https://www.intelligentdatacentres.com/2024/04/08/zutacores-hypercool-liquid-cooling-technology-to-support-nvidias-advanced-gpus-for-sustainable-ai/

163.NVIDIA. (n.d.). NVIDIA Blackwell Platform Arrives to Power a New Era of Computing. NVIDIA Newsroom. https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing

164.DataBank. (n.d.). Data Center Scalability For Growing Tech Companies. https://www.databank.com/resources/blogs/data-center-scalability-for-growing-tech-companies/

165.Splunk. (n.d.). Moore's Law: A Complete Introduction. https://www.splunk.com/en_us/blog/learn/moores-law.html

166.Kindig, B. (2024, June 7). Here's Why Nvidia Stock Will Reach $10 Trillion Market Cap By 2030. Forbes. https://www.forbes.com/sites/bethkindig/2024/06/07/prediction-nvidia-stock-will-reach-10-trillion-market-cap-by-2030/

167.Nebius. (n.d.). InfiniBand in focus: bandwidth, speeds and high-performance computing. https://nebius.com/blog/posts/what-is-infiniband

168.XenonStack. (n.d.). Distributed Machine Learning Frameworks and its Benefits. https://www.xenonstack.com/blog/distributed-ml-framework

169.IBM. (2024, February 12). What is transfer learning? https://www.ibm.com/topics/transfer-learning

170.Nebius. (n.d.). What is AutoML? Understanding automated machine learning. https://nebius.com/blog/posts/what-is-automl

171.Tech Xplore. (2024, September 1). AI is 'accelerating the climate crisis,' expert warns. https://techxplore.com/news/2024-09-ai-climate-crisis-expert.html

172.Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. Journal of Machine Learning Research, 21(248), 1-43. https://jmlr.org/papers/volume21/20-312/20-312.pdf

173.Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., So, D., Texier, M., & Dean, J. (2022). The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. arXiv. https://arxiv.org/abs/2204.05149

174.Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54-63. https://cacm.acm.org/research/green-ai/

175.AFNOR Group. (n.d.). A benchmark for measuring and reducing the environmental impact of AI. https://www.afnor.org/en/news/referential-for-measuring-and-reducing-environmental-impact-of-ia/

176.Le Pape-Gardeux, C., & Kluska, J. (2024, July 2). AI on a diet: How to apply frugal AI standards. Schneider Electric Blog. https://blog.se.com/digital-transformation/artificial-intelligence/2024/07/02/ai-on-a-diet-how-to-apply-frugal-ai-standards/

177.Epoch AI. (n.d.). Can AI Scaling Continue Through 2030? https://epochai.org/blog/can-ai-scaling-continue-through-2030

178.Diana, F. (2023, August 7). The Chip Crisis: Implications For Generative AI's Progress. https://frankdiana.net/2023/08/07/the-chip-crisis-implications-for-generative-ais-progress/

179.Digitimes. (2024, October 31). Generative AI faces bottlenecks from limited data, rising resource demands. https://www.digitimes.com/news/a20241031PD212/genai-data-demand-copper.html

180.Data Center Dynamics. (n.d.). OpenAI pitched White House on multiple 5GW data centers. https://www.datacenterdynamics.com/en/news/openai-pitched-white-house-on-multiple-5gw-data-centers/

181.Sevilla, J., & Roldán, E. (2024). Training Compute of Frontier AI Models Grows by 4-5x per Year. Epoch AI. https://epochai.org/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year

182.A&O Shearman. (n.d.). Generative AI hardware - the other arms race. https://www.aoshearman.com/en/insights/generative-ai-hardware-the-other-arms-race

183.SemiAnalysis. (2024, February 21). Groq Inference Tokenomics: Speed, But At What Cost? https://semianalysis.com/2024/02/21/groq-inference-tokenomics-speed-but/

184.Epoch AI. (n.d.). Can AI Scaling Continue Through 2030? https://epochai.org/blog/can-ai-scaling-continue-through-2030

185.Semiengineering. (n.d.). Higher Density, More Data Create New Bottlenecks In AI Chips. https://semiengineering.com/higher-density-more-data-create-new-bottlenecks-in-ai-chips/

186.Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). Will we run out of data? Limits of LLM scaling based on human-generated data. arXiv. https://arxiv.org/abs/2211.04325

References (9/12)

187.Webster, T. (2024, January 2). AI Faces a Data Drought … for Real. PYMNTS.com. https://www.pymnts.com/artificial-intelligence-2/2024/facing-a-data-drought-ai-industrys-race-to-find-high-quality-information/

188.FutureBeeAI. (n.d.). Why is Training Data Diversity Important for Machine Learning, AI. https://www.futurebeeai.com/blog/why-is-training-data-diversity-important-for-machine-learning-ai

189.MOHARA. (n.d.). Should You Build vs. Buy Generative AI? The Pros and Cons. https://mohara.co/should-you-build-vs-buy-generative-ai-the-pros-and-cons/

190.ITRex Group. (n.d.). Calculating the Cost of Generative AI. https://itrexgroup.com/blog/calculating-the-cost-of-generative-ai/

191.RINF.tech. (n.d.). Unveiling the Costs of Generative AI Projects. https://www.rinf.tech/unveiling-the-costs-of-generative-ai-projects/

192.Erdil, E., & Besiroglu, T. (2023). Increased Compute Efficiency and the Diffusion of AI Capabilities. arXiv. http://arxiv.org/pdf/2311.15377.pdf

193.All About Circuits. (n.d.). System Challenges of Generative AI Inference Acceleration. https://www.allaboutcircuits.com/industry-articles/system-challenges-of-generative-ai-inference-acceleration/

194.S&P Global. (2023, October 16). Wild predictions of power demand from AI put industry on edge. https://www.spglobal.com/commodityinsights/en/market-insights/latest-news/electric-power/101623-power-of-ai-wild-predictions-of-power-demand-from-ai-put-industry-on-edge

195.Semiengineering.com. (n.d.). How Inferencing Differs From Training in Machine Learning Applications. https://semiengineering.com/how-inferencing-differs-from-training-in-machine-learning-applications/

196.ClearML. (2024). The State of AI Infrastructure at Scale 2024. https://clear.ml/blog/the-state-of-ai-infrastructure-at-scale-2024

197.The Register. (2024). Cloud providers underutilizing GPUs for AI - report. Data Center Dynamics. https://www.datacenterdynamics.com/en/news/cloud-providers-underutilizing-gpus-for-ai-report/

198.Penguin Solutions. (n.d.). AI Factories and Creative GPU Utilization for AI. HPCwire. https://www.hpcwire.com/2023/12/04/ai-factories-and-creative-gpu-utilization-for-ai/

199.Smartly.AI. (n.d.). What is the CO2 emission per ChatGPT query? https://smartly.ai/blog/the-carbon-footprint-of-chatgpt-how-much-co2-does-a-query-generate

200.Motivair Corporation. (n.d.). AI Workloads and Liquid Cooling: The Future of HPC. https://www.motivaircorp.com/news/ai-workloads-and-liquid-cooling-the-future-of-hpc/

201.World Economic Forum. (2024, April). How to manage AI's energy demand today, tomorrow and in the future. https://www.weforum.org/stories/2024/04/how-to-manage-ais-energy-demand-today-tomorrow-and-in-the-future

202.Morton, J. (2024). Generative AI Adoption and Three Traps for Organizational Agility. California Management Review. https://cmr.berkeley.edu/2024/03/generative-ai-adoption-and-three-traps-for-organizational-agility/

203.Goldman Sachs. (2024). Gen AI: Too Much Spend, Too Little Benefit? https://www.goldmansachs.com/insights/top-of-mind/gen-ai-too-much-spend-too-little-benefit

204.Gartner. (2024). THE Journal. https://thejournal.com/articles/2024/08/06/gartner-30-of-gen-ai-projects-will-be-abandoned.aspx

205.McKinsey & Company. (2023). How generative AI could add trillions to the global economy. World Economic Forum. https://www.weforum.org/stories/2023/07/generative-ai-could-add-trillions-to-global-economy/

206.EMSNOW. (2024). Gartner predicts 30% of AI Projects will be abandoned by 2025. https://www.emsnow.com/gartner-predicts-30-of-ai-projects-will-be-abandoned-by-2025/

207.Garrido, G., Gidaris, S., Ponce, J., Thome, N., & Cord, M. (2024). Learning and Leveraging World Models in Visual Representation Learning. arXiv. https://arxiv.org/abs/2403.00504

208.Forbes. (2024). AI Power Consumption: Rapidly Becoming Mission-Critical. https://www.forbes.com/sites/bethkindig/2024/06/20/ai-power-consumption-rapidly-becoming-mission-critical/

209.Supply Chain Beyond. (n.d.). How the AI Boom is Impacting Supply Chains. https://supplychainbeyond.com/how-the-ai-boom-is-impacting-supply-chains/

210.Vespignani, J., & Smyth, R. (2024). Artificial intelligence investments reduce risks to critical mineral supply. Nature Communications, 15(1), 1-12. https://www.nature.com/articles/s41467-024-51661-7

211.Amodei, D., & Hernandez, D. (2018). AI and Compute. OpenAI. https://openai.com/research/ai-and-compute

References (10/12)

212.Data Center Dynamics. (n.d.). Meta details AI data center redesign that led to facilities being scrapped. https://www.datacenterdynamics.com/en/analysis/meta-details-ai-data-center-redesign-that-led-to-facilities-being-scrapped/

213.World Nuclear Association. (2024). Economics of Nuclear Power. https://world-nuclear.org/information-library/economic-aspects/economics-of-nuclear-power

214.MIT Sloan Management Review. (n.d.). Don't Get Distracted by the Hype Around Generative AI. https://sloanreview.mit.edu/article/dont-get-distracted-by-the-hype-around-generative-ai/

215.Anadolu Agency. (2024). World's top 10 tech companies worth more than China's GDP. https://www.aa.com.tr/en/economy/world-s-top-10-tech-companies-worth-more-than-china-s-gdp/3283342

216.Voronoi App. (n.d.). TSMC Hits the Trillion Dollar Milestone. https://www.voronoiapp.com/technology/-TSMC-Hits-the-Trillion-Dollar-Milestone-2783

217.Stock Analysis. (n.d.). Broadcom Market Cap. https://stockanaysis.com/stocks/avgo/market-cap/

218.Companies Market Cap. (n.d.). Samsung Electronics Market Cap. https://companiesmarketcap.com/eur/samsung/marketcap/

219.Daniel, W. (2024). China's Dominance in Critical Minerals: The Latest Battleground in the U.S.-China Trade War. Fortune. https://fortune.com/2024/06/10/china-near-monopoly-many-critical-minerals-jpmorgan-says-next-battleground-usa/

220.Mining Technology. (n.d.). US graphite tariffs. https://mine.nridigital.com/mine_oct24/us-graphite-tariffs

221.International Labour Organization. (n.d.). Mind the AI Divide: Shaping a Global Perspective on the Future of Work. https://www.ilo.org/publications/major-publications/mind-ai-divide-shaping-global-perspective-future-work

222.Feldman, M. (2023, May 18). Meta's GPU Buildout Is Bigger Than We Thought. The Next Platform. https://www.nextplatform.com/2023/05/18/metas-gpu-buildout-is-bigger-than-we-thought/

223.Smolaks, M. (2023, August 22). Tesla claims to have world's 5th most powerful supercomputer. Data Center Dynamics. https://www.datacenterdynamics.com/en/news/tesla-claims-to-have-worlds-5th-most-powerful-supercomputer/

224.Wikipedia. (2024). Leonardo (supercomputer). https://en.wikipedia.org/wiki/Leonardo_(supercomputer)

225.UX Tigers. (n.d.). AI Productivity. https://www.uxtigers.com/post/ai-productivity

226.Nielsen Norman Group. (n.d.). AI Tools Productivity Gains. https://www.nngroup.com/articles/ai-tools-productivity-gains/

227.McKinsey & Company. (n.d.). The economic potential of generative AI: The next productivity frontier. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

228.Wang, P., et al. (2024). The environmental impact of generative artificial intelligence: Forecasting e-waste generation from 2020 to 2030. Nature Computational Science. https://www.nature.com/articles/s41599-023-01234-5

229.Tzachor, A., Wang, P., & Others. (2024). The environmental impact of generative artificial intelligence: E-waste generation projections. Nature Computational Science. https://www.nature.com/articles/s41599-023-01234-5

230.Schwaller, F. (2024, October 28). E-waste from AI computers could 'escalate beyond control'. DW. https://www.dw.com/en/e-waste-from-ai-computers-could-escalate-beyond-control/a-70619724

231.ECS Environment. (n.d.). Hazardous substances in e-waste. https://www.ecsenvironment.com/what-is-e-waste/hazardous-substances-in-e-waste/

232.IEA. (2024). What the data centre and AI boom could mean for the energy sector. https://www.iea.org/commentaries/what-the-data-centre-and-ai-boom-could-mean-for-the-energy-sector

233.Baker, D., & Bhatia, S. (2024). Exploding AI Power Use: An Opportunity to Rethink Grid Planning. ACM Transactions on Internet Technology, 24(1), 1-20. https://dl.acm.org/doi/fullHtml/10.1145/3632775.3661959

234.IEA. (2023). Electricity Grids and Secure Energy Transitions. https://www.iea.org/reports/electricity-grids-and-secure-energy-transitions/executive-summary

235.Data Center Frontier. (2024). Dominion: Virginia's Data Center Cluster Could Double in Size. https://www.datacenterfrontier.com/energy/article/33013010/dominion-virginias-data-center-cluster-could-double-in-size

236.Columbia University. (2024). Projecting the Electricity Demand Growth of Generative AI Large Language Models in the US. https://www.energypolicy.columbia.edu/projecting-the-electricity-demand-growth-of-generative-ai-large-language-models-in-the-us/

237.Aurora Energy Research. (2024). Aurora report finds Northern Virginia data center demand could incentivize up to 15 GW of additional natural gas generators by 2030. https://auroraer.com/media/new-aurora-report-finds-northern-virginia-data-center-demand-could-incentivize-up-to-15-gw-of-additional-natural-gas-generators-by-2030/

References (11/12)

238.Data Center Dynamics. (2024). Meta signs 330MW renewable energy agreements in Illinois and Arkansas. https://www.datacenterdynamics.com/en/news/meta-signs-330mw-renewable-energy-agreements-in-illinois-and-arkansas/

239.World Nuclear News. (2024). Amazon invests in X-energy, unveils SMR project plans. https://world-nuclear-news.org/articles/amazon-invests-in-x-energy-unveils-smr-project-plans

240.Government Technology Insider. (2024). AI Implications – Power Requirements Going Nuclear on Local Grids. https://governmenttechnologyinsider.com/ai-implications-power-requirements-going-nuclear-on-local-grids/

241. Lohrmann, D. (2024). AI's Energy Appetite: Challenges for Our Future Electricity Supply. Government Technology. Retrieved from https://www.govtech.com/blogs/lohrmann-on-cybersecurity/ais-energy-appetite-challenges-for-our-future-electricity-supply

242.Source: Delta Power Solutions. (n.d.). Modular Data Centers: The Rise and the Advantages. Retrieved from https://www.deltapowersolutions.com/en/mcis/technical-article-modular-data-centers-the-rise-and-the-advantages.php

243.IEA. (2024). Data centres energy demand – a growing challenge. Retrieved from https://www.smart-energy.com/industry-sectors/energy-grid-management/data-centres-electricity-demand-is-increasing-finds-iea/

244.Data Centre Magazine. (2024). Power-Hungry Data Centres Put Pressure on Ireland's Grid. Retrieved from https://datacentremagazine.com/critical-environments/power-hungry-data-centres-put-pressure-on-irelands-grid

245.TeleGeography. (2024). The Data Center Sector: An Uncertain Juncture. Retrieved from https://blog.telegeography.com/the-data-center-sector-an-uncertain-juncture

246.Epoch AI. (2024). Can AI Scaling Continue Through 2030? Retrieved from https://epochai.org/blog/can-ai-scaling-continue-through-2030

247.McKinsey & Company. (2024). The role of power in unlocking the European AI revolution. Retrieved from https://www.mckinsey.com/industries/electric-power-and-natural-gas/our-insights/the-role-of-power-in-unlocking-the-european-ai-revolution

248.WTOP. (2024). Northern Virginia is again the No. 1 data center market, but challenges are mounting. Retrieved from https://wtop.com/business-finance/2024/03/northern-virginia-again-ranks-no-1-data-center-market-but-challenges-are-mounting/

249.International Energy Agency (IEA). (2024). Southeast Asia Energy Outlook 2024. Retrieved from https://www.iea.org/reports/southeast-asia-energy-outlook-2024

250. CFA Institute. (2024). The Hidden Environmental Costs of Tech Giants' AI Investments. Retrieved from https://blogs.cfainstitute.org/investor/2024/10/31/the-hidden-environmental-costs-of-tech-giants-ai-investments/

251.Yale E360. (2024). As Use of A.I. Soars, So Does the Energy and Water It Requires. Retrieved from https://e360.yale.edu/features/artificial-intelligence-climate-energy-emissions

252.The Innovator. (2024). Data Limitations Are Constraining AI Development. https://theinnovator.news/data-limitations-are-constraining-ai-development/

253.Gartner. (2024). The Future of Data: Synthetic Data Will Replace Real Data in AI Models. https://www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic-data-the-future-of-ai

254.SSRN. (2024). Synthetic Data and the Future of AI. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4722162

255.DataCamp. (2024). What is Multimodal AI? https://www.datacamp.com/blog/what-is-multimodal-ai

256.ITRex Group. (2024). Synthetic Data Generation Using Generative AI. https://itrexgroup.com/blog/synthetic-data-generation-using-generative-ai/

257.SAS. (2024). Harnessing synthetic data to fuel AI breakthroughs. https://www.sas.com/en_gb/insights/articles/analytics/synthetic-data-fuels-ai-breakthroughs.html

258.Jones, G. (2024). Energy is ripe for synthetic data disruption. BusinessGreen. https://www.businessgreen.com/opinion/4145287/energy-ripe-synthetic-disruption

259.International Energy Agency. (2024). What the data centre and AI boom could mean for the energy sector. https://www.iea.org/commentaries/what-the-data-centre-and-ai-boom-could-mean-for-the-energy-sector

260.Zhang, Y., & Zhou, Z. (2023). Rethinking deep learning's energy-performance relationship. arXiv. https://arxiv.org/abs/2310.06522

261.Brownlee, A. E. I., Adair, J., Haraldsson, S. O., & Jabbo, J. (2021). Exploring the accuracy–energy trade-off in machine learning. Proceedings of the 2021 IEEE/ACM International Conference on Software Engineering (ICSE), 1-12. https://doi.org/10.1109/ICSE43902.2021.00009

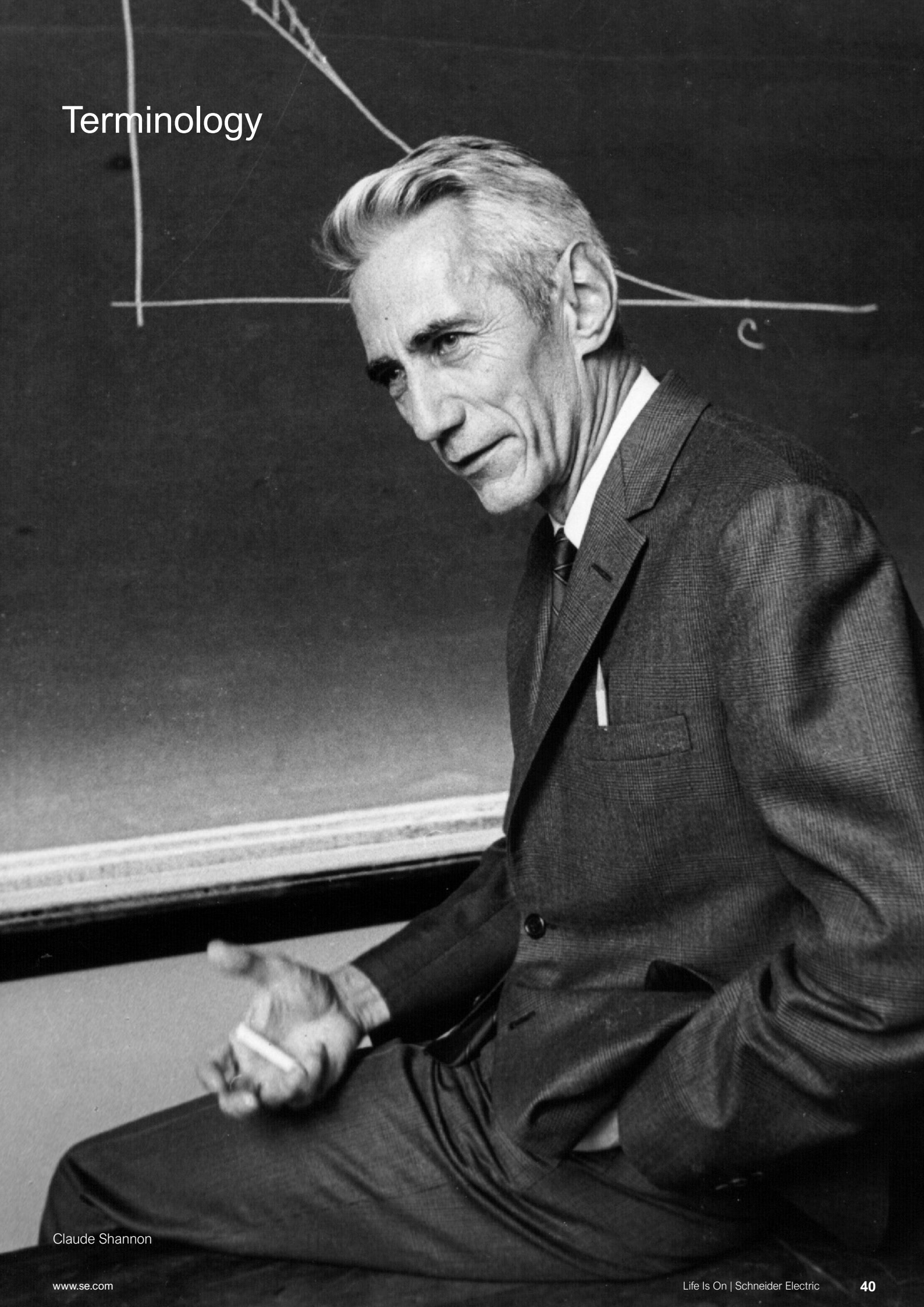
References (12/12)

262.Hoffman, A., & Hargreaves, T. (2024). The energy implications of AI development: Geographic concentration and local challenges. Energy Research & Social Science, 113, 102-115. https://doi.org/10.1016/j.erss.2023.102115

263.SemiAnalysis. (2024, March 13). AI datacenter energy dilemma race. https://semianalysis.com/2024/03/13/ai-datacenter-energy-dilemma-race/#gigawatt-dreams-and-matroyshka-brains-limited-by-datacenters-not-chips

264.Kindig, B. (2024, June 20). AI power consumption rapidly becoming mission critical. Forbes. https://www.forbes.com/sites/bethkindig/2024/06/20/ai-power-consumption-rapidly-becoming-mission-critical/

Terminology



Claude Shannon

Terminology (1/3)

• Algorithm: A step-by-step procedure (series of instructions) used for solving a problem or performing a computation. The execution of the algorithm unrolls an exact list of instructions leading to a sequence of specified actions in either hardware- or software-based routines.

• AMD MI300X: A competing high-performance GPU from AMD, which is considered superior to the NVIDIA H100 in some aspects.

• Analytics: Algorithms that exploit data to produce information of higher user value.

• Artificial Intelligence (AI): The ability of a system to handle representations, both explicit and implicit, and procedures to perform tasks that would be considered intelligent if performed by a human. [ETSI 2020]

• Artificial Intelligence Provider: An organization that provides products or services that use one or more AI systems, or a seller of AI components.

• Artificial Intelligence System: A machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment. [EU-US 2024] [OECD 2024]

• Atypical Data: Data representing a rare event. [LNE 2021]

• Bias: Systematic or disproportionate difference in the treatment of certain objects, people, or groups in comparison to others.

• BF16: Brain Float 16, a 16-bit floating-point format optimized for deep learning.

• Chatbot: Artificial Intelligence assistant that communicates via text chat.

• Computer Vision: A branch of Artificial Intelligence aimed at processing data available as images or videos, giving computers the ability to process and interpret visual data.

• Corner Case: Scenario located at the limits of the effective domain of use of a system, thus constituting a situation difficult to handle. [Adapted from LNE 2021]

• Data Curation: Preparation of collected data for use with the intended analytic approach. This can include integrating data from multiple sources and formats, identifying missing components of the data, removing errors and sources of noise, conversion of data into new formats, labelling the data, data augmentation using real and synthetic data, or scaling the data set using data synthesis approaches. [Adapted from ETSI 2020]

• Data Center: A facility used to house computer systems and associated components, such as telecommunications and storage systems. They consume significant amounts of electricity for both computing and cooling.

• Deep Learning: A special case of machine learning based on the use of a multi-layer neural network algorithm. The greater the number of layers, the deeper the network. [LNE 2021] Deep learning is a technique for implementing machine learning that relies on deep artificial neural networks to perform complex tasks such as image recognition, object detection, and natural language processing (NLP).

• Drift: Change in the distribution of data (data drift) or in the statistical relationships between the target variables and other variables (concept drift), which occurs over time, either instantaneously or gradually, predictably, or unpredictably. These changes can make a model built on old data incompatible with new data, requiring regular updates. [Adapted from LNE 2021]

• Efficiency: The degree to which a system or component performs its designated functions with minimal resource consumption. [LNE 2021]

• Embodied Emissions: The total amount of Sustainablehouse gas emissions generated to produce a product. In the context of AI, it refers to the emissions associated with manufacturing and disposing of AI hardware.

• Expert System: An Artificial Intelligence system that encapsulates knowledge provided by one or several human expert(s) in a specific domain to infer solutions to problems.

• Explainable Artificial Intelligence (XAI): A set of processes and methods that allows human users to comprehend and trust the results and output created by AI algorithms.

• Explicability: The ability to explicitly account for the elements leading or having led to an evaluation or a decision, based on known data and characteristics of the situation. [Adapted from LNE 2021]

• Fault Detection and Diagnostic (FDD): The identification and (possibly partial) understanding of an abnormal situation. Examples: detection of drifts in a factory, suggesting energetic inefficiency or impacts on product quality; detection of leaks in a water network.

Terminology (2/3)

• Federating Learning: A machine learning technique performed on data sets distributed across multiple devices or decentralized servers. [LNE 2021]

• FLOPS (Floating Point Operations Per Second): A measure of computer performance, particularly in fields of scientific calculations that make heavy use of floating-point calculations. It's often used to measure the performance of AI systems.

• Forecasting: The estimation of the value of some data in the future, based on past, present, and/or other forecasted data. Examples: building energy consumption forecasting; energy price forecasting.

• Foundation Model: A machine learning model trained on vast datasets so that it can be applied to across a wide range of use cases.

• GB200 superchips: The building blocks of the NVL72 system, combining Grace CPUs and B200 GPUs.

• General Purpose Technology (GPT): A term used to describe a new method of producing and inventing that is important enough to have a protracted aggregate impact. AI is often considered a GPT due to its potential to impact multiple sectors of the economy.

• Generative Artificial Intelligence (GenAI): An advanced technological approach that enables the creation of content including text, images, and videos. By analyzing and discerning patterns within extensive training datasets, Generative AI can autonomously construct material (like text, images, video and software) that shares comparable characteristics to its training input.

• GPU (Graphics Processing Unit): A specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images. GPUs are increasingly used for AI computations due to their parallel processing capabilities.

• Hopper architecture: The GPU architecture used in the H100, named after computer scientist Grace Hopper. It introduces various improvements over the previous Ampere architecture.

• Hybrid Artificial Intelligence System: Artificial Intelligence system integrating both machine learning techniques from data and models allowing to express constraints and to perform logical reasoning. [LNE 2021]

• HBM3e: High Bandwidth Memory, a type of computer memory with high-speed data transfer.

• Hyperparameter: A parameter whose value is used to control the learning process in machine learning models.

• H100 SXM5: A variant of the H100 GPU that uses NVIDIA's SXM form factor, offering higher performance and power consumption (up to 700W) compared to the PCIe version. It is designed for large-scale AI and HPC applications.

• H100 PCIe: A variant of the H100 GPU that uses the PCIe interface, offering more flexibility and easier installation in existing server infrastructure. It has lower power consumption (up to 350W) compared to the SXM version.

• Incremental Learning: Automatic learning performed on data grouped in batches, the batches being renewed periodically, as new data accumulates throughout the life cycle of the Artificial Intelligence functionality. [LNE 2021]

• Inferencing: The process of using a trained machine learning model to make predictions or decisions based on new input data. GenAI inferencing produces content on basis of a prompt.

• Input Data: Data provided to or directly acquired by an Artificial Intelligence (or Analytics) system, based on which the system produces an output. [Adapted from EC 2021]

• Interpretability: The ability to make the operation of an Artificial Intelligence (or Analytics) system understandable to a given category of users. [Adapted from LNE 2021]

• Invention of a Method of Invention (IMI): A concept referring to technologies or methodologies that can accelerate the process of innovation itself. AI is sometimes described as an IMI due to its potential to automate and accelerate research and development processes.

• Jevons Paradox: An economic theory stating that as technological progress increases the efficiency with which a resource is used, the rate of consumption of that resource may increase due to increasing demand.

• Large Language Model (LLM): A foundation model trained on immense amounts of data, making it capable of "understanding" and generating natural language and other types of content to perform a wide range of tasks.

• Learning (Training) Data: Data used for training an Artificial Intelligence system through fitting its learnable parameters, including the weights of a neural network. [EC 2021]

Terminology (3/3)

• Machine Learning: A branch of Artificial Intelligence in which a computer generates a model (e.g., a set of rules) based on raw data. Machine Learning refers to a broad set of techniques to train a computer to learn from its inputs, using existing data, and one or more "training" methods, instead of being explicitly programmed.

• Memristors: Electrical components that remember the amount of charge that has flowed through them.

• Mixed precision training: A technique in deep learning that uses different numerical precisions.

• Model: An abstract mathematical representation of a real-world event, system, behavior, or natural phenomenon, created on a computer to enable calculations and predictions.

• Model Execution Algorithm: An algorithm that applies a model to a provided set of input data to generate an output (prediction, recommendation, ...).

• Model Learning (Training) Algorithm: An algorithm that generates, improves, or adapts a model from a provided set of learning (training) data.

• Natural Language Processing: A branch of Artificial Intelligence aimed at processing data available in natural language, giving computers the ability to process and interpret text and spoken words..

• NVIDIA A100: The previous generation GPU from NVIDIA, based on the Ampere architecture. It is often used as a comparison point for the H100's performance improvements.

• NVIDIA H100 GPU: The latest high-performance graphics processing unit from NVIDIA, designed specifically for AI, deep learning, and high-performance computing workloads. It is built on the Hopper architecture.

• NVIDIA GB200 NVL72 system: A high-performance AI system based on NVIDIA's Blackwell architecture.

• NVLink: NVIDIA's high-speed interconnect technology for GPUs.

• Overfitting: The case in which the learned model matches the training data closely but does not generalize and fails to make correct predictions on new data. [Adapted from LNE 2021]

• Performance: The degree to which a system or component performs its designated functions within a given set of constraints, such as speed, accuracy, or memory usage. [LNE 2021]

• Prompt: An input word or statement with possible requirements for a GenAI system to produce output consisting of a meaningful ordered range of tokens, like a software program or piece of creative text.

• PUE (Power Usage Effectiveness): A metric used to determine the energy efficiency of a data center. It is calculated by dividing the total amount of power used by a data center by the power used for computing.

• Rack power density: The amount of power consumed per rack in a data center.

• Rare Event: Event of non-zero risk of occurrence over an infinite time, but which is observable only a few times over a large number of observations. [LNE 2021]

• Reinforcement Learning: Machine learning in which a policy defining how to act is learned by agents through experience to maximize their reward; and agents gain experience by interacting in an environment through state transitions. [Adapted from ETSI 2020]. Example: learning how to satisfy varying user comfort preferences.

• Representativeness: Quality of a sample constituted in such a way as to correspond to the population from which it is taken. [LNE 2021]

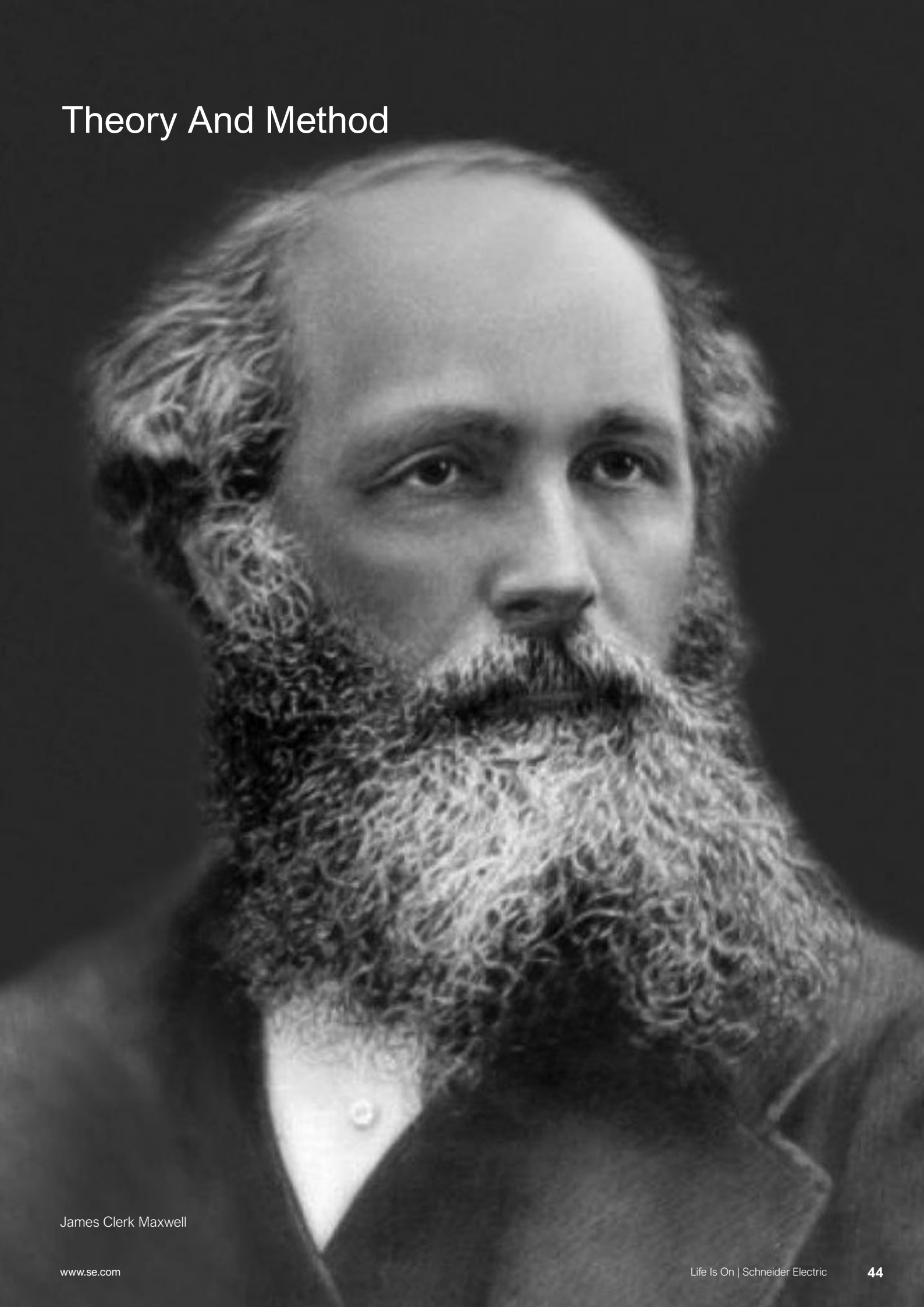
• Reproducibility: Reproducibility describes whether an experiment exhibits the same behaviour when repeated under the same conditions. [EC 2019]

• Resilience: The ability of a system to maintain compliance with expected requirements in the presence of inputs outside of its intended use domain. [Adapted from LNE 2021]

• Robustness: The ability of a system to maintain compliance with expected requirements in the presence of inputs within its intended use domain. [Adapted from LNE 2021]

• Semi-Supervised Learning: Machine learning in which the data set is partially labeled. Semi-supervised learning techniques are such that the unlabeled data can be used to improve the quality of the model. [Adapted from ETSI 2020]

Theory And Method



James Clerk Maxwell

Abstract and Preliminary Elements

Abstract

Forecasting the value of a social and economic phenomenon is difficult but useful if the influencing variables and their impacts are known, and if forecasters are willing to accept that there are multiple scenarios that may possibly unfold. This article presents the generic mechanisms (demand growth, efficiency, constraints, rebounds, and crunch), also known as archetypes, for increases or decreases in the volume of a phenomenon from a system dynamics perspective. These mechanisms are applied to the stock of electricity that data centers may need over a couple of years when the usage of data center facilities increases and when data centers experience continuous efficiency improvements that reduce this stock. Additionally, we propose the possibility of a disruption when the stock cannot be sufficiently supplied from the electricity resource, including the workings of exogenous mechanisms from the broader social contexts of data centers.

Preliminary elements

Artificial intelligence (AI) has become a focal point of intense debate, primarily due to its potential contributions, inherent risks, and the substantial resources required for its implementation. The AI High-Level Expert Group (HLEG) defines AI systems as software and possibly hardware systems created by humans. These systems, given a complex goal, operate in both physical and digital environments by perceiving their surroundings through data acquisition, interpreting structured and unstructured data, reasoning on the knowledge or processing the information derived from this data, and deciding the best actions to achieve the set goal. AI systems can use symbolic rules or learn numeric models, and can adapt their behavior by analyzing how their previous actions have affected the environment (AI HLEG, 2019a).

However, as noted by the EU Joint Research Centre (JRC), this definition may soon require updates due to the rapid advancements in AI. In the long term, AI systems might not necessarily be human-designed, and not every AI system may act autonomously; some may function merely as input-output modules or components.

AI's application domains are vast and diverse, encompassing smart home technologies, consumer health apps, cybersecurity, social networking, online shopping and recommender systems, search and answer applications like ChatGPT, automotive AI in vehicles and factories, healthcare services, educational tools, financial systems, entertainment, route planning, logistics optimizations, factory scheduling, and customer relationship management (CRM) through chat services. While there is significant interest in generative AI (GenAI) and large language models (LLMs), which are developed using extensive datasets and involve billions of model hyperparameters, many AI applications effectively operate on-device as TinyML, utilizing minimal data with efficient algorithms. Reinforcement learning, for example, employs smart algorithms to discover better pathways for achieving goals without necessitating large databases. These low data AI systems, which include TinyML, consume relatively low electricity but may incur substantial embodied emissions during their production and end-of-life stages. Despite their lower electricity consumption, low data AI can play a crucial role in reducing the energy usage of data centers by optimizing workload schedules and cooling infrastructure.

This study primarily focuses on the debates surrounding GenAI, particularly the electricity consumption associated with training and using these complex models with billions of hyperparameters. LLMs differ significantly from low data AI and data science or rule-based systems, as shown in Table 1.

They require substantial electricity for training and even more for widespread inferencing, exemplified by applications like CoPilot and ChatGPT. The deployment of LLMs necessitates significant data processing in data centers, leading to increased electricity consumption. These models have prompted revisions of core technical components in data centers, such as the shift from CPUs to GPUs, along with extensive updates to cooling technologies. The focus on GenAI and LLMs underscores the need to balance the benefits of advanced AI models with their environmental and resource implications.

Table 1: Types of AI

	Simple models	Complex models
Low data AI	TinyML and simple decision tools Low electricity usage For example thermostats	Rule-based systems Little energy usage for development and more for usage For example operations research optimization algorithms and reinforcement learning
Large datasets AI	Analytics models More electricity usage for analytics but less than for GenAI/LLMs For example: customer classifiers	Main focus of this study. GenAI/LLM use large volumes of electricity for training and very large volumes of electricity during widespread inferencing (11) For example Co Pilot and ChatGPT

Narratives About AI Impact

The data-science community is heavily involved in predictions based on correlations, neural networks, and other algorithms. Predicting the future, however, especially multiple years ahead, is a different challenge entirely, yet essential for investment decisions and company strategy formation. The strategic management literature is rich with methods for forecasting or foresighting through scenarios, as longer-term trends may be disrupted by difficult-to-foresee influencing factors and disruptions. Consequently, strategic decisions must work with multiple competing scenarios (Ari de Geus; Van der Heijden).

Recently, the rise of a new AI summer has resulted in future analyses of AI's impact on employment, profit, and resources (especially electricity) needed. These concerns have led to multiple forecasts of electricity needs for data centers, such as those by Goldman Sachs, Epoch AI, SemiAnalysis, and Koot and Wijnhoven (2021). While some of these predictions may be right or wrong, many of them (except Koot and Wijnhoven) are based on expert views but lack transparent reasoning for their forecasts. Freitag et al. (2021) is one of the few analyses that provide reasoned ways of looking at these forecasts, albeit without any calculation method. They provide four scenarios instead of one stream of thinking, exploring the possibilities of increasing demand and improving efficiencies.

Systemic impacts (16) are further formalized in the academic literature through the description of key variables influencing data centre (DC) electricity usage. Both professional and academic literature have attempted to quantify the longer-term effects of (Gen)AI. In this article, we focus on AI's electricity needs impact for data centers, although indirect and systemic impacts extend beyond data center electricity requirements.

Regarding GenAI's direct electricity needs, narratives can be either positive or negative. Positive narratives assert that AI training and inferencing, if done efficiently, will consume minimal energy, rendering GenAI's electricity impact nearly negligible compared to other energy uses (17). This positive narrative is also presented as the "AI will save ICT" hypothesis, suggesting AI will help data centres become more efficient, enabling further growth of the ICT sector (18). The distinction between low and high data AI is crucial here, as low data AI may significantly contribute to energy savings without substantially increasing total DC electricity consumption, potentially earning the label "Sustainable AI" (9). Conversely, negative narratives claim GenAI will consume large and increasing volumes of electricity if current trends of expanding training and inferencing volumes persist (14, 19). These narratives also suggest ICT electricity growth will continue, primarily due to extensive processing required for training and inferencing by billions of people in coming years (11). This could potentially result in growth without efficiency, diverting scarce electricity capacity from other social functions like mobility and heating (20), or reaching growth limits due to social, ecological, or technical constraints (21).

Indirectly, GenAI's electricity impact on the economy could surpass its direct impact, yielding a significantly different final electricity outcome. However, an assessment of GenAI's electricity impact would be incomplete without considering potential large efficiency gains in the economy offsetting increased GenAI electricity consumption. Currently, data centres account for about 2-3% of global electricity resources (22, 23)(Schneider Electric , 2024).

This implies that if AI were used to increase economic efficiency, its indirect impact could easily outweigh the additional electricity costs of GenAI in data centres. Thus, high electricity-consuming AI might still prove efficient for the economy, while low energy-consuming AI could paradoxically result in high economy-wide electricity consumption. The scenario where high energy usage of AI is fully compensated by a larger reduction in electricity consumption across the economy is predicted by the enablement hypothesis (18).

The situation where this is not realized—possibly due to increased consumption of existing products and services or more consumption of newly innovative products and services—is termed the global Jevons paradox (18). In the most optimistic case, more efficient usage of low-data AI will further reinforce a trend towards a lower energy-consuming economy. This is predicted by the electricity efficiency reinforcement hypothesis. The systemic narratives on AI impact assume that AI usage may result in other costs for the system as a whole, such as pollution, CO2 emissions, and climate change (9). The new opportunities of AI are part of a new fourth industrial revolution (24) that affects the operations of industries, governments, and society in a yet difficult-to-foresee future (25).

Regarding the systemic impact of AI, the electricity needed for AI and the AI-stimulated economy can be converted to a greenhouse gas emissions (GHGE) volume. Many expect that more renewable electricity will become available, and thus, despite a large growth in electricity usage, the total GHGE will decline (26, 27). This is predicted by the electricity renewables hypothesis. Few propose that the growth of renewables will not happen, but theoretically, the growth of electricity needs could be so high that it outpaces the growth of renewables. This narrative is the carbon footprint accelerator hypothesis. Finally, even if the renewables hypothesis becomes reality, further growth in economic demand could result in significant waste increase and material exhaustion. This prediction is the waste accelerator and exhaustion hypothesis. Alternatively, AI could contribute to less waste and exhaustion by helping find new sustainable products and reduce consumption—the waste and exhaustion reduction hypothesis. We summarize these narratives in Table 2.

More precisely focused on AI and ICT, we examine AI's impact on data center (DC) electricity consumption. H1 suggests that DC electricity consumption can be reduced by GenAI or low data AI (Symbiotic and sustainable AI) applied in data center operations, freeing more capacity for further growth of GenAI training and inferencing. Negatively (H2), a decrease in DC electricity consumption or its growth can also be realized by reaching certain constraints, such as lack of available electricity or hardware for growth. Increases in DC electricity consumption by AI may also realize a Jevons paradox (H3), which posits that efficiency gains will further accelerate the growth of AI-based DC electricity consumption, potentially unbounded. Alternatively, H4 states that AI's impact on DC electricity consumption can increase, but this growth may result in conflict over scarce electricity resources between data centers and other economic functions, like mobility or health centers. These data center electricity consumption scenarios can be reinforced or balanced as hypotheses 1-4 suggest, and can be further modeled and simulated using a system dynamics view (28–30), as we will explain further in this article.

Table 2: A classification of narratives and 11 related hypotheses on AI sustainability					
Direct impact		Indirect impact		Systemic impact	
AI reduces ICT electricity needs	1. H1. AI saves ICT by GenAI or low data AI 2. H2. AI causes ICT limits to growth; AI ICT stalled	Reduction of electricity needs for economy	5. Electricity efficiency reinforcement	Waste reduction Lower carbon footprint	8. Electricity renewables. 9. Waste & exhaustion reduction.
				Increased waste. Increased carbon footprint	10. Carbon footprint acceleration 11. Waste & exhaustion acceleration.
		Increase of electricity needs for economy	6. Global Jevons paradox	Waste reduction Lower carbon footprint	8. Electricity renewables. 9. Waste & exhaustion reduction.
				Increased waste. Increased carbon footprint	10. Carbon footprint acceleration 11. Waste & exhaustion acceleration.
AI increases ICT electricity needs	3. H3 AI Jevons paradox 4. H4 GenAI with little or no efficiency gains	Reduction of electricity needs for economy	7. Enablement	Waste reduction Lower carbon footprint	8. Electricity renewables. 9. Waste & exhaustion reduction.
				Increased waste. Increased carbon footprint	10. Carbon footprint acceleration 11. Waste & exhaustion acceleration.
		Increase of electricity needs for economy	6. Global Jevons paradox	Waste reduction Lower carbon footprint	8. Electricity renewables. 9. Waste & exhaustion reduction.
				Increased waste. Increased carbon footprint	10. Carbon footprint acceleration 11. Waste & exhaustion acceleration.

Research Questions

For many of these narratives, we lack conclusive evidence, and even worse, their logic is fragmented. In this article, we focus on the question of direct impact: What is the direct impact of AI on data centre electricity needs?

To answer this question, we conduct a broader literature search for statements that provide more precise and systematic insights into the system dynamics of AI's direct impact on DC electricity. This article focuses less on concrete empirical facts, such as current DC electricity consumption and its short-term projections, and more on developing a theory and model for answering "what if" questions about DC electricity volumes. More important than a forecast, this article codifies and integrates knowledge in the field of AI's direct impact on DC electricity by creating a model for reasoning about the future.

This approach further deepens the previous narratives through causal statements and data scenarios. A scenario is a configuration of values of independent variables that determine a sequence of events to help understand and prepare for possible future situations (31–33). Scenarios thus state how expected futures may become reality under a set of assumptions about future-determining variables (34, 35). More importantly for actionable knowledge, however, is the possibility of creating behavioral data scenarios, i.e., descriptions of the values of key variables over time, so that decision-makers know when specific decisions need to be made.

Derbyshire and Giovannetti (32) for example put emphasis on knowing when a certain critical threshold value will be passed and thus when an organization needs to have collected the resources and actions to avoid that the critical threshold will turn in an irreversible unwanted condition. An example may be the needs for electricity grid extensions, which needs to be based on a longer-term forecast of electricity demand because of the delay of about 5 to 8 years in upgrading a grid's capacity. Good data scenario's must be built by simulation models that are consistent with qualitative causal scenarios.

Given the inherent uncertainty about the future of social reality, simulation models are designed to empower its users to envision future data scenarios based on their own assumptions or newly acquired information. The simulation model by that also will enable the model's user to work with alternative and competing assumptions. By this the model may become a useful tool for scenario management and strategic decision making (35, 36). Although impact studies of the digital economy on environmental sustainability have been done before (20), also with many levels of details and for the full product life cycle (8), system dynamics modelling of it offer many opportunities of understanding future scenarios under the influence of technological innovations and possible policy interventions (37).

Future Studies

Studying the future is an important concern because knowing something about the future helps people make decisions. Natural science has shown great capabilities in predicting future events, resulting in highly predictable, if not lawlike, statements (38). For social and economic reality, knowing the future is more problematic; even if people's behavior can be predictable, it is not certain that such predictability will persist across different contexts and new realities (39).

Son (40) identifies the mid-1940s as the start of three stages of scientific approaches to the future. The first scientific stage contains methods and theories for forecasting, based on data and possible alternative futures that can be systematically explored in terms of their impacts. This first stage is gradually replaced after 1960 by future studies that included worldwide discourses on global futures and the development of normative futures with the involvement of the business community. This, for example, resulted in new global institutional norms and questions regarding the dominant discourse of unlimited growth in the mid-1970s. While the first stage was mainly technical forecasting and empirical, the second stage became more normative, either optimistic or pessimistic. The third stage began in the early 1990s with a focus on critical future studies and the identification of multiple competing future scenarios that can be synthesized towards joint strategic visions. In this article, we follow this third stage type of studies.

Chiasson et al. (41) argue that many forecasts say more about the current time than the future because they are beliefs or have been built on current data with the assumption that if everything stays the same, the future will be as predicted. As a set of normative beliefs at one moment, forecasts can be integrated into technological codes, like a set of appreciable features at one moment that will determine the reality for the next couple of years to come.

Alternatively, Chiasson et al. (41) state that the actual realization of technologies and their environments may also be a matter of involvement of intended users and stakeholders. Therefore, the actual shaping of the future will be the outcome of a more or less democratic rationalist process. Many future studies also attempt to predict how the future will look given a current understanding of it. This is especially a popular way of thinking from a positivist perspective, where a belief in lawlike statements is the foundation for knowing what will happen in the future (39, 42). Finally, Chiasson et al. (41) identify future studies that explore future potentials or opportunities, sometimes suggesting multiple scenarios of what can happen. These scenario studies emphasize the likelihood of wanted or unwanted futures and provide means for decision-makers to act appropriately when one of the scenarios seems to materialize. Table 3 summarizes these views on futures studies.

Although creating visions for the future needs some factual basis rooted in current knowledge, we believe that, especially for highly innovative processes like AI adoption, innovations may cause disruptions in trends. Thus, a focus on "now" may result in misleading insights. For technological adoptions and innovations, we believe that they are caused by human intentions, investments, and willingness to adopt. Therefore, in the context of our study, the future is more the outcome of change processes than a fact caused by technological codes or lawlike insights. This leads us to choose "potential scenarios" as the most relevant approach to studying AI impact. These scenarios may include nonlinear trends and reinforcement and feedback mechanisms over time, which are key assumptions of system dynamics (29, 44).

Table 3: Epistemologies for future studies, based on Chiasson et al (Chiasson et al., 2018), updated by Schneider Electric

	Future as a fact	Future as an normative process
Focus on now	Technological codes Interpretive studies to uncover the code	Democratic rationalism A democratic process involving many actors with different views who present and debate each other's scenarios to create actional knowledge for decisions
Focus on the future	Forecasting Positivist studies of finding facts and valid predictive (lawlike) models to predict (extrapolate) the future	**Potential scenarios (focus of this article)** Multiple scenarios as outcomes of research and debate with available data as input to predict future outcomes. These outcomes are evaluated used as new inputs for prediction or for model revisions (Wijnhoven et al., 2024).

Understanding the Future with System Dynamics

Understanding the future with system dynamics

Freitag et al. (17) posit that the direct energy consumption effect of the ICT sector may grow due to larger volumes of ICT use and more intense (AI) processing. This trend may be balanced by an increase in efficiency of the ICT sector through more energy-efficient processors and better-designed software systems. This efficiency growth, however, may stimulate demand for ICT, potentially resulting in a reinforcement of energy consumption by ICT, known as the Jevons paradox of ICT (37). Both these reinforcements of electricity consumption and the balancing feedback mechanism through efficiency improvements are system dynamics processes (38), i.e., "... the reciprocal and temporal causal mechanisms that underlie many complex and dynamic systems ..." (39). By causally modeling these system dynamics processes, we will be able to virtually answer "what if" questions about possible futures that are difficult to create and experiment with in a non-virtual way.

Even if there is an electricity consumption-reinforcing Jevons paradox for the ICT sector, AI may indirectly help reduce a much larger source of electricity consumption and emissions in industry (this is Freitag et al.'s "enablement" scenario). Alternatively, it may also result in a multiplier for more industry activities and greenhouse gas emissions (GHGE) (this is Freitag et al.'s Global Jevons paradox scenario). System dynamics enables the analysis of the non-linear direct and indirect longer-term effects of AI, which also includes possible longer-term rebound effects of short-term positive actions. For example, the Club of Rome studies showed the longer-term exhaustion of the planet and longer-term degradation of health and wealth effects of economic growth (40).

Although the original Club of Rome report has been criticized for its lack of consideration of possible technological solutions, recent IPCC reports do point to the system dynamics of mutually influencing factors like emissions and population growth on long-term temperature and resulting disruptions of the global ecosystem.

These reports have been accepted by many policymakers as a foundation for their sustainability policies (41). Understanding and visualizing these longer-term Jevons and rebound effects is difficult for decision-makers without a model that can predict possible long-term effects of many variable interactions.

Valid system dynamics models may be able to simulate future scenarios, presenting trends over a certain time period. This information can be useful for decision-makers and may also simulate the outcomes of alternative choices. System dynamics models are not forecasts in themselves, but rather models into which alternatively chosen expectations (i.e., scenarios) can be inserted and used to calculate longer-term outcomes if these expectations were to be true (42). Given possible views on future trends, system dynamics projections may also be useful for decisional actions or identifying constraints to growth to prevent worst-case scenarios from becoming reality (43). By inserting values and formulas into the variables, flows, and stocks, we can simulate outcomes over multiple time periods to observe their system dynamics impacts. With this model, we can also run multiple scenarios and analyze the impact of different growth rates in smart industries and the effect of growing artificial intelligence and machine learning workloads connected to smart industry applications in data centers.

Future study research designs

For studying the future impact of some intervention in the social world, multiple research approaches are possible, linking the intervention (independent variable(s)) with the dependent variable(s) in different ways. Often, multiple independent and dependent variables are connected causally, as shown in Table 4.

Our study of AI and other data center (DC) services' impact on future DC TWh consumption fits the lower left italicized quadrant, but additionally has a diachronic nature (44) that is common to system dynamics.

Table 4: Independent-dependent variables designs

	Single dependent variable	Multiple dependent variables
Single independent variable	(Quasi)-experimental design; the study of the impact of one intervention (e.g., new production facility) on one variable (e.g., costs reduction)	Technology assessment studies; i.e., one new technology is assessed in multiple possible impacts, like labour productivity, ethics, profit, sustainability
Multiple independent variables	Impact of a configuration of independent variables (e.g., a new socio-technical design of a workplace) on one dependent variable (e.g., work motivation)	The impact of configurations of independent variables (e.g., a redesign of electricity markets and their price settings) on a configuration of dependent variables (e.g., total electricity consumption, electricity prices, sustainability, peak capacity)

Creating system dynamics future models

System dynamics models can be created using existing data but will likely go beyond data by incorporating literature, experience, and (creative) reasoning to identify causal mechanisms behind observable behavior (51). By understanding these causal mechanisms, it can answer "what if" questions about possible futures. In the system dynamics methodology, a problem or system is represented by a causal loop diagram (i.e., a causal scenario) that captures its structure and interactions. The causal effects may be positive, with reinforcing feedback loops strengthening a given trend, or negative, with balancing trends of negative reinforcements. Both feedback loops may act simultaneously or at different times (e.g., as rebound effects that manifest themselves at a later stage) and they may have different strengths. Because the system dynamics method is primarily a technique for business and policy simulation (37, 52, 53), its focus is not on generating precise point-predictions of the future, but rather on creating models that are useful for thinking about the future by combining possible trends and interactions of influencing variables (51).

System dynamics identifies balancing and reinforcing patterns of change both in influences on an existing stock and as feedback or rebound effects. Comparable to the water level in a lake, a system dynamics model views a stock as the lake and identifies inflows and outflows from this stock. In a balanced state, the volume of inflows equals the volume of outflows. Through reinforcements, both the inflow or outflow speed can be accelerated, resulting in an increased or decreased water level in the lake. As a secondary effect, feedback mechanisms can inform the volume of inflow for the purpose of balancing or reinforcing, while feedforward can inform the level of outflows needed for balancing and reinforcement (54).

Applied to policy research, system dynamics research starts with problem articulation and ends with policy design or policy recommendations based on 'what if' analyses. Duggan (52) and Morecroft (44) articulate three other inter-related activities in system dynamics model building.

The first activity is dynamic hypothesis creation that identifies relevant stocks, flows, and relations for a problem. The second is simulation model building, in which initial states of the system (i.e., parameters and variables' assumed values) are inserted into the model to determine their joint impact. The third activity is testing of the simulation model, through which inconsistencies and consistency with existing knowledge and data can be checked. Morecroft (44) defines a dynamic hypothesis as a "... preliminary guess at the sort of relationships likely to explain a given pattern of behaviour through time". Dynamic hypotheses need to be updated if simulation outcomes indicate inconsistencies with logic and facts. However, in a system dynamics study, not all hypotheses will always be tested, as some will be proposed for follow-up studies. The structure of the system dynamics model itself is already a useful research outcome, comparable to an explanatory or predictive theory.

The different activities of a system dynamics study are presented as a sequence in Figure 1, but in practice, these activities will go through many iterations.

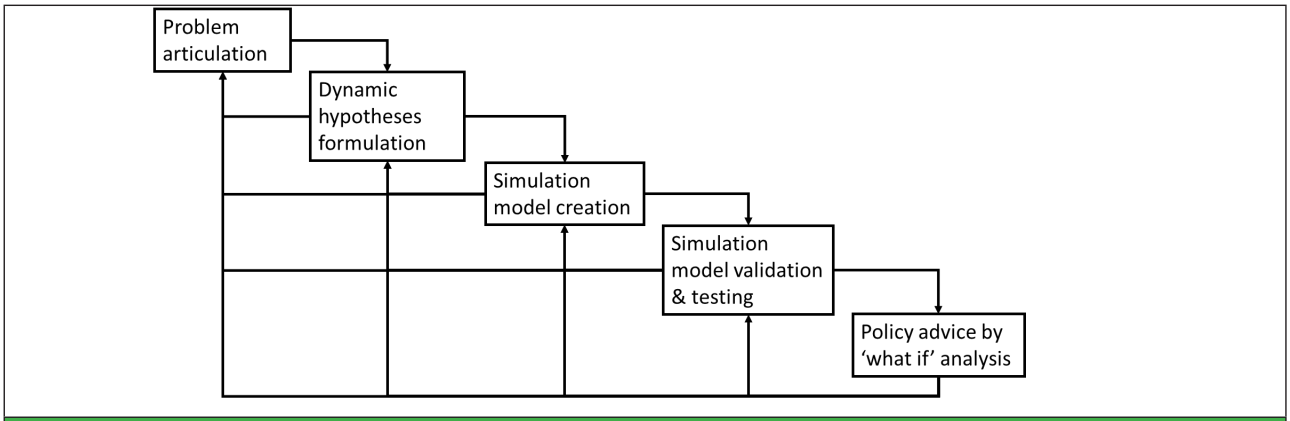


Figure 1. Steps in the system dynamics modeling method (Duggan, 2016)

These steps are performed as follows:

Problem articulation through a literature study and expert interviews.
Dynamic hypotheses: Following Freitag et al.'s 4 main hypotheses.
Simulation creation: Using principles of system dynamics and the InsightMaker.com tool.
Model validation: Triangulation of simulation outcomes with Schneider Electric studies and other sources, corrections of model errors in 10+ rounds.
Policy advice: Provided after correct model outcomes are achieved.

For the simulation model creation, we depend on a combination of top-down and bottom-up approaches. We use a top-down approach to remain relevant for the macro variables of importance to C-level executives and external decision-makers.

We employ a bottom-up approach to be able to use the more detailed analyses that have been recently published. We combine these approaches and delve into the black box of Generative AI (GenAI), the focus of this study, while keeping traditional AI and traditional data center (DC) services as a black box by relying on the expertise of experienced DC managers and analysts.

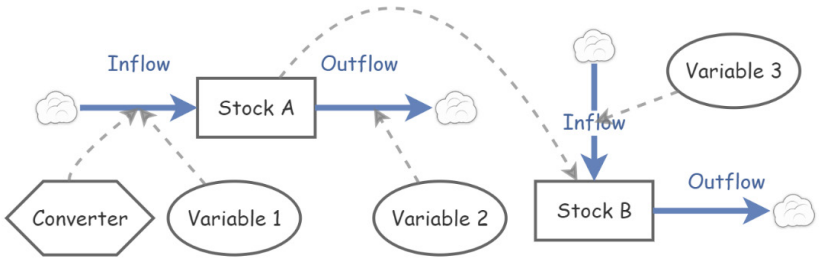
Creating system dynamics future models

As part of the creation and testing of a simulation model, system dynamics software is important. Hristosky and Mitrevski (55) state that there are many software tools for system dynamics modeling, but they argue in favor of InsightMaker (which we use throughout this study) because it is a "...free-of-charge, Web 2.0-based, multi-user, general-purpose, online modeling and simulation environment, completely implemented in JavaScript, which promotes online sharing and collaborative working. (...)

To the best of our knowledge, it is the first, and yet the only free-of-charge Web 2.0-based Internet service that can deliver a plethora of advanced features to its online users, including Causal Loop Diagrams, Rich Pictures Diagrams, Dialogue Mapping, Mind Mapping, as well as Stock & Flow simulation" (Hristoski & Mitrevski, 2016, p. 5).

System dynamics, as a causal modeling language, has multiple primitives to express causal relations. One type of primitive is the stock, which represents entities that accumulate or deplete over time. Related to stocks are flows, which add to or deplete a stock. The third type of primitive is the variable, which defines the speed or volume of flows. InsightMaker represents stocks graphically by rectangles. Flows are represented by solid lines with a directed edge that indicates the direction of the content flows. Variables are graphically portrayed by ovals; they can be dynamically calculated values that change over time (governed by an equation), or they can be constants (fixed values), e.g., e-Customer arrival rate.
Links in InsightMaker connect variables with each other and they connect variables with flows and stocks. Links are graphically shown by dashed lines with a directed edge in the model. The InsightMaker primitives and representation objects are summarized in Figure 2.

InsightMaker primitive objects



Primitive descriptions

1. Stock (box): Contains a volume of something that can increase or decrease over time.
2. Flow (filled line with edge): Includes incoming flows that add to a stock's volume and outgoing flows that reduce a stock's volume.
3. Variable (oval): Gives the strength or volume of an inflow or outflow. Variables can also mutually contribute to each other. They can contain fixed values or values that change over time.
4. Link (broken line with edge): Links variables with flows and possibly also links stocks to reuse information from one stock as a value for the volume of another stock.
5. Converter (hexagon): A data set with values that change over time.
6. A cloud represents the environment of the system.

Figure 2: Description of InsightMaker's modelling primitives

Creating system dynamics future models

Using this system dynamics modeling "language", we structure our problem of direct GenAI impact on electricity usage of data centers (DC), indirect Large Language Model (LLM) effects on the economy (growth and electricity usage of the economy), and systemic impact of AI (CO2/GHGE and waste). For this, we present Figure 3, which shows stocks, flows, and variables for the AI direct, indirect, and systemic impact at the highest level of abstraction.

Following system dynamics reasoning, the volume of a stock, such as DC electricity consumption, can be reinforced by an increase in AI usage (training and inferencing; which corresponds to H4) and further reinforced via a rebound effect. This rebound effect occurs because of the extra opportunities and demand for AI as a consequence of increased efficiency of DCs, which frees DC capacity for new innovative AI applications (this is H3 and the AI Jevons paradox).

In addition to reinforcements, system dynamics also recognizes balancing effects, which aim to control the total level of electricity consumption of the DC through increased efficiency via Sustainable AI that saves capacity for the DC (this is H1 in Table 1).

There may also be a possible limitation in AI demand that slows down (i.e., balances) the reinforcements due to AI and DC constraints (this is H2, the Jevons stalled). These constraints may be a rebound from the size of DC electricity utilization but may also be caused by limitations in hardware manufacturing or ecological constraints.

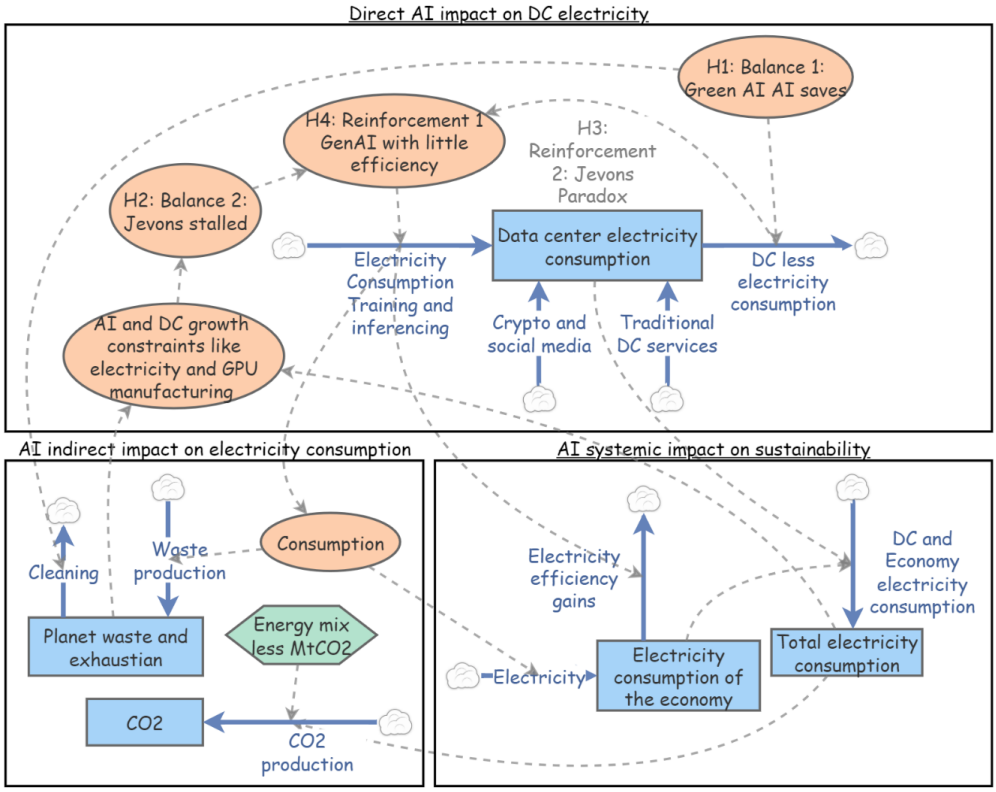


Figure 3: High level structural system dynamics model of AI impact

Note that this model only presents the main variables as stocks, flows, and converters, along with their relations (flows and links). The model does not present data or behavioral patterns such as volumes and the relationships between inflows and outflows that are part of a behavioral model.

Scenarios Descriptions and Potential Impacts

Each of our scenarios has behavioral and managerial implications that we summarize below.

H1: “Symbiotic-sustainable AI Revolution”

In this scenario, AI-driven innovations in energy efficiency and resource optimization lead to a significant reduction in the ICT sector’s electricity consumption. Sustainable AI advocates successfully promote sustainable practices, resulting in widespread adoption of energy-efficient algorithms, hardware, and data center designs. The synergy between AI and renewable energy technologies creates a virtuous cycle, where AI enhances the efficiency of clean energy systems, which in turn power more sustainable AI development. This scenario sees AI as a solution to ICT’s energy challenges rather than a problem, suggesting that AI adoption in data centers will lead to reduced energy consumption through increased efficiency and optimized resource allocation. The key hypothesis is that widespread adoption of sustainable AI technologies will lead to symbiotic growth in the ICT sector, balanced with environmental resources and electricity availability. Major tech companies focus on energy-efficient AI solutions, and method-driven AI approaches improve operational efficiency and resource optimization within data centers. Importantly, this scenario prioritizes responsible AI development and reduced carbon footprint in data centers, driven by the intentional control of DC electricity consumption by developers and engineers, ultimately presenting a future where AI technology and sustainability goals are harmoniously aligned.

H2: “Limits To Growth”

In this scenario, shaped by Demand Dynamics Analysts and Regulatory Proponents, AI capabilities expand but encounter natural limits in energy availability, computational resources, and regulatory constraints. This leads to a more measured and sustainable growth trajectory for AI and data centers, inspired by the work of researchers who posit that while demand for AI services will grow, data centers will face limitations. The key hypothesis is that the AI industry will experience various growth constraints, resulting in a more balanced and controlled economic model. As data centers confront expansion limitations, there’s a shift towards cautious investments and controlled AI innovation. The focus moves to application-driven AI and traditional computing methods that are less resource-intensive, as rising costs make large-scale, energy-intensive AI projects less feasible. Regulatory frameworks evolve to balance innovation with environmental and social concerns, prioritizing efficiency and sustainability over rapid expansion. This scenario’s dynamics hypothesis involves a balanced control of DC electricity consumption, mainly forced by external factors, resulting in a controlled expansion of AI technologies that aligns with broader sustainability goals. Ultimately, this measured approach leads to a more stable, albeit slower, growth trajectory for AI and data center technologies, addressing social and environmental concerns while maintaining progress in the field.

H3: “AI Abundance Without Boundaries”

This scenario embodies the Jevons Paradox, where improvements in AI efficiency paradoxically lead to increased overall energy consumption. Inspired by the Jevons paradox concept, it posits that AI and Data Centres will grow without barriers, with efficiency gains accelerating AI growth and consumption of DC electricity. Techno-optimists drive rapid AI deployment across all sectors, believing that technological advancements will solve any resource constraints. The key hypothesis is that massive investments fuel continuous, unrestrained growth in AI and data center technologies, with no significant barriers to expansion. As the increased efficiency of AI systems lowers computation costs, an explosion in AI applications and data center proliferation occurs. While individual AI systems become more energy-efficient, the total energy consumption of the AI sector grows dramatically due to massively increased usage and new applications. This scenario is characterized by aggressive AI investment strategies and a strong belief in technological solutions to overcome potential limitations. The rapid economic growth fueled by AI advancements leads to increased income inequality as benefits are unevenly distributed. Environmental concerns are largely overlooked in favor of technological progress, potentially resulting in ecological degradation. The reinforcing mechanism of investment-driven growth and efficiency improvements creates a cycle of ever-expanding AI capabilities and data center infrastructure, embodying a dynamic hypothesis of accelerated reinforcement of electricity consumption by AI.

H4: “AI Energy Crisis”

This scenario, shaped by Alarmists and Regulatory Proponents, anticipates and reacts to potential “black swan” events in AI energy consumption. Inspired by research highlighting the risks of unchecked AI growth, it posits that AI and Data Centers will expand beyond a certain threshold, conflicting with other critical electricity functions in the economy. The key hypothesis is that the rapid growth of AI and data centers reaches a tipping point, triggering a cascade of negative consequences. As AI’s electricity demand begins to compete with other essential sectors, it leads to an unforeseen energy crunch, resulting in economic downturns and severe operational challenges for AI-dependent industries. This overexpansion causes significant financial losses and operational difficulties as electricity demand outstrips supply or becomes prohibitively expensive. Regulators scramble to implement strict controls on AI development and deployment, while researchers grapple with a “data crunch,” struggling to balance the need for massive datasets with energy constraints. The economic downturn leads to substantial cuts in AI investment, making many existing business models less viable. This scenario’s dynamic hypothesis suggests a difficult-to-control reinforcement mechanism that may ultimately result in a crisis, highlighting the potential risks of unchecked AI growth and the critical need for proactive risk management in AI development and deployment.

Schools of Thought: Perspectives on AI Impacts

There are multiple exogenous and endogenous data center (DC) factors that influence each of these scenarios in different ways and thus are potential "interventions" that can influence the trends of each scenario. Exogenous factors are "outside the Data Center," which typically refer to power availability, chip manufacturing capacity, supply chain resiliency, infrastructure deployment life cycle, data scarcity, latency walls, AI investment dynamics, and regulations. Endogenous factors are "inside the Data Centers" and refer to hardware efficiency, software efficiency, algorithmic progress evolutions, computing power, rack and server efficiency evolutions, and quantum computing development.

By inserting values and formulas into the variables, flows, and stocks, we can simulate outcomes over multiple time periods to observe their system dynamics impacts. With this model, we can also run multiple scenarios and analyze the impact of different growth rates in smart industries and the effect of growing artificial intelligence and machine learning workloads connected to smart industry applications in data centers.

A scenario that follows H1 Sustainable AI saves DC electricity may also be called a Sustainable AI advocacy scenario. For H2, the possibility that AI development will be constrained and stalled may also be referred to as a Black Swan scenario, as the initial AI optimism may turn into more realism and perhaps pessimism due to resource limitations. The H3 AI Jevons paradox may be the outcome of a techno-optimism scenario, as the efficiency gains of DC by AI will be fully used to try out new AI ideas and further stimulate the concept of AI development without constraints. H4 forecasts a scenario of continued unlimited AI growth that at some point will be so high that other electricity usages will be halted. This scenario is the alarmism scenario.

H1 Core Factors

Endogenous Factors
• Hardware Evolution: Development of highly efficient, AI-specific hardware that significantly outperforms general-purpose processors
• Software Optimization: Widespread adoption of AI-driven energy management systems in data centers
• Algorithmic Efficiency: Breakthrough in neuromorphic computing leading to drastically reduced energy consumption for AI tasks
• Data Center Design and Architecture: Implementation of AI-optimized cooling systems that reduce overall energy requirements

Exogenous Factors: Regulatory Landscape: Global standards for Sustainable AI certification driving industry-wide adoption of energy-efficient practices for the realizing efficiency gains, and economic demand or increased of the volume of data center usage.

H2 Core Factors

Endogenous Factors
• Data Center Design and Architecture: Emergence of decentralized edge computing reducing reliance on large-scale data centers
• Algorithmic Efficiency: Shift towards "small data" and transfer learning techniques to reduce computational requirements

Exogenous Factors
• Geopolitical and Economic Conditions: Resource scarcity (e.g., rare earth elements) constraining hardware production and data center expansion
• Regulatory Landscape: Implementation of strict regulatory frameworks limiting data center energy consumption
• Market Demand: Public pressure and corporate responsibility initiatives leading to voluntary limitations on AI energy use

H3 Core Factors

Endogenous Factors
• Hardware Evolution: Development of room-temperature superconductors revolutionizing data center energy efficiency
• Data Center Design and Architecture: Creation of a global, high-capacity dark fiber network enabling more efficient distributed computing
• Algorithmic Efficiency: Advances in bio-computing allowing for ultra-low power AI operations

Exogenous Factors
• Technological Advancements: Breakthrough in quantum computing making certain AI tasks exponentially more efficient
• Energy Sector Developments: Widespread adoption of AI-driven fusion reactors providing abundant clean energy
• General beliefs in AI opportunities
• Willingness to invest in AI.

H4 Core Factors

Endogenous factors
• Technology push and AI developers with limited perspective outside of their own project
• Powerful AI industry that can claim scarce electricity resources
• Problem of the electricity production and distribution networks of scaling up their productivity. As stated before (Figure 2), system dynamics studies the temporal changes of a stock under influence of inflow and outflows of theses stocks, whose volumes and speeds can be speeded up or delayed by influencing variables. Each of the four hypotheses and scenarios described above, are based on different generic system dynamic mechanisms, which describe below.

Exogenous Factors
• Geopolitical and Economic Conditions:
• Severe climate events disrupting power grids and forcing limitations on non-essential energy use
• Global energy crisis leading to strict rationing of electricity for data centers
• Geopolitical conflicts disrupting global supply chains and limiting access to necessary hardware components
• Regulatory Landscape:
• Cybersecurity concerns resulting in mandatory air-gapping of critical infrastructure from AI systems
• Public backlash against AI's environmental impact leading to restrictive legislation

Key System Dynamics Mechanisms

System dynamics studies the temporal changes of a stock under the influence of inflows and outflows, whose volumes and speeds can be accelerated or delayed by influencing variables. In our study, this stock is the volume of DC electricity used (see the highest listed model in Figure 4). This stock increases due to greater demand for AI and DC services and decreases because of efficiency improvements in DC and efficient methods of AI training. Both these increases may be "normal" or accelerated by the influence of exogenous factors outside of the DC.
Figure 1 presents, for each "school of thought" and related scenario, their main endogenous (what happens inside the DC) and exogenous (outside DC) factors. Each of these scenarios is further detailed in the main report for the different components of DC electricity consumption (i.e., GenAI training and inferencing, Traditional AI training and inferencing, and Traditional DC services) and the main influencing factor per component.

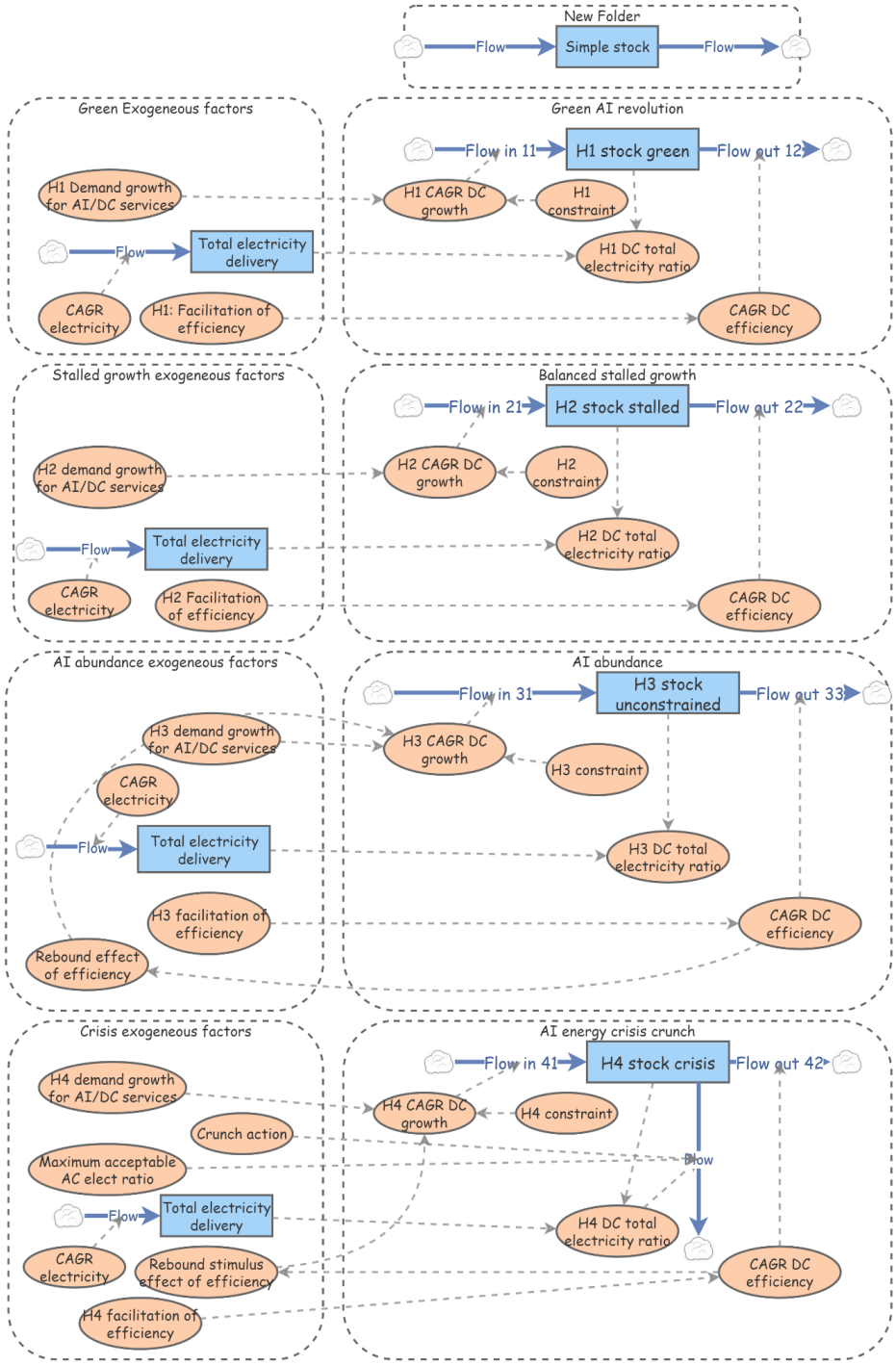


Figure 4. Main system dynamics models for schools of thought. See legend in Figure 2

Top-down and bottom up model integration

Multiple authors have been developing models and estimates of electricity consumption of GenAI from different angles that can be summarized as top-down versus bottom-up approaches. Sometimes top down and bottom approaches are presented as competing alternatives (56, 57), but we argue that both approaches serve different stakeholders in their decision making in different levels of uncertainty (58–60).

Top down approaches start with viewing the whole system, aims at solving the problems that are present for the whole system, and may go in more depth by one or more subsystem layers for analyzing details per subsystem, especially those subsystems that may cause the largest problems for the whole (61, 62). Top-down approaches are applied in strategic management and investment strategies where the focus is on macro variables about the organization (like ROI, revenue and strategic plans) and its environment, like market cycles, public policy impacts, environmental and ecosystem impact, and technologies (63). Measures thus focus on broad corporate and economic trends. Alternatively, bottom-up management approaches start with local or company-specific variables, identify individual opportunities, needs and requests, but may miss broader strategic and market trends (64).

The tensions that may exist between top-down approaches and bottom-up approaches are often manifest in conflicts between the managerial apex and work floors. Middle-up-down principles have been development, in which middle management has the role of aligning visions of the top with opportunities and needs of the work floor (60). The main variables for the top-down approach are related to the effectiveness of the system as a whole by which longer term demand views are well connected to longer term supply opportunities with high ROI as an outcome. For bottom-up approaches, the main variables are efficiency related, i.e., output divided by resources and quality as a key factor that determines the level of the output/input efficiency measure.

The output/input indicators can be calculated using existing data from the organization's bookkeeping system, and are mainly historic data that can be used to predict outcomes in the short term with time series analyses. For the top-down indicators, macroeconomic measures about market developments from both the supplier side as well as the demand side are important. The data needed are especially about the future, but are mostly not available, and thus key metrics are based on scenarios.

Masanet et al. (65) state that two decades of data center energy analysis have demonstrated that "bottom-up" modeling on the basis of installed IT hardware stocks, cooling infrastructure, and operating characteristics is the most reliable and transparent approach for estimating electricity use (22). Moreover, because of their technological details, such models can also generate "what if" scenarios that explore how changes in key drivers, such as the numbers, types, and locations of deployed AI servers, could affect future electricity demand. The bottom-up approach to DC electricity consumption thus calculates the total rack power usage in watts for workload (e.g., dedicated AI servers), storage, and network devices, times the number of hours that these servers run, multiplied by a PUE for infrastructure. These calculations could be done if the data were available and reliable. Masanet et al. (65) also state that these data are often not there because of a lack of incentives for DC companies to provide them. However, multiple attempts have been made to go into even more detail, i.e., the estimation of AI training and inferencing electricity costs for data centers.

Going down to the elementary steps in AI training and inferencing, GenAI/LLM applications have large training efforts of processing large data sets and realizing models with billions of parameters with high volumes of inferencing. The workload related to training and inferencing depends on the size of the data set to be processed, the number of hyperparameters in the models, the number of prompts, and the size of the output they deliver (16, 66).

Table 5: Criteria for comparing Top down and Bottom up approaches

Criterion	Top down	Bottom up
Goal	Long term alignment of system with its supply and demand environment	Short term ability to deliver in an efficiency way
For whom	Strategic management	Operational management
Main variables	Macro economic, like ROI and market size	Operational, like costs and efficiency
Estimates and data	About the future based on scenarios	About the past from reporting systems
Opportunities and weaknesses	Long term alignment going into a not yet existing reality with assumption instead of facts	Short alignment based on data from the past from which we do not know when they will become invalid

Top-down and bottom up model integration

Starting from the dependent variable of total electricity for training of a GenAI/LLM measured in watt-hours (GWh), we delve into more details on variables that cause the volume of electricity needed. From a bottom-up perspective, this is a direct result of the calculated Model Electricity multiplied by the Power Usage Effectiveness (PUE) of the DCs that train the GenAI/LLM, which for an advanced data center is close to 1.5 and 1.10 (16). The electricity needed to train the LLM is the outcome of multiplying the GPU's thermal design power (TDP) used for training by the Training Time. TDP is regarded as a decent approximation for the power required during model training (67).

Training Time represents how long it would take to train the model on a single GPU - measured in hours. This is calculated by dividing the Total Compute by the Actual GPU Speed (in hertz, that is per second) and transforming from seconds to hours by further dividing by 3600 (as there are 3600 seconds in an hour). The Actual GPU Speed is the GPU Speed reported by the manufacturer (mostly NVIDIA, which has 95% of the GPU market), multiplied by the Utilization Rate. The Utilization Rate shows how efficiently the GPUs are actually used. Research shows that this value is usually between 30-70% depending on the model and data center architecture (https://lambdalabs.com/blog/demystifying-gpt-3).

AI tasks usually require floating point operations (FLOPs) at different precision levels (64 bits, 32 bits, and 16 bits). Occasionally, when very well optimized, they may also make use of tensor cores, i.e., specialized hardware in GPUs that can perform mixed precision calculations, such as combining 16-bit floating point precision (FP16) and FP32 (https://lambdalabs.com/blog/demystifying-gpt-3). GPUs have different speeds for the different precision levels. In practice, models use all of these speeds at different parts of the training (https://lambdalabs.com/blog/demystifying-gpt-3).

The Total Compute is the number of FLOPs required to fully train the GenAI/LLM. It is calculated by multiplying the Training Steps by the number of FLOPs per Step. The number of Training Steps represents the total number of times the weights (parameters) of the model are updated. It is calculated by multiplying the Dataset Size (expressed in tokens) by the number of Epochs. An epoch is one complete pass of the training dataset through the algorithm and shows how many times the dataset is used to train the model. This is one hyperparameter (a setting) of the training model.

Finally, the number of FLOPs per Step represents how many operations have to be performed in a Forward Pass and Backward Pass of the model. To get this value, the number of Model Parameters is multiplied by the number of FLOPs per parameter per token and divided by 1,000 to express it in TFLOPs. The number of FLOPs per parameter per token can be approximated using OpenAI's LLM scaling law (68) to a value of 6.

GenAI/LLM training thus has 6 input values to estimate GenAI/LLM training electricity costs per model: PUE, GPU TDP, GPU Speed, Epochs, Dataset Size, and Model Parameters. The final result is the Total Electricity. The list of the formulas and the corresponding variables is listed here:

- Total Electricity: $Model\ Electricity \times PUE$
- Model Electricity: $GPU\ TDP \times Training\ Time$
- Training Time: $Total\ Compute \div Actual\ GPU\ Speed \div 3600$
- Actual GPU Speed: $Utilization\ Rate \times GPU\ Speed$
- Total Compute: $Training\ Steps \times FLOPs\ per\ Step$
- Training Steps: $Dataset\ Size \times Epochs$
- TFLOPs per Step: $(FLOPs\ per\ parameter\ per\ token \times Model\ Parameters)*1012$

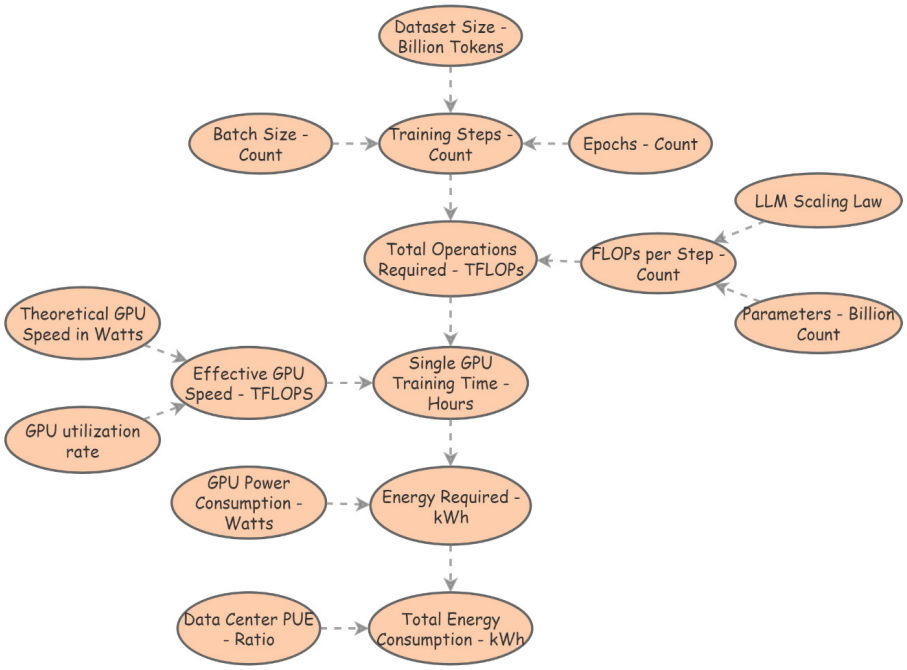


Figure 5. Causal model of electricity consumption of training

Top-down and bottom up model integration

For AI inferencing, the electricity usage per prompt depends on (11):
• Size of the output: number of tokens
• Energy consumption per token
• The type of output token production tasks: i.e., mainly classification versus generation, where the latter is about 10 times more energy-intensive
• The type of LLM architecture: i.e., seq2seq or decoder-only (like GPT) is about 1.5 to 2 times more electricity-consuming

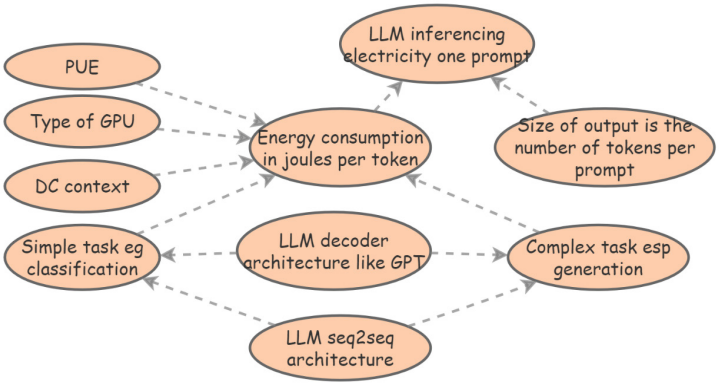


Figure 6 GenAI/LLM inferencing per prompt

The top-down approach for DC electricity usage focuses on an outside-in perspective and starts with estimations of service demand by industry and individual users. This approach also has a longer-term future view and thus is interested in annual growth (CAGR) based on scenario assumptions. Finally, top-down approaches can have a multi-layered perspective, i.e., global versus regional versus local. These approaches confront demand with electricity supplies at global, regional, or local levels.

Both the estimates for the future and the translation of service demand to electricity needs are highly uncertain. Regarding this, one has to rely on scenarios, which are possible futures based on expert insights like those of DCoF and best wishes (69, 70). In the top-down approach, the total electricity need of AI depends not only on characteristics of the DC itself but especially on the frequencies and intensity of use of these facilities for training and inferencing. For longer-term forecasts, DCs may be confronted with resource limitations (such as lack of power, hardware, and people) that can be partially relieved by Sustainable AI methods (71, 72), thus enabling even more use of the limited DC resources (i.e., Jevons paradox (45)).

This causal model can be presented in a top-down AI training and inferencing model, where consumer demand and business demand growth volumes determine the total electricity used for training and inferencing by the DC. Additionally, the DC also provides other services like social media, business applications, and cryptocurrency mining. This total electricity consumption is drawn from the power available from the grid.

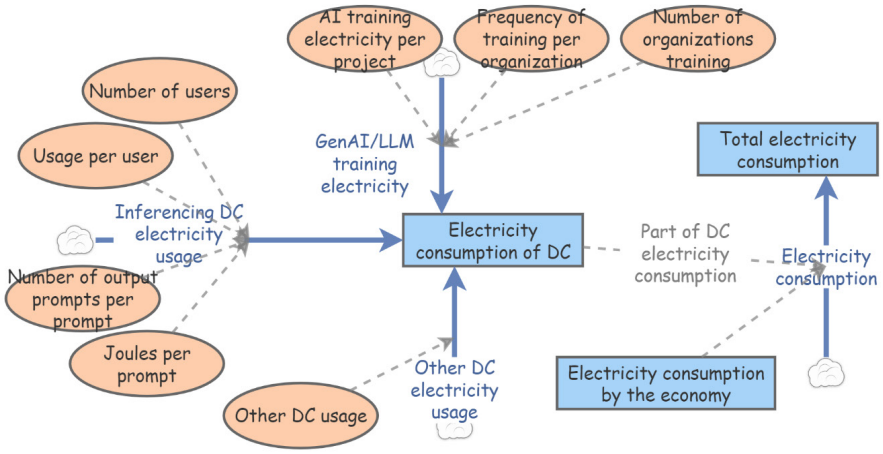


Figure 7. Top down AI electricity causal diagram

Top-down and bottom up model integration

We propose that bottom-up and top-down deliver different but complementary views on reality. These complementarities are summarized in Table 6.

Table 6: Top down and bottom up		
	Top down	Bottom up
Goal	An outside-in perspective	An inside-out perspective
For whom	Top executives, grid suppliers, external infrastructure decision makers	Operational DC managers and workload, network or storage customers
Main variables	Consumer and industry demand; market forecasts; type of applications; DC demand translated in electricity volume	Workload and server utilization, storage and number of storage servers, networking and network devices, total rack power usage, hours, PUE
Estimates and data	Scenarios about the future	Registered data and time series
Opportunities and weaknesses	Future looking with little empirical evidence	Status and short term looking with unclearly valid data from the past.

We integrate top-down and bottom causal models in figure 8 where we keep some details aways to avoid an overcrowded model.

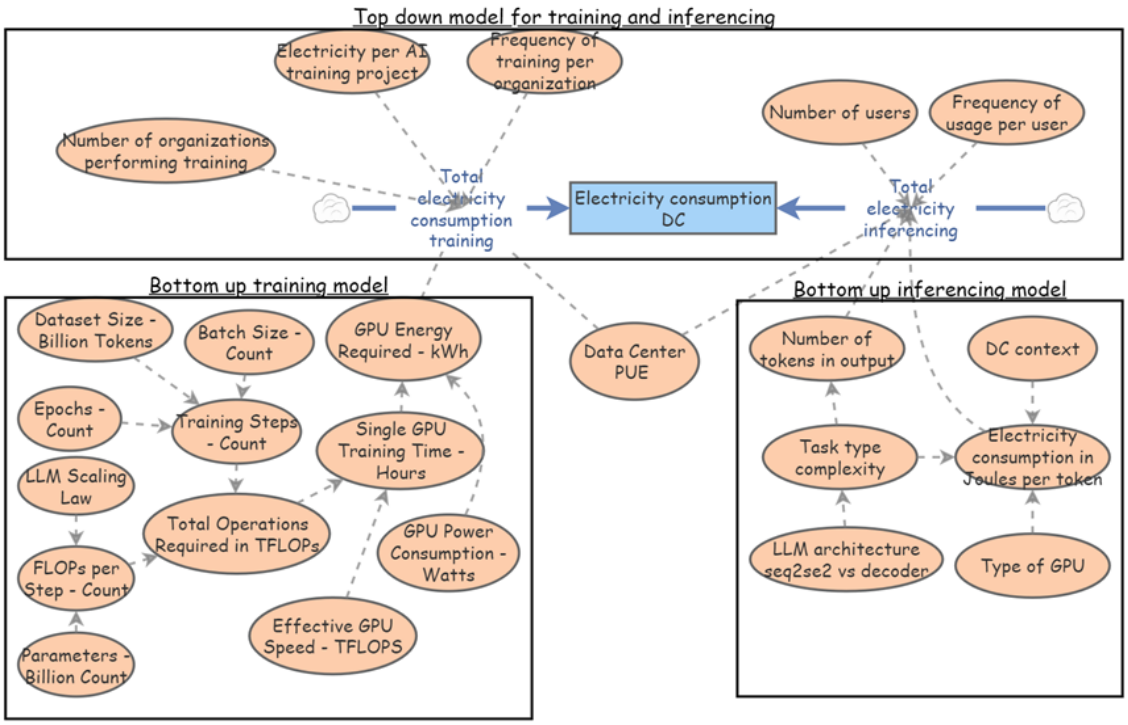


Figure 8: Integrated top-down-bottom up causal model

Note that this model is not yet a system dynamics model as it does not include time and possible reinforcements and balances dynamics.

Validating system dynamics models

Because we apply system dynamics to possible future scenarios, data are not always available, or if they exist, the disruptions of modern ICT make it risky to rely on them. Therefore, we aim at more theory-based models and validations via triangulation instead of empirical tests.

Following the work of Denzin et al. (51, 52) and Wijnhoven et al. (53–55), scientific and practical insights can be triangulated in four ways: 1) via comparison with other data, 2) by involving multiple alternative human investigators, 3) by applying different theoretical foundations as alternative views on the domain, and 4) by applying different research methods. In Table 6, we give short descriptions of the triangulation methods and discuss their usefulness for this study.

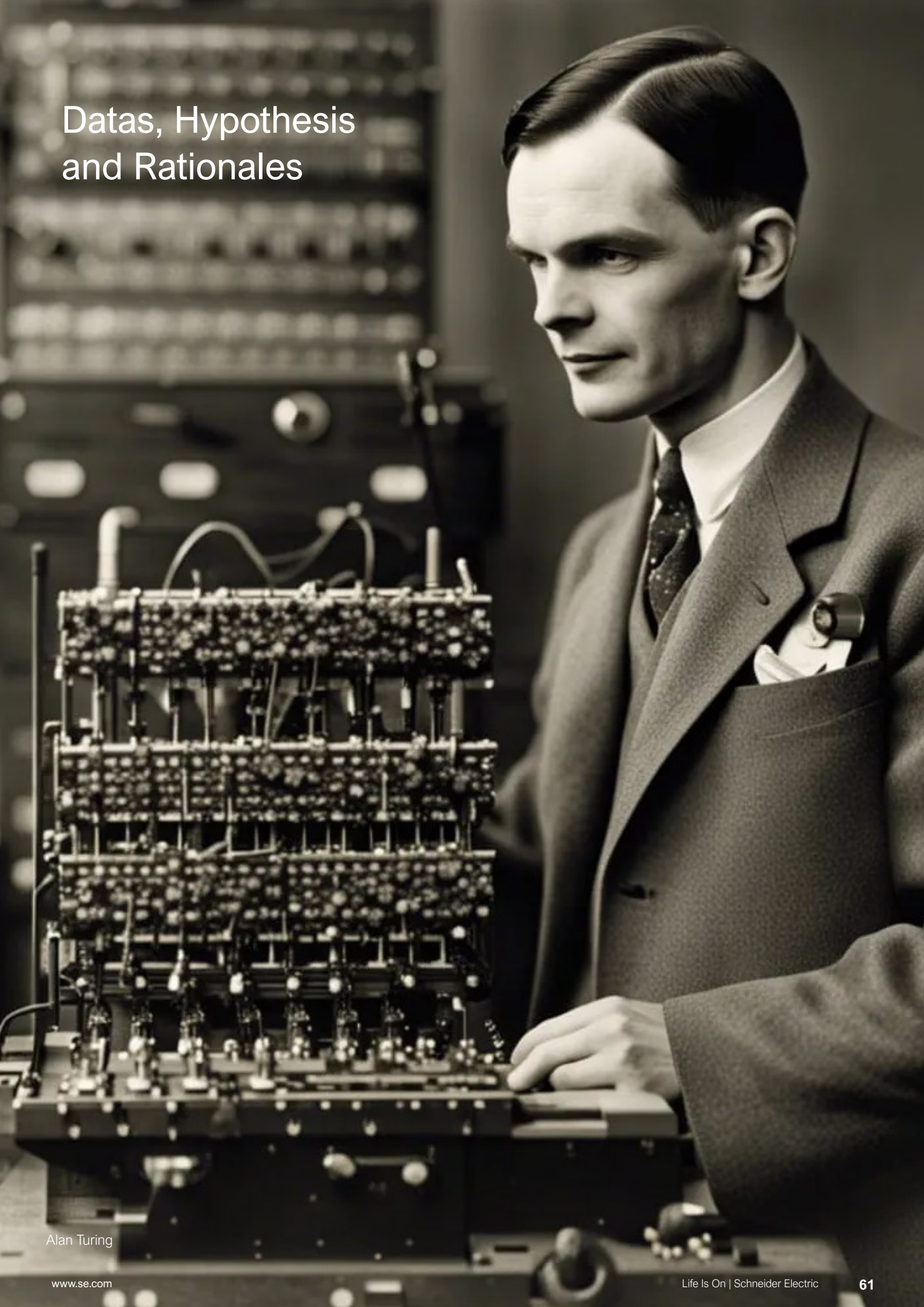
The Total Compute is the number of FLOPs required to fully train the GenAI/LLM. It is calculated by multiplying the Training Steps by the number of FLOPs per Step. The number of Training Steps represents the total number of times the weights (parameters) of the model are updated. It is calculated by multiplying the Dataset Size (expressed in tokens) by the number of Epochs. An epoch is one complete pass of the training dataset through the algorithm and shows how many times the dataset is used to train the model. This is one hyperparameter (a setting) of the training model.

Finally, the number of FLOPs per Step represents how many operations have to be performed in a Forward Pass and Backward Pass of the model. To get this value, the number of Model Parameters is multiplied by the number of FLOPs per parameter per token and divided by 1,000 to express it in TFLOPs. The number of FLOPs per parameter per token can be approximated using OpenAI's LLM scaling law (68) to a value of 6.

Table 7 Validation by triangulation

	Application to system dynamics models
Data	Behavioural model validation Data are representations of the past and in disruptive economic contexts possibly misleading. We will therefore not apply a data science approach to this study, but we will include data as possible patterns, e.g., as best guesses by professionals like the Schneider Electric 2024 report. Behavioral system dynamics models can produce data outputs with different input data. These outcomes can be assessed on "reasonableness". If outcomes are seen as not realistic, other input data, formulas or a restructuring of the model are needed. More details on data triangulation are given in the last part of this section.
Investigator	Search for alternative sources, like authors, experts and organizations Through intensive collaboration with Schneider Electric™ Sustainability Research Institute, we have access to many experts on AI and data centers. We will thus use these individuals for investigator triangulation by constructively working with their comments. Experts can comment on both the structural causal model (e.g., what variables are included or excluded and what relations are part of the model) and the behavioral model (e.g., commenting on outcomes).
Theory	Structural model validation and theory integration We integrate a demand-driven (top-down) approach with a supply-driven (bottom-up) approach to estimating energy needs and will compare the differences in multiple parts of our analysis. We integrate top-down and bottom-up approaches not to compare the difference of insights, but because we see them as two parts of the same reality. More details are provided in the next subsection.
Method	Identify a method used to ground reasoning and conclusions. The insights for our modeling and parameter selections are based on academic literature and professional publications. The number of academic publications in this field is scarce, so instead of a systematic literature search (which would result in few documents that are intensively used), we also apply a non-systematic literature search via the network of researchers in this field at Schneider and its academic and professional ecosystem.

Datas, Hypothesis and Rationales



Alan Turing

Endogenous and Exogenous Factors

Endogenous Factors

In the context of system dynamics modeling, this framework presents a structured approach for quantifying and analyzing the multifaceted factors influencing the potential growth of AI use, specifically focusing on trends originating from within data centers. The framework delineates six endogenous factor types, each representing a critical domain of technological and operational evolution inside the data center ecosystem. These factor types are further disaggregated into 34 micro factors, providing a granular level of detail essential for robust system dynamics modeling. To facilitate referencing within the model, each micro factor is assigned a unique identifier.

The factor types encompass:

• Hardware Evolution (EN1): This factor type focuses on the advancements in physical components of data centers, including networking, cooling systems, server optimization, storage solutions, and power management.
• Software and Algorithmic Efficiency (EN2): This category covers improvements in virtualization, workload management, AI-driven optimizations, and algorithm efficiency.
• Data Center Design and Infrastructure (EN3): This factor type addresses architectural trends, including modular designs, edge computing, and network optimizations.
• Operational Practices and Management (EN4): This category focuses on AI-driven maintenance, circular economy principles, cybersecurity, and energy efficiency standards.
• Research, Development, and Education (EN5): This factor type highlights investment patterns, technological advancements, and interdisciplinary collaborations in AI and energy efficiency.
• Workforce and Skills (EN6): This category examines the human aspect of data center operations, including skill availability, training programs, and occupational health standards.

Weighting System:
• -10: High negative trend (strongly reduces AI demand)
• -5: Moderate negative trend (moderately reduces AI demand)
• 0: Neutral/No significant change
• 5: Moderate positive trend (moderately increases AI demand)
• 10: High positive trend (strongly increases AI demand)

This weighting system allows for a nuanced assessment of how each factor might influence AI electricity consumption in different future scenarios. The weights range from -10 to 10, with negative values indicating a reduction in electricity consumption growth and positive values indicating an increase.

Example: Let's consider the microfactor EN1.2: "Cooling and HVAC technology advancements", part of the Hardware Evolution (EN1) factor, across the four scenarios:

Sustainable AI (H1): Weight = 10

Rationale: In this scenario, significant advancements in cooling and HVAC technologies are prioritized to enhance energy efficiency. These improvements allow for increased AI computational capacity within the same energy envelope, effectively enabling greater AI use without proportional increases in energy consumption. The high positive weight reflects the dual impact of these advancements: they both reduce energy consumption per unit of computation and facilitate expanded AI applications.

Limits to Growth (H2): Weight = 5

Rationale: Moderate progress in cooling and HVAC technologies is observed, contributing to incremental improvements in energy efficiency. While these advancements allow for some increase in AI use, their impact is limited by various constraints. The moderate weight indicates a positive but restrained influence on AI adoption, reflecting the scenario's overall theme of constrained growth.

Abundance without boundaries (H3): Weight = 5
Rationale: Rapid advancements in cooling and HVAC technologies are achieved, primarily driven by the need to support exponential growth in AI infrastructure. These improvements significantly reduce cooling-related energy consumption, allowing for substantial increases in AI computational capacity and use. The high positive weight reflects the critical role of these advancements in enabling widespread AI adoption and deployment, albeit with potential environmental trade-offs.

AI Energy Crisis (H4): Weight = 5
Rationale: Despite the energy crisis, cooling and HVAC technology advancements remain crucial for maintaining AI operations under severe energy constraints. These improvements are primarily focused on optimizing existing infrastructure rather than enabling expansion. The moderate weight reflects the importance of these advancements in preserving current AI capabilities and potentially allowing for limited growth, even in a resource-constrained environment.

Exogenous Factors

Exogenous factors are external influences that affect AI electricity consumption but are not directly controlled by the AI industry or data center operations. These factors shape the environment in which AI systems operate and can significantly impact their development, deployment, and energy use. They are categorized into ten main groups, each representing a different aspect of the external environment that can influence AI electricity consumption. These groups range from geopolitical and economic conditions to specific material and mineral availability.

List of Endogenous Factors, Microfactors and Associated Weights per Scenario

Factor type	Factor	Name	Micro Factor	#	Description of factor	Impacts on Scenarios			
						H1	H2	H3	H4
Hardware Evolution						**7**	**5**	**8**	**4**
Endogenous	EN1	Hardware Evolution	EN1.1	55	Chip efficiency and architecture innovations	10	0	10	5
			EN1.2	56	High-speed networking component evolution	5	5	10	**0**
			EN1.3	57	Cooling and HVAC technology advancements	10	5	5	5
			EN1.4	58	Rack and server optimization trends	5	5	10	**0**
			EN1.5	59	High-density storage solution developments	5	5	10	**0**
			EN1.6	60	Power distribution and management system improvements	10	5	5	10
Software and Algorithmic Efficiency						**8**	**5**	**7**	**4**
Endogenous	EN2	Software and Algorithmic Efficiency	EN2.1	60	Virtualization and containerization technique advancements	10	5	5	**0**
			EN2.2	61	Workload management and orchestration progress	10	5	5	5
			EN2.3	62	AI-driven resource allocation and scheduling evolution	10	5	5	10
			EN2.4	63	AI training technique efficiency improvements	10	5	5	10
			EN2.5	64	Inference algorithm optimization trends	10	5	5	10
			EN2.6	65	Federated learning and distributed AI advancement patterns	5	5	10	0
			EN2.7	66	AI and machine learning algorithm breakthrough dynamics	5	5	10	**0**
			EN2.8	67	Latency reduction techniques in AI systems	5	5	10	0
Data Center Design and Infrastructure						**7**	**5**	**8**	**1**
Endogenous	EN3	Data Center Design and Infrastructure	EN3.1	68	Modular and scalable data center design trends	10	5	5	0
			EN3.2	69	Edge data center proliferation patterns	5	5	10	**0**
			EN3.3	70	Hybrid and multi-cloud architecture evolution	5	5	10	0
			EN3.4	71	AI data center physical deployment planning	10	5	5	10
			EN3.5	72	AI/Gen AI deployment trends (cloud providers, colocation, ent..)	5	5	10	-5
			EN3.6	73	Network architecture optimizations for reducing latency	5	5	10	**0**
Operational Practices and Management						**9**	**5**	**4**	**4**
Endogenous	EN4	Operational Practices and Management	EN4.1	74	AI-driven predictive maintenance implementation trends	10	5	5	**0**
			EN4.2	75	Circular economy principle adoption in data center lifecycle	10	5	0	5
			EN4.3	76	Cybersecurity measure and resilience enhancement patterns	5	5	10	**0**
			EN4.4	77	Energy efficiency standard implementation for AI systems	10	5	0	10
Research, Development, and Education						**8**	**5**	**4**	**5**
Endogenous	EN5	Research, Development, and Education	EN5.1	78	Energy-efficient AI technology investment patterns	10	5	0	10
			EN5.2	79	Specialized AI hardware development trends	5	5	10	**0**
			EN5.3	80	AI model compression technique advancement dynamics	10	5	0	10
			EN5.4	81	AI education and skill development program evolution	5	5	10	**0**
			EN5.5	82	Research funding allocation trends for AI and energy efficiency	10	5	0	10
			EN5.6	83	Interdisciplinary collaboration between AI and energy sectors	10	5	**0**	5
			EN5.7	84	Research into overcoming data scarcity and latency challenges	5	5	10	0
Workforce and Skills						**6**	**5**	**6**	**1**
Endogenous	EN6	Workforce and Skills	EN6.1	85	Skilled personnel availability trends in AI and data center	5	5	10	-5
			EN6.2	86	Training and upskilling program evolution	10	5	5	**0**
			EN6.3	87	Routine task automation and human-AI collaboration progress	5	5	10	0
			EN6.4	88	Occupational health standard for AI-intensive industries	5	5	**0**	10

Table 8: List of endogenous factors, microfactors and associated weights per scenario.

List of Exogenous Factors, Microfactors and Associated Weights per Scenario

Factor type	Factor	Name	Micro Factor	#	Description of factor	Impacts on Scenarios			
						H1	H2	H3	H4
Geopolitical and Economic Conditions						5	3	-1	-7
Exogenous	EX1	Geopolitical and Economic Conditions	EX1.1	1	Global events dynamics (trade wars, pandemics, political instability)	0	5	-5	-10
			EX1.2	2	Economic cycles and market dynamics	5	0	10	-10
			EX1.3	3	International relations and technology transfer policy trends	5	0	-5	-10
			EX1.4	4	International cooperation or competition in AI development	5	5	-5	-10
			EX1.5	5	Global AI governance framework evolution	10	5	0	5
Regulatory and Policy Landscape						8	5	0	-9
Exogenous	EX2	Regulatory and Policy Landscape	EX2.1	6	Data privacy and protection law developments	5	5	0	-10
			EX2.2	7	Cybersecurity regulations evolution	5	5	0	-10
			EX2.3	8	Environmental and energy efficiency standard changes	10	5	0	-10
			EX2.4	9	Carbon pricing and emissions trading scheme dynamics	10	5	0	-10
			EX2.5	10	Circular economy initiative trends in the tech sector	10	5	0	-5
			EX2.6	11	Health & safety regulation changes (electromagnetic radiation…)	5	5	0	-10
Technological Advancements (External to AI Industry						8	4	9	1
Exogenous	EX3	Technological	EX3.1	12	Quantum computing development trends	5	0	10	-5
			EX3.2	13	Networking technology evolution (e.g., 5G, 6G)	10	5	10	0
			EX3.3	14	Energy storage technology progress	10	5	5	10
			EX3.4	15	Standardization and interoperability protocol developments	5	5	10	0
Energy Sector Developments						7	3	8	3
Exogenous	EX4	Energy Sector	EX4.1	16	Renewable energy adoption and grid integration trends	10	5	5	-5
			EX4.2	17	Smart grid implementation progress	10	5	10	0
			EX4.3	18	Nuclear energy development dynamics	5	0	10	10
			EX4.4	19	Electricity cost fluctuations and their impact on AI deployment	-5	0	5	10
			EX4.5	20	Power generation capacity expansion patterns	10	5	10	-5
			EX4.6	21	Energy storage technology advancements	10	5	10	5
Power Availability and Infrastructure						8	3	1	-6
Exogenous	EX5	Power Availability and Infrastructure	EX5.1	22	Grid stability and reliability trends	5	0	-5	-10
			EX5.2	23	Power transmission and distribution network development	10	5	5	-5
			EX5.3	24	Regional variations in power availability and quality	5	0	-5	-10
			EX5.4	25	Energy mix evolution (fossil fuels vs. renewables)	10	5	5	-5
			EX5.5	26	Microgrid and distributed energy resource adoption patterns	10	5	0	5
			EX5.6	27	Power purchase agreement (PPA) trends for data centers	10	5	5	-10
Market Demand and Economic Factors						8	5	7	-3
Exogenous	EX6	Market Demand and Economic Factors	EX6.1	28	Cloud computing service growth patterns	5	5	10	-5
			EX6.2	29	IoT and edge computing evolution	10	5	10	0
			EX6.3	30	AI-driven application demand trends	5	5	10	-5
			EX6.4	31	Investment trends in AI and Sustainable technologies	10	5	5	-10
			EX6.5	32	Economic incentives for energy-efficient AI development	10	5	0	5

Table 9: List of exogenous factors, microfactors and associated weights per scenario.

List of Exogenous Factors, Microfactors and Associated Weights per Scenario

Supply Chain and Resource Constraints						2	0	1	-7
Exogenous	EX7	Supply Chain and Resource Constraints	EX7.1	34	Chip packaging capacity dynamics	5	0	10	-5
			EX7.2	35	HBM chips production capacity trends	5	0	10	-5
			EX7.3	36	Wafer production capacity changes	0	5	10	-5
			EX7.4	37	Major disruption occurrences and impacts	-5	0	-5	0
			EX7.5	38	Availability trends of rare earth elements for AI hardware	0	-5	-5	-5
			EX7.6	39	Water scarcity impacts on data center cooling	5	0	-5	-5
			EX7.7	40	Land use consideration patterns for AI infrastructure	5	0	-5	0
Social and Ethical Considerations						7	2	2	-5
Exogenous	EX8	Social and Ethical	EX8.1	41	Public perception and acceptance trends of AI technologies	5	0	10	5
			EX8.2	42	Ethical guideline developments for AI deployment	10	5	0	5
			EX8.3	43	Social equity concern evolution in AI access and energy distribution	5	0	-5	0
Data Availability and Quality						6	1	5	-2
Exogenous	EX9	Data Availability and Quality	EX9.1	44	High-quality training data availability trends	5	0	10	-5
			EX9.2	45	Data collection and curation methodologies evolution	10	5	5	0
			EX9.3	46	Data privacy regulations impact on data accessibility	0	-5	0	0
			EX9.4	47	Synthetic data generation capabilities advancement	10	5	10	5
			EX9.5	48	Domain-specific data scarcity challenges	5	0	10	0
Materials & Minerals						5	3	4	9
Exogenous	EX10	Materials & Minerals	EX10.1	49	Critical mineral availability for AI hardware components	5	0	5	10
			EX10.2	50	Rare earth element supply chain dynamics	0	0	5	10
			EX10.3	51	Recycling and circular economy trends for tech materials	10	5	0	10
			EX10.4	52	New material development for energy-efficient AI hardware	10	5	10	5
			EX10.5	53	Geopolitical tensions impacting mineral/material access	0	0	5	10
			EX10.6	54	Sustainable mining practices for AI-relevant materials	5	5	0	10

Integrating Endogenous and Exogenous Factors in a System Dynamics Model

Using the IIASA framework, we categorize exogenous factors into four main categories : Energy and Material, Economy and Industry, Society and Behavior, and Governance and Markets to organize and simplify the complex array of external influences on AI development from trends outside the data center. This method aligns with system dynamics modeling practices, allowing for a more structured and comprehensible analysis of the various factors affecting AI development and use.

- A higher number (3) means that the exogenous factor has a strong impact on that particular category.
- A lower number (1) means the factor has minimal impact on that category.
- A middle number (2) indicates a moderate level of influence.

Factor / Category	#	Energy and Material	Economy and Industry	Society and Behavior	Governance and Markets
Geopolitical and Economic Conditions	EX1	3	2	2	3
Regulatory and Policy Landscape	EX2	2	3	2	3
Technological Advancements (External to AI	EX3	2	2	2	3
Energy Sector Developments	EX4	3	2	1	2
Power Availability and Infrastructure	EX5	3	2	1	2
Market Demand and Economic Factors	EX6	2	3	2	2
Supply Chain and Resource Constraints	EX7	2	3	2	1
Social and Ethical Considerations	EX8	1	2	3	2
Data Availability and Quality	EX9	1	2	2	2
Materials & Minerals	EX10	3	2	1	1

Table 10: Exogenous factors and weights in our model

GenAI Inferencing Sub Model

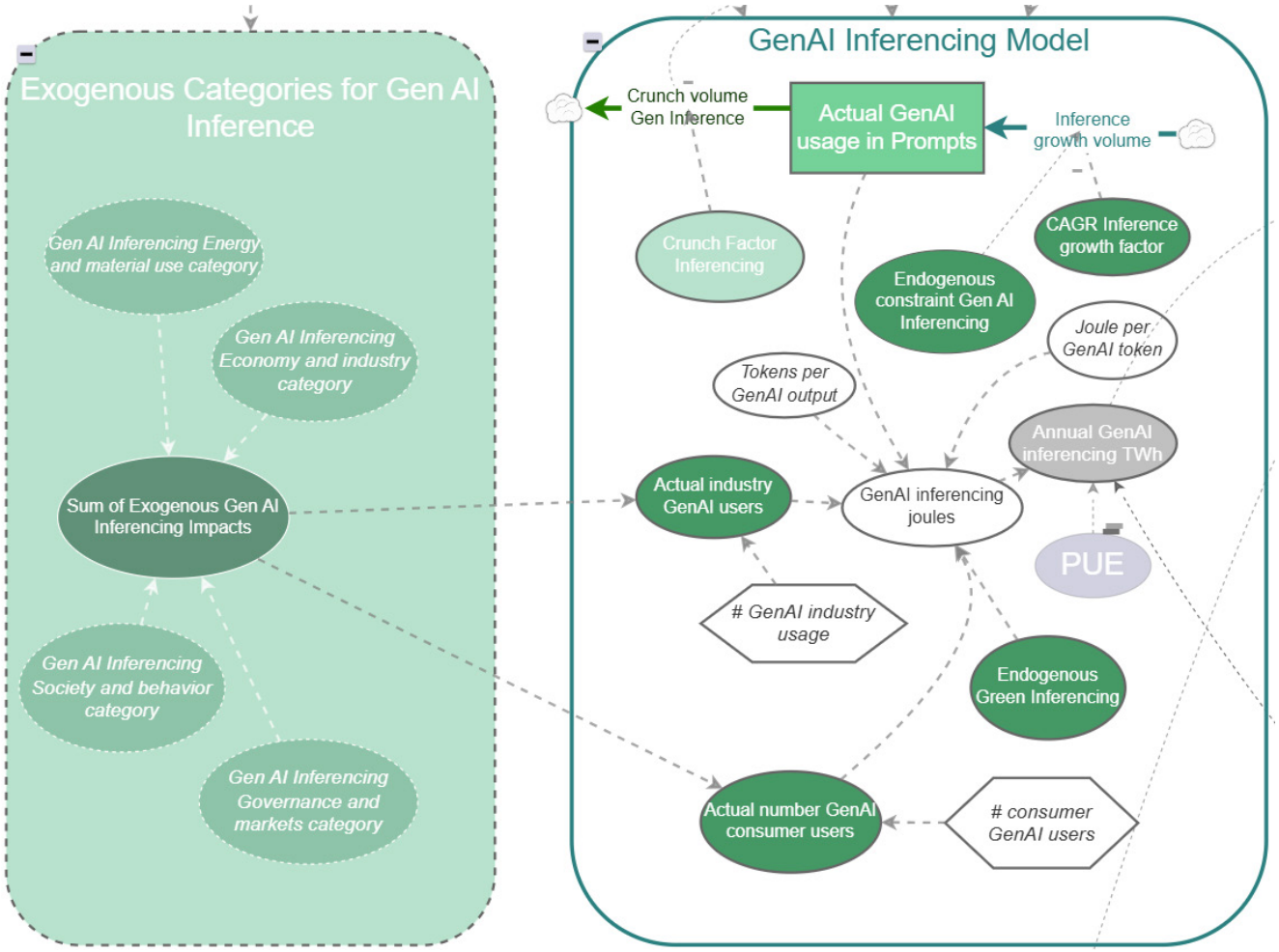


Figure 9: The GenAI inferencing submodel

The GenAI Inferencing Sub Model depicted in the image illustrates the relationships between various parameters that drive and constrain the electricity use evolution of Generative AI (GenAI) inferencing. Here's a breakdown of the key components and their interrelations:

Key Components

• Actual GenAI Usage in Prompts: This is the central node of the model, representing the actual volume of electricity use in TWh of GenAI inferencing tasks processed (e.g., number of prompts). It is influenced by several factors, including Endogenous Growth Factor for Gen AI Inferencing and Endogenous Constraint for GenAI Inferencing.

• GenAI Inferencing Joules: This node represents total energy consumption for performing GenAI inference tasks, measured in joules. It is influenced by factors like PUE (Power Usage Effectiveness), which measures how efficiently energy is used in data centers, and the number of tokens processed per output. Higher energy consumption negatively impacts sustainability unless offset by improvements in Sustainable inferencing.

• Annual GenAI Inferencing TWh: This represents the total energy consumption over a year, measured in terawatt-hours (TWh). It is directly related to how much energy is consumed per inference task and how many tasks are performed.

• Endogenous Growth Factor for Gen AI Inferencing: This represents the Compound Annual Growth Rate (CAGR) of GenAI inferencing usage. It drives the inference growth volume, which increases the overall demand for GenAI tasks. A higher growth factor leads to an increase in Actual GenAI Usage in Prompts.

• Endogenous Constraint for GenAI Inferencing: This represents internal limitations or bottlenecks within the system, such as computational resources, energy efficiency, and hardware capacity. Higher constraints reduce the system's ability to scale efficiently, limiting both Actual GenAI Usage in Prompts and overall system performance.

GenAI Inferencing Sub Model

• Endogenous Efficiency for Gen AI Inferencing: Endogenous Efficiency for Gen AI Inferencing refers to the efficiency of the GenAI inferencing process, influenced by factors like energy efficiency and hardware performance. For instance, ML hardware has significantly improved, with computational performance doubling every 2.8 years. These advancements, driven by optimized number formats and specialized tensor cores, contribute to higher overall efficiency.

• Crunch Factor Inferencing: The Crunch Factor in the context of GenAI Inferencing refers power availability limitations that restrict the development and efficient scaling of GenAI systems.

• Power Usage Effectiveness (PUE): PUE measures how efficiently a data center uses energy; a lower PUE indicates better efficiency. PUE influences both GenAI Inferencing Joules and overall sustainability (Endogenous Sustainable Inferencing).

• # GenAI industry users: number of GenAI industry usage represents the total count of businesses or organizations actively employing generative AI technologies in their operations, products, or services. This metric reflects the adoption and integration of GenAI across various industrial sectors, influenced by factors such as technological advancements, resource availability, regulatory environments, and market dynamics.

• # Gen AI consumer users: The number of consumer GenAI users represents the total count of individuals who regularly interact with or utilize generative AI technologies for personal or non-professional purposes. This includes users of AI-powered chatbots, content generation tools, personal assistants, and other consumer-facing generative AI applications.

• Joule per GenAI Token: Refers to the amount of energy, measured in joules, required to process a single token during a Generative AI (GenAI) inferencing task. A token typically represents a word or part of a word in natural language processing tasks, and inferencing involves generating predictions or outputs based on a trained AI model.

• Tokens per GenAI Output: Refers to the number of tokens generated by a Generative AI (GenAI) model in response to a single input or prompt during inferencing. A token typically represents a word, part of a word, or a character, depending on the language model and the task. The number of tokens per output is an important metric as it directly influences the computational load and energy consumption of the model. More tokens generally require more computational resources, leading to higher energy consumption.

• The Actual Industry GenAI Users Industry node in the diagram refers to the number of enterprise or industrial users actively utilizing Generative AI (GenAI) systems for a variety of applications. These users typically represent companies, organizations, or industries that leverage GenAI for tasks such as automation, decision-making, content generation, and optimization of business processes. This parameter is influenced by four key exogenous categories (see right column).

1. GenAI Inferencing Energy and Material Use Category
The energy consumption and material resources required for inferencing play a crucial role in determining how scalable and sustainable GenAI adoption is for industrial users. High energy demands or resource constraints may limit the number of industry users.

2. GenAI Inferencing Economy and Industry Category
Economic factors, such as the cost of deploying and maintaining GenAI systems, as well as industry-specific trends (e.g., automation in manufacturing or AI-driven analytics in finance), influence how widely GenAI is adopted in different sectors.

3. GenAI Inferencing Society and Behavior Category
Societal acceptance of AI technologies within industries, along with workforce behavior (e.g., willingness to integrate AI into workflows), affects the rate at which industries adopt GenAI solutions.

4. GenAI Inferencing Governance and Markets Category
Regulatory frameworks, market competition, and policies around AI usage (e.g., data privacy laws or industry standards) also impact how many enterprises can or will adopt GenAI technologies.

• The Actual Number of GenAI Consumer Users node represents the total number of individual, non-enterprise users who actively engage with Generative AI systems for personal or consumer-level applications. These users typically interact with Gen AI-powered tools for tasks such as content creation, personal assistance (e.g., chatbots), entertainment (e.g., AI-generated media), and other consumer-oriented activities. This parameter is similarly influenced by the same four exogenous categories:

1. GenAI Inferencing Energy and Material Use Category
The energy efficiency and material costs associated with running consumer-facing GenAI applications (e.g., cloud-based services) affect how accessible these technologies are to everyday consumers. High energy costs may limit the scalability of consumer applications.

2. GenAI Inferencing Economy and Industry Category
The affordability of consumer-facing AI products and services plays a major role in adoption rates. Economic conditions such as disposable income levels or market pricing for AI-powered tools influence how many consumers can engage with these technologies.

3. GenAI Inferencing Society and Behavior Category
Consumer behavior, including attitudes toward AI technology (e.g., trust in AI-generated content) and societal trends (e.g., demand for personalized digital experiences), directly affects the number of people using GenAI tools.

4. GenAI Inferencing Governance and Markets Category
Policies surrounding data privacy, ethical AI use, and market regulations can either encourage or hinder consumer adoption of GenAI technologies. For example, stricter data protection laws might slow down adoption if they increase compliance costs for service providers.c

Actual GenAI Usage in Prompts

Definition
The Actual GenAI Usage in Prompts is the central node of the GenAI Inferencing Model, representing the total volume of electricity consumed (in TWh) by Generative AI (GenAI) systems to process inferencing tasks (i.e., prompts). It reflects the actual energy used to handle the number of prompts processed by GenAI models, such as text generation, image generation, or multimodal tasks.

Role in the Model
> Key Indicator of Energy Demand: This node acts as a direct measure of how much electricity is consumed by GenAI systems based on the number of prompts processed. It encapsulates the core energy consumption resulting from GenAI operations.
> Influence on Other Nodes: The Actual GenAI Usage in Prompts drives several other important nodes in the model:
• It directly impacts GenAI Inferencing Joules, which represents the energy consumed per task.
• It influences Annual GenAI Inferencing TWh, which tracks total yearly energy consumption.
• It affects the number of Actual Industry GenAI Users and Actual Number of GenAI Consumer Users, as higher energy demands could limit scalability and adoption.
> Value: TWh

Key Assumptions for 2023 Calculation
We use the following structuration from Luccioni et al, 2024.

Task (kWh/1000 queries)	mean	sd
Text classification	0.002	0.001
Extractive QA	0.003	0.001
Image classification	0.007	0.001
Object detection	0.038	0.02
Text generation	0.047	0.03
Summarization	0.049	0.01
Image captioning	0.063	0.02
Image generation	2.907	3.31

Energy Consumption Calculation for 2023
> Number of Prompts in 2023: Based on industry estimates, Gen AI systems (like ChatGPT) are handling approximately 78 billion prompts annually.
> Task Distribution: We assume that 80% of these prompts are text-based tasks, while 20% are image-based tasks.

Text-based tasks (80% of total prompts):
> Text classification: 0.002 kWh per 1000 queries
> Extractive QA: 0.003 kWh per 1000 queries
> Text generation: 0.047 kWh per 1000 queries
> Summarization: 0.049 kWh per 1000 queries
> Weighted average for text-based tasks: (0.002 + 0.003 + 0.047 + 0.049) / 4 = 0.02525 kWh per 1000 queries or 0.00002525 kWh per query

Image-based tasks (20% of total prompts):
> Image classification: 0.007 kWh per 1000 queries
> Object detection: 0.038 kWh per 1000 queries
> Image captioning: 0.063 kWh per 1000 queries
> Image generation: 2.907 kWh per 1000 queries
> Weighted average for image-based tasks: (0.007 + 0.038 + 0.063 + 2.907) / 4 = 0.75375 kWh per 1000 queries or 0.00075375 kWh per query

Final Energy Consumption Calculation for 2023
> Total prompts: 78 billion
> Text-based prompts (80%): 62.4 billion
> Image-based prompts (20%): 15.6 billion
> Text-based prompts: 62.4 billion × 0.00002525 kWh = 1.5756 TWh
> Image-based prompts: 15.6 billion × 0.00075375 kWh = 11.7585 TWh
> Total Energy Consumption in 2023: 1.5756 TWh + 11.7585 TWh = 13.3341 TWh

Projected Energy Consumption of Generative AI in 2025

Methodology
Our projection is based on the following key assumptions:
oGrowth in AI Usage: We anticipate a Compound Annual Growth Rate (CAGR) of 40% in the number of AI prompts processed, increasing from 78 billion in 2023 to approximately 153 billion by 2025.
oTask Distribution: We project a shift towards more energy-efficient tasks, with 70% text-based and 30% image-based prompts by 2025.
oEnergy Efficiency Improvements: Significant advancements in AI technology and hardware are expected to reduce energy consumption per task substantially.

Detailed Calculations
•Projected Number of Prompts (2025)
•2023 Base: 78 billion prompts
•CAGR: 40%
•2025 Projection: 78 billion × $(1 + 0.40)^2 \approx 153$ billion prompts
•Energy Consumption by Task Type

Task Type	Prompts (billions)	Energy per Prompt (kWh)	Total Energy (TWh)
Text-based	107.1 (70%)	0.00000505	0.5408
Image-based	45.9 (30%)	0.0003015	13.8389
Total	153.0		14.3797

Key Factors Influencing the Projection
> Technological Advancements: We anticipate an 80% reduction in energy consumption for text-based tasks and a 60% reduction for image-based tasks, driven by:
• Improved natural language processing models
• Optimized image processing algorithms
• More efficient hardware (e.g., advanced GPUs)
> Shift in Task Distribution: The projection assumes a higher proportion of less energy-intensive text-based tasks (70%) compared to more energy-intensive image-based tasks (30%).
> Industry-wide Efficiency Measures:
• Implementation of energy-efficient scheduling in AI systems
• Increased use of shared data centers and cloud computing resources
• Advancements in computational storage and CXL technology

Conclusion
Our analysis projects that GenAI systems will consume approximately 14.38 TWh of energy in 2025, aligning with the initial value for the system dynamic model of 15 TWh.

This projection is based on anticipated rapid technological advancements and widespread adoption of energy-efficient practices in the AI industry.

Endogenous Growth Factor for Gen AI Inferencing

Definition
The Endogenous Growth Factor for Gen AI Inferencing represents the rate at which Gen AI inferencing usage is expected to grow over time. It is typically expressed as a compound annual growth rate (CAGR), which measures how much the volume of Gen AI inference increases year over year.

Role in the Model
This factor directly influences the inference growth volume, which drives how much more energy, resources, or infrastructure will be required to support the growing demand for Gen AI inferencing.

Scale (0 to 1)
A value closer to 1 indicates a higher growth rate, meaning that the demand for Gen AI inferencing is expanding rapidly.
A value closer to 0 indicates slower or minimal growth.

Calculation methodology
> The approach for quantifying this value is to link to weighting each micro factor involves the following steps.

Weighting system
-10: High negative trend (strongly reduces Gen AI inferencing growth)
-5: Moderate negative trend (moderately reduces Gen AI inferencing growth)
0: Neutral/No significant change
5: Moderate positive trend (moderately increases Gen AI inferencing growth)
10: High positive trend (strongly increases Gen AI inferencing growth)

Factor evaluation: Each micro factor is evaluated for its impact on Gen AI inferencing growth in each scenario. Positive weights indicate factors that increase AI demand (potentially leading to more consumption), while negative weights indicate factors that decrease AI demand (potentially leading to less consumption).

Scenario-specific weighting: The weights are assigned based on how each factor is expected to evolve in each scenario. For example, in the Sustainable AI scenario, many factors related to efficiency improvements have high positive weights, as they are expected to enable more AI usage while managing energy consumption.

Aggregation: The weights for each macro factor are averaged to provide an overall trend for that category. Then, an average of all endogenous factors is calculated for each scenario.

Endogenous constraint calculation: The Endogenous constraint Gen AI Inferencing values (0 to 1) are derived from these averages. A higher average indicates lower constraints (closer to 0), while a lower average indicates higher constraints (closer to 1). The exact formula for this conversion is not provided, but it involves normalizing the averages to fit the 0-1 scale.

Conversion Formula: The average impact scores to the 0-1 scale is: Constraint Value = 1 - ((Average Impact Score + 10) / 20)
This formula normalizes the -10 to 10 scale to a 0 to 1 scale and inverts it so that higher impact scores result in lower constraint values.

Scenario	Average Endogenous Microfactors Weights	Growth Factor
H1	7	0.7 (70% / year)
H2	3	0.3 (30% / year)
H3	6	0.6 (60% / year)
H4	6	0.6 (60% / year)

Results for Endogenous Growth Factor for Gen AI Inferencing

Endogenous Constraint for GenAI Inferencing

Definition
The Endogenous Constraint for GenAI Inferencing refers to the physical and operational limitations that restrict the ability to deploy and scale Generative AI (GenAI) systems in practice. While growth in GenAI usage is driven by demand and technological advancements, these constraints represent the system's capacity to physically deliver this growth. Examples of endogenous constraints include delays in AI data center (AI-DC) construction, shortages in chip production (e.g., GPU delays), power infrastructure bottlenecks, cooling capacity limitations, hardware inefficiencies, and supply chain disruptions.

Role in the Model
> Capacity Limitation: Endogenous constraints act as a limiting factor on the system's ability to scale GenAI inferencing. Even if demand (inference growth volume) is high, these constraints determine whether the system can physically handle the increased workload.
> Impact on Energy Efficiency: Higher endogenous constraints lead to inefficiencies in energy usage, such as higher Joules per GenAI Token, which increases overall electricity consumption. This directly impacts nodes like GenAI Inferencing Joules and Annual GenAI Inferencing TWh, as more energy is required to compensate for inefficiencies.
> Influence on Scalability: These constraints also affect how many industry and consumer users can adopt GenAI technologies. If constraints are too high, it limits the number of Actual Industry GenAI Users and Actual Number of GenAI Consumer Users, slowing down overall adoption.

Physical Examples of Endogenous Constraints
> Delays in AI Data Center Construction: For example, power infrastructure bottlenecks in key markets like Northern Virginia have led to delays of over three years for new data center builds
> Chip Production Delays: Nvidia's Blackwell AI chip delays due to design flaws have pushed back hyperscale data center deployments by several months, affecting companies like Microsoft, Google, and Meta
> Supply Chain Shortages: Shortages in critical components like chip packaging (CoWoS) have delayed GPU availability, impacting AI data center builds and increasing backorders
> Hardware Failures and Inefficiencies: Issues such as overheating GPUs, network interface card (NIC) failures, or optical transceiver malfunctions can degrade performance or cause system outages, further constraining capacity
> Cooling Capacity Limitations: The need for advanced cooling systems to manage heat dissipation in high-performance computing clusters can limit data center scalability

Scale (0 to 1)
> A value of 0 represents no constraint, meaning the system can perform Gen AI inferencing without any internal limitations.A value closer to 0 indicates slower or minimal growth.
> A value of 1 represents severe constraints, indicating significat internal limitations that hinder Gen AI inferencing capabilities.

The Endogenous Constraint for GenAI Inferencing values have been set up based on a combination of current technological trends, expert projections, and scenario-specific assumptions. Here's a brief rationale for each:

> Value: 0.18 (Sustainable AI scenario): This relatively low constraint value likely reflects optimistic projections about technological advancements and energy efficiency improvements. It assumes significant progress in hardware and software optimization, as well as strategic investments in sustainable AI infrastructure.

> Value: 0.51 (Limits to Growth scenario): The higher constraint value here suggests a more conservative outlook, where improvements in AI technology and infrastructure are offset by resource limitations and regulatory constraints. This balanced approach accounts for both advancements and challenges in AI development.

> Value: 0.12 (Abundance scenario): This lowest constraint value indicates an extremely optimistic view of AI development, assuming rapid and substantial breakthroughs in AI hardware, software, and infrastructure. It likely considers aggressive adoption of cutting-edge technologies and minimal regulatory barriers.

> Value: 0.36 (AI Energy Crisis scenario): This moderate to high constraint value reflects a scenario where AI development outpaces infrastructure capabilities. It likely considers factors such as energy supply limitations, cooling challenges, and potential supply chain disruptions that could hinder AI deployment.

Endogenous Constraint for GenAI Inferencing

H1	2025: 0.18 2026: 0.17 2027: 0.16 2028: 0.15 2029: 0.14 2030: 0.13 2031: 0.12 2032: 0.11 2033: 0.10 2034: 0.09 2035: 0.08	In the Sustainable AI scenario, the decreasing endogenous constraint on Gen AI inferencing from 0.18 in 2025 to 0.08 in 2035 is driven by a combination of hardware and software advancements, as well as strategic investments in energy efficiency. Key factors include improvements in cooling and HVAC technology (EN1.2), power distribution systems (EN1.5), and software efficiency (EN2). Specific advancements contributing to this trend are virtualization and containerization techniques (EN2.1), AI-driven resource allocation and scheduling (EN2.3), more efficient AI training techniques (EN2.4), and optimized inference algorithms (EN2.5). These developments are supported by ongoing hardware improvements, such as Google's TPUv4 AI chips being 2.7 times more efficient than TPUv3 (Jouppi et al., 2021), and software optimization techniques aimed at reducing energy and computational costs while maintaining performance (Schwartz et al., 2020). The implementation of energy efficiency standards for AI systems (EN4.4) and increased investment in energy-efficient AI technologies (EN5.1) further contribute to this trend. Additionally, strategies for improving the energy efficiency of large language models (Patterson et al., 2021) indicate a focus on sustainable practices in AI development. Collectively, these factors lead to a more efficient and sustainable AI infrastructure, gradually reducing constraints on Gen AI inferencing over the forecast period.
H2	2025: 0.51 2026: 0.52 2027: 0.53 2028: 0.54 2029: 0.55 2030: 0.56 2031: 0.57 2032: 0.58 2033: 0.59 2034: 0.60 2035: 0.61	For the H2 - Limits to Growth scenario, the slight increase in endogenous constraint on Gen AI inferencing from 0.51 in 2025 to 0.61 in 2035 reflects a balanced but constrained growth trajectory. This scenario is characterized by moderate improvements across multiple areas, but with no single factor driving significant breakthroughs. Key aspects include steady advancements in hardware evolution (EN1), such as incremental improvements in high-speed networking components (EN1.1) and cooling technologies (EN1.2). Software and algorithmic efficiency (EN2) see balanced progress, with gradual enhancements in workload management (EN2.2) and AI training techniques (EN2.4). Data center design and infrastructure (EN3) evolve steadily, with modest advancements in modular designs (EN3.1) and hybrid cloud architectures (EN3.3). Operational practices and management (EN4) show consistent development, particularly in energy efficiency standards implementation (EN4.4). Research, development, and education efforts (EN5) maintain a balanced approach, with ongoing investments in energy-efficient AI technologies (EN5.1) and specialized hardware development (EN5.2). However, these improvements are offset by growing challenges. Strubell et al. (2019) highlight the significant computational costs of training large AI models, suggesting limits to scaling. Sun et al. (2017) discuss data scarcity challenges in AI, potentially constraining large-scale AI model growth. Additionally, regulatory restrictions, such as the EU AI Act (Veale and Borgesius, 2021), may limit certain AI applications. The combination of these factors results in a gradual increase in constraints on Gen AI inferencing, reflecting a scenario where technological advancements are balanced by resource limitations and societal concerns.
H3	2025: 0.12 2026: 0.11 2027: 0.10 2028: 0.09 2029: 0.08 2030: 0.07 2031: 0.06 2032: 0.05 2033: 0.04 2034: 0.03 2035: 0.02	In the H3 - Abundance without Boundaries scenario, the rapid decrease in endogenous constraint on Gen AI inferencing from 0.12 in 2025 to 0.02 in 2035 is driven by aggressive advancements in hardware, infrastructure, and AI capabilities. Key factors contributing to this trend include high-speed networking component evolution (EN1.1), rack and server optimization (EN1.3), and high-density storage solution developments (EN1.4), which collectively enhance the physical infrastructure supporting AI systems. On the software and algorithmic front, federated learning and distributed AI advancement patterns (EN2.6) and AI and machine learning algorithm breakthrough dynamics (EN2.7) play crucial roles in expanding AI capabilities and efficiency. The proliferation of edge data centers (EN3.2) and the evolution of hybrid and multi-cloud architectures (EN3.3) further support this trend by optimizing AI deployment and resource utilization. This scenario aligns with Brynjolfsson and McAfee's (2017) vision of AI's transformative potential across all sectors, suggesting widespread adoption and investment. The trend of increasing AI model sizes, as demonstrated by Brown et al. (2020) with GPT-3's 175 billion parameters, indicates a trajectory of expanding computational capabilities. Additionally, emerging fields like AI-driven scientific discovery, highlighted by Jumper et al. (2021), could drive increased AI computation demand and efficiency improvements. These factors collectively contribute to the significant reduction in constraints on Gen AI inferencing, reflecting a future where technological advancements and widespread AI adoption lead to a near-elimination of endogenous constraints by 2035.
H4	2025: 0.36 2026: 0.39 2027: 0.43 2028: 0.48 2029: 0.54 2030: 0.61 2031: 0.69 2032: 0.78 2033: 0.88 2034: 0.95 2035: 0.98	In the H4 - AI Energy Crisis scenario, the sharp increase in endogenous constraint on Gen AI inferencing from 0.36 in 2025 to 0.98 in 2035 is driven by a combination of infrastructure limitations and energy management challenges. Cooling and HVAC technology advancements (EN1.2) become critical as data centers struggle to manage increased heat generation from more powerful AI systems. Power distribution and management systems (EN1.5) face difficulties in keeping pace with the rapidly growing energy demands. AI-driven resource allocation and scheduling (EN2.3) becomes crucial but struggles to mitigate the overall energy crisis. Despite efforts to improve AI training technique efficiencies (EN2.4), these advancements fail to offset the dramatically increased demand. Inference algorithm optimization (EN2.5) becomes essential but cannot fully compensate for the energy-intensive nature of large-scale AI deployments. AI data center physical deployment planning (EN3.4) faces significant challenges in balancing expansion with local power constraints, as highlighted by Masanet et al. (2020). The urgent need for energy efficiency standards (EN4.4) for AI systems emerges as a critical factor. Bauer et al. (2021) point out how semiconductor shortages and supply chain disruptions contribute to inefficiencies, exacerbating the energy crisis. Additionally, Andrae and Edler's (2015) research on ICT efficiency improvements leading to increased overall energy consumption due to expanded use illustrates the potential rebound effects in AI energy consumption. These factors collectively contribute to a scenario where AI development outpaces infrastructure capabilities, leading to severe constraints on Gen AI inferencing and a potential energy crisis by 2035.

Endogenous Constraint for GenAI Inferencing

These forecasts and rationales are based on current research, expert discussions and trends in AI development. However, it is important to note that the field of AI is rapidly evolving, and future developments may significantly impact these projections.

Supporting sources for Endogenous Constraint for GenAI Inferencing

Jouppi, N. P., Yoon, D. H., Kurian, G., Li, S., Patil, N., Laudon, J., ... & Patterson, D. (2021). Ten lessons from three generations shaped Google's TPUv4i: Industrial product. In 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA) (pp. 1-14). IEEE. https://ieeexplore.ieee.org/document/9499913

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54-63. https://dl.acm.org/doi/10.1145/3381831

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350. https://arxiv.org/abs/2104.10350

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243. https://arxiv.org/abs/1906.02243

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE international conference on computer vision (pp. 843-852). https://openaccess.thecvf.com/content_iccv_2017/html/Sun_Revisiting_Unreasonable_Effectiveness_ICCV_2017_paper.html

Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. Computer Law Review International, 22(4), 97-112. https://doi.org/10.9785/cri-2021-220402

Brynjolfsson, E., & McAfee, A. (2017). The business of artificial intelligence. Harvard Business Review, 7, 3-11. https://hbr.org/2017/07/the-business-of-artificial-intelligence

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165. https://arxiv.org/abs/2005.14165

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589. https://www.nature.com/articles/s41586-021-03819-2
Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). Recalibrating global data center energy-use estimates. Science, 367(6481), 984-986. https://www.science.org/doi/10.1126/science.aba3758

Bauer, H., Burkacky, O., Kenevan, P., Mahindroo, A., & Patel, M. (2021). Semiconductor shortage: How the automotive industry can succeed. McKinsey & Company. https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/semiconductor-shortage-how-the-automotive-industry-can-succeed

Andrae, A. S., & Edler, T. (2015). On global electricity usage of communication technology: trends to 2030. Challenges, 6(1), 117-157. https://www.mdpi.com/2078-1547/6/1/117

Endogenous Efficiency for Gen AI Inferencing

Definition
The Endogenous Efficiency Factor for GenAI Inferencing measures how Efficient the GenAI inferencing process is. It reflects the system's ability to use resources such as energy efficiently while minimizing environmental impacts, including carbon emissions and waste heat. This factor captures how well the system integrates energy-efficient technologies and practices to reduce its overall environmental footprint.

Role in the Model
> Indicator of Sustainability: The Endogenous Efficiency Factor serves as a key indicator of the environmental sustainability of the GenAI inferencing process. It shows how efficiently energy is used during inferencing tasks and how much effort is put into minimizing negative environmental impacts.
> Influenced by System Efficiency: This factor is influenced by both endogenous constraints (e.g., hardware inefficiencies or power limitations) and system efficiency metrics like Power Usage Effectiveness (PUE). A highly efficient system with low PUE values will have a higher Sustainable factor, indicating better sustainability performance.
> Impact on Energy Consumption: Higher values of the Endogenous Sustainable Factor suggest that the system is operating efficiently with lower energy consumption per inference task (e.g., fewer joules per GenAI token). This helps reduce overall energy consumption (GenAI Inferencing Joules) and improves sustainability.

Scale (0 to 1)
> A value closer to 0 indicates poor sustainability performance, with high energy consumption and a significant environmental impact. This reflects inefficient use of resources, higher carbon emissions, and more waste heat.
> A value closer to 1 indicates a highly efficient and sustainable process with minimal environmental impact. This suggests that the system is using energy-efficient technologies and practices to minimize its carbon footprint

Impact on Model
These constraints affect variables such as the joules per GenAI token and overall energy efficiency (e.g., power usage effectiveness or PUE).

Calculation Method
> We identify key factors: Hardware evolution, software efficiency, data center design, operational practices, research and development, and workforce skills.
> We assign weights to these factors:
- Hardware Evolution: 0.25
- Software and Algorithmic Efficiency: 0.20
- Data Center Design and Infrastructure: 0.15
- Operational Practices and Management: 0.15
- Research, Development, and Education: 0.15
- Workforce and Skills: 0.10

> We calculate the weighted average: Endogenous Constraint = Σ (Factor Weight × Factor Score)
> We adjust the values over time based on scenario-specific trends and assumptions.

Efficiency Scores for Each Scenario:
> H1 (Sustainable AI): (0.25×0.8) + (0.20×0.7) + (0.15×0.6) + (0.15×0.6) + (0.15×0.5) + (0.10×0.5) = 0.64
• Hardware Evolution: High due to cutting-edge chips like NVIDIA A100, with performance doubling every 2.8 years
• Software Efficiency: Software improvements contribute to a 30% increase in performance per dollar each year
• Data Center Design: Leading Design and Operational practices allow 10% more energy-efficient each year
• Operational Practices: High with integrated energy-efficient practices.
• Research and Development: High investment in new technologies.
• Workforce Skills: High availability of skilled personnel.

> H2 (Limits to Growth): (0.25×0.1) + (0.20×0.1) + (0.15×0.2) + (0.15×0.2) + (0.15×0.1) + (0.10×0.2) = 0.13

> H3 / H4 (Abundance without Boundaries / AI Energy Crisis): (0.25×0.6) + (0.20×0.5) + (0.15×0.4) + (0.15×0.4) + (0.15×0.5) + (0.10×0.4) = 0.48

Crunch Factor Inferencing

The Crunch Factor in the context of GenAI Inferencing refers power availibilityavailability limitations that restrict the development and efficient scaling of GenAI systems.

Definition
This factor captures the constraints imposed by energy shortages or grid capacity issues, which limit how much computational power can be allocated to GenAI tasks. These limitations are particularly critical when the demand for AI inferencing grows rapidly, as it can lead to conflicts with other sectors of the economy that also rely on electricity.

Role in the Model
> Power Constraint on Growth: The Crunch Factor acts as a bottleneck that restricts how much GenAI systems can scale due to limited power availability. Even if demand for GenAI usage is high (as indicated by the CAGR Inference Growth Factor), a high Crunch Factor will limit the system's ability to meet this demand.
oImpact on Energy Efficiency: When power availability is constrained, it forces inefficient use of available resources, increasing energy consumption per task and slowing down overall system performance. This directly impacts nodes like Tokens per GenAI Output, Joules per GenAI Token, and ultimately, total energy consumption (Annual GenAI Inferencing TWh).
> Influence on Adoption: High Crunch Factor values can slow down the adoption of GenAI technologies by both industry and consumer users because power shortages make it difficult to maintain reliable AI services. This impacts nodes such as Actual Industry GenAI Users and Actual Number of GenAI Consumer Users.

Physical Examples of Crunch Factor Constraints

> Energy Shortages in Data Centers: Rapid growth in AI workloads can lead to energy shortages in data centers, especially in regions where grid capacity is already stretched thin.
> Power Grid Conflicts with Other Sectors: As AI's electricity demand rises, it may conflict with other critical sectors (e.g., healthcare, manufacturing) that also rely on stable power supplies.
> Regulatory Restrictions Due to Energy Crises: In scenarios like an "AI Energy Crisis," regulators may impose strict controls on AI development and deployment to manage energy consumption across sectors.

Scale (0 to 1)
> A value closer to 0 represents minimal or no power constraints, meaning there is sufficient electricity available for GenAI systems to scale without limitations.
> A value closer to 1 represents severe power constraints, where energy shortages significantly hinder the ability of GenAI systems to scale and operate efficiently.

Scenario 1/2/3 : Value = 0, suggesting there are no significant power availability limitations for GenAI inferencing.

Scenario 4: Value = 0.6, suggesting that while there are serious constraints, the situation is not yet at a complete gridlock (which would be closer to 1).

GenAI industry users

The number of GenAI industry usage represents the total count of businesses or organizations actively employing generative AI technologies in their operations, products, or services. This metric reflects the adoption and integration of GenAI across various industrial sectors, influenced by factors such as technological advancements, resource availability, regulatory environments, and market dynamics.

H1	2025: 58,000,000 2026: 70,000,000 2027: 84,000,000 2028: 101,000,000 2029: 121,000,000 2030: 145,000,000 2031: 174,000,000 2032: 209,000,000 2033: 251,000,000 2034: 289,000,000 2035: 324,000,000	This scenario aligns with McKinsey's 2024 State of AI report, which found that 65% of organizations are regularly using GenAI in at least one business function, nearly double from 10 months prior. The steady growth reflects increasing adoption across sectors, with IDC projecting worldwide AI spending to reach $154 billion by 2027 at a CAGR of 27%. 1. McKinsey & Company. (2024). The state of AI in 2024: Generative AI's breakout year. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-early-2024-gen-ai-adoption-spikes-and-starts-to-pay-off 2. Statista. (2024). Generative AI market size worldwide 2022-2032. https://www.statista.com/statistics/1365145/worldwide-generative-ai-market-size/
H2	2025: 58,000,000 2026: 65,000,000 2027: 72,000,000 2028: 80,000,000 2029: 88,000,000 2030: 97,000,000 2031: 106,000,000 2032: 116,000,000 2033: 127,000,000 2034: 139,000,000 2035: 152,000,000	This scenario reflects more constrained growth, aligning with Gartner's 2023 survey finding that 45% of organizations were piloting GenAI programs. The slower adoption rate considers challenges highlighted by Accenture, where 61% of companies report their data assets are not ready for GenAI, and 70% struggle to scale projects using proprietary data. 3. Bloomberg Intelligence. (2023). Generative AI Market Outlook. https://www.bloomberg.com/professional/blog/generative-ai-market-to-reach-1-3-trillion-by-2032/ 4. Grandview Research. (2024). Generative AI Market Size, Share & Trends Analysis Report. https://www.grandviewresearch.com/industry-analysis/generative-ai-market
H3	2025: 58,000,000 2026: 75,000,000 2027: 97,000,000 2028: 126,000,000 2029: 164,000,000 2030: 213,000,000 2031: 277,000,000 2032: 360,000,000 2033: 468,000,000 2034: 608,000,000 2035: 790,000,000	This rapid growth scenario is supported by Bloomberg Intelligence's projection that the GenAI market could reach $1.3 trillion by 2032. It also aligns with Accenture's finding that 98% of global executives believe AI foundation models will play an important role in their organizations' strategies in the next 3-5 years. 5. Salesforce. (2023). State of IT Report. https://www.salesforce.com/resources/research-reports/state-of-it/ 6. IDC. (2024). Worldwide Artificial Intelligence Spending Guide. https://www.idc.com/getdoc.jsp?containerId=prUS52530724
H4	2025: 58,000,000 2026: 70,000,000 2027: 84,000,000 2028: 101,000,000 2029: 121,000,000 2030: 145,000,000 2031: 130,000,000 2032: 117,000,000 2033: 105,000,000 2034: 95,000,000 2035: 85,000,000	The AI Energy Crisis scenario (H4) initially follows the growth trend seen in other scenarios but then experiences a sharp decline due to energy constraints. The forecast shows growth from 58 million users in 2025 to a peak of 145 million in 2030, followed by a decline to 85 million by 2035. The peak in 2030 reflects the maximum adoption before energy constraints become critical. The subsequent decline represents the impact of these constraints on AI adoption, supported by EPRI's 2024 study projecting that data centers could consume up to 9% of U.S. electricity generation by 2030. 7. Accenture. (2024). Reinventing Enterprise Operations with Gen AI. https://newsroom.accenture.com/news/2024/new-accenture-research-finds-that-companies-with-ai-led-processes-outperform-peers 8. Electric Power Research Institute (EPRI). (2024). Powering Intelligence Study. https://www.epri.com/research/products/000000003002025917

Gen AI consumer users

The number of consumer GenAI users represents the total count of individuals who regularly interact with or utilize generative AI technologies for personal or non-professional purposes. This includes users of AI-powered chatbots, content generation tools, personal assistants, and other consumer-facing generative AI applications.

H1	2025: 1.0 billion 2026: 1.8 billion 2027: 2.8 billion 2028: 4.0 billion 2029: 5.2 billion 2030: 6.0 billion 2031: 6.5 billion 2032: 6.8 billion 2033: 7.0 billion 2034: 7.1 billion 2035: 7.2 billion	This scenario envisions a balanced and sustainable growth of GenAI adoption, with user numbers increasing from 1.0 billion in 2025 to 6.0 billion in 2030, and reaching 7.2 billion by 2035. The rapid initial growth reflects the increasing accessibility and efficiency of GenAI technologies, aligning with the scenario's focus on sustainable AI development. This trajectory is supported by data points such as ChatGPT reaching 100 million monthly active users just two months after its launch in 2022, and having an estimated 180.5 million users worldwide by January 2024.
H2	2025: 1.0 billion 2026: 1.5 billion 2027: 2.0 billion 2028: 2.4 billion 2029: 2.7 billion 2030: 3.0 billion 2031: 3.2 billion 2032: 3.3 billion 2033: 3.4 billion 2034: 3.4 billion 2035: 3.5 billion	In this scenario, GenAI user growth is more constrained, increasing from 1.0 billion in 2025 to 3.0 billion in 2030, and reaching 3.5 billion by 2035. This slower adoption rate aligns with the scenario's theme of controlled growth and technocratic control, reflecting various limiting factors such as regulatory restrictions, data scarcity, and infrastructure limitations. The scenario is supported by data points such as the 35% of companies using AI in 2023, up from 25% in 2022, indicating growth but at a measured pace. The implementation of regulations like the EU AI Act and similar restrictions may limit certain AI applications, potentially increasing constraints. This scenario is further validated by research highlighting the significant computational costs of training large AI models, suggesting potential limits to scaling.
H3	2025: 1.0 billion 2026: 2.0 billion 2027: 3.5 billion 2028: 5.0 billion 2029: 6.5 billion 2030: 7.5 billion 2031: 8.2 billion 2032: 8.7 billion 2033: 9.0 billion 2034: 9.2 billion 2035: 9.3 billion	This scenario assumes rapid, unchecked growth in GenAI adoption, with user numbers surging from 1.0 billion in 2025 to 7.5 billion in 2030, and reaching 9.3 billion by 2035. The accelerated adoption rate reflects the scenario's theme of abundance and widespread AI deployment across all sectors. This trajectory is supported by data points such as McKinsey's research finding that GenAI features stand to add up to $4.4 trillion to the global economy annually, and the projection that the global AI market size will reach $2.25 trillion by 2030 from $428 billion in 2022. The scenario is further validated by the rapid adoption rates observed, with almost 40% of the U.S. population aged 18 to 64 using generative AI to some degree by August 2024.
H4	2025: 1.0 billion 2026: 1.7 billion 2027: 2.5 billion 2028: 3.0 billion 2029: 2.5 billion 2030: 2.0 billion 2031: 1.7 billion 2032: 1.5 billion 2033: 1.4 billion 2034: 1.3 billion 2035: 1.2 billion	This scenario depicts an initial rapid growth followed by a sharp decline in GenAI users, peaking at 3.0 billion in 2028 before decreasing to 2.0 billion in 2030 and further declining to 1.2 billion by 2035. This trajectory reflects the scenario's theme of an energy crunch and the resulting limitations on AI deployment and usage due to energy constraints and infrastructure challenges. The scenario is supported by data points such as Goldman Sachs Research forecasting data center power demand to grow 160% by 2030, potentially rising from 1-2% of overall power consumption to 3-4% by the decade's end. Local examples further validate this scenario, with data centers in Northern Virginia's "Data Center Alley" already consuming 25% of the region's electricity, and projections suggesting data centers could consume up to 32% of Ireland's electricity by 2026.

Gen AI consumer users

Sources for 2025 starting points

Generative AI Market Size and Growth:
Global value: $44.89 billion in 2023 (Statista)
Expected to exceed $66 billion by end of 2024 (Statista)
Projected to reach $1.3 trillion by 2032 (Bloomberg Intelligence)
North America leads with 40.2% of global revenue share (Grandview Research)

AI Adoption Rates:
65% of organizations regularly using generative AI in at least one business function (McKinsey, 2024)
72% overall AI adoption rate, up from about 50% in previous years (McKinsey, 2024)
92% of Fortune 500 firms have adopted generative AI (Exploding Topics, 2024)
45% of organizations piloting generative AI programs (Gartner, October 2023)

Consumer Usage:
Almost 40% of U.S. population ages 18 to 64 used generative AI to some degree (St. Louis Fed, August 2024)
53% of Americans have utilized generative models (Master of Code, 2024)
41% of regular users engage with AI daily (Master of Code, 2024)

Industry-Specific Adoption:
37% adoption in U.S. marketing and advertising industry (Statista, 2024)
51% of marketing specialists using or experimenting with generative AI (Master of Code, 2024)
One-third of salespeople applying or planning to use generative models (Master of Code, 2024)

Energy Consumption and Data Centers:
Data center electricity demand could grow 160% by 2030 (Goldman Sachs Research, 2023)
Data centers could consume 9% of U.S. electricity generation by 2030, double the current amount (EPRI, 2024)
U.S. data center load expected to grow from 19 GW in 2023 to 21 GW in 2024 (FERC, 2024)
Global data center electricity demand projected to more than double between 2022 and 2026, reaching over 1,000 TWh (IEA, 2023)

AI Energy Efficiency:
A ChatGPT query requires 2.9 watt-hours of electricity, compared to 0.3 watt-hours for a Google search (IEA, 2023)
Generating 1,000 images with Stable Diffusion XL produces as much CO2 as driving 4 miles in a gas-powered car (MIT Technology Review, 2024)

Future Projections:
85% of business leaders expect to use generative AI for low-value tasks by end of 2024 (MIT/Telstra, 2024)
95% of customer interactions may involve AI by 2025 (Exploding Topics, 2024)
AI could generate up to 97 million jobs by 2025 (Exploding Topics, 2024)

Joule per GenAI Token

Joule per GenAI Token refers to the amount of energy, measured in joules, required to process a single token during a Generative AI (GenAI) inferencing task. A token typically represents a word or part of a word in natural language processing tasks, and inferencing involves generating predictions or outputs based on a trained AI model.

This metric is critical for understanding the energy efficiency of AI models during inferencing. It reflects how much energy is consumed for each token processed by the model and is influenced by several factors, including:
> Model architecture: Different architectures (e.g., seq2seq vs. decoder-only) have varying computational and energy demands.
> Hardware efficiency: The type of GPU or TPU used for inferencing can significantly impact energy consumption.
>Task complexity: More complex tasks (e.g., reasoning vs. simple text completion) require more computational power, thus consuming more energy per token.
> Data center efficiency: The overall efficiency of the data center (measured by metrics like Power Usage Effectiveness, or PUE) also affects the energy consumed per token.

	Values	Rationale
H1	2025: 2.35 Joules 2026: 2.25 Joules 2027: 2.15 Joules 2028: 2.05 Joules 2029: 1.95 Joules 2030-2035: 1.85 Joules	In the Sustainable AI scenario, there is a strong focus on sustainability and energy efficiency. The gradual reduction in joules per token reflects the adoption of energy-efficient hardware, improvements in AI algorithms, and the use of renewable energy sources. The yearly reduction of approximately 4-5% shows continuous improvements in infrastructure, cooling technologies (e.g., liquid cooling), and optimization of inferencing tasks. By 2030, the efficiency gains start to plateau as physical limits are approached, but the focus remains on minimizing energy consumption.
H2	2025: 5.05 Joules 2026: 5.00 Joules 2027: 4.90 Joules 2028: 4.80 Joules 2029-2035: 4.70 Joules	In the Limits to Growth scenario, energy efficiency improves slowly due to resource constraints, regulatory restrictions, and slower adoption of advanced AI hardware and cooling technologies. The initial value is high (5.05 Joules) due to inefficiencies in data center operations and limited access to cutting-edge technology. The yearly reduction is modest (~1%), reflecting incremental improvements driven by regulatory pressures rather than proactive innovation. By 2035, the value remains relatively high (4.70 Joules) due to persistent inefficiencies and a lack of significant breakthroughs in energy-saving technologies.
H3	2025: 4.42 Joules 2026: 4.35 Joules 2027: 4.25 Joules 2028-2030: 4.20 Joules 2031-2035: 4.10 Joules	In the Abundance without Boundaries scenario, rapid expansion of AI workloads leads to increased energy consumption despite some improvements in hardware efficiency and algorithmic optimization. The initial value (4.42 Joules) reflects a balance between performance and energy use, but the focus on scaling AI models results in slower improvements compared to H1 (Sustainable AI). The yearly reduction is moderate (~2%), with efficiency gains primarily driven by technological advancements rather than energy-conscious practices. By 2030, efficiency improvements slow down as demand for larger models increases, leading to a plateau in energy savings.
H4	2025: 3.88 Joules 2026-2027: 3.85 Joules 2028-2030: 3.80 Joules 2031-2035: 3.75 Joules	In the AI Energy Crisis scenario, early signs of an energy crunch lead to a focus on reducing energy consumption through crisis-driven optimizations rather than proactive planning or innovation. The initial value (3.88 Joules) is relatively low due to immediate efforts to conserve energy as resources become scarce. However, improvements are slow (~1% yearly reduction), as the focus is on managing existing infrastructure rather than investing in new technologies or optimizing processes for long-term sustainability. By 2035, the value stabilizes at around 3.75 Joules due to continued crisis management but limited innovation.

Main sources

Epoch AI Blog on Predicting GPU Performance

SemiAnalysis Report on Nvidia Blackwell Performance Analysis

Tokens per GenAI Output

Tokens per GenAI Output refers to the number of tokens generated by a Generative AI (GenAI) model in response to a single input or prompt during inferencing. A token typically represents a word, part of a word, or a character, depending on the language model and the task. The number of tokens per output is an important metric as it directly influences the computational load and energy consumption of the model. More tokens generally require more computational resources, leading to higher energy consumption.

The Tokens per GenAI Output metric is influenced by:
> Model architecture: Larger models with more parameters may generate longer responses with more tokens.
> Task complexity: Complex tasks, such as answering questions in detail or generating long-form content, will result in more tokens.
> Optimization and efficiency: More efficient models or those optimized for specific tasks may generate fewer but more relevant tokens, reducing energy consumption.

H1	2025: 840 Tokens 2026: 860 Tokens 2027: 880 Tokens 2028: 900 Tokens 2029: 920 Tokens 2030-2035: 950 Tokens	In the Sustainable AI scenario, there's a significant focus on energy efficiency and sustainability. The gradual increase in tokens per output from 840 in 2025 to 950 by 2030 (and maintained through 2035) reflects successful optimization efforts. These improvements aim to enhance performance quality while managing computational load and energy consumption. Key factors contributing to this trend include: • Continuous improvements in model architecture, allowing for more efficient token generation while increasing output quality. • Task-specific optimizations that enable models to produce more comprehensive and relevant outputs. • Advancements in hardware, such as more efficient GPUs, supporting increased token processing capabilities with optimized energy consumption during inferencing tasks. • Evolving user expectations and increasing task complexity, necessitating a gradual increase in output tokens to provide more detailed and contextually rich responses.
H2	2025: 722 Tokens 2026: 725 Tokens 2027: 727 Tokens 2028: 731 Tokens 2029-2035: 735 Tokens	In the Limits to Growth scenario, resource constraints and regulatory restrictions hinder rapid optimization of models for token efficiency. The slow increase from 722 tokens in 2025 to 735 tokens by 2029 (maintained through 2035) shows a much slower pace of improvement compared to the Sustainable AI scenario. Key aspects of this scenario include: • Focus on maintaining functionality rather than aggressive optimization. • Regulatory restrictions potentially limiting the deployment of more advanced, efficient models. • Resource constraints slowing down research and development in AI efficiency. • A minimal yearly growth of about 0.4-0.5%, reflecting the challenges in improving model performance under strict limitations
H3 H4	2025: 900Tokens 2026: 950 Tokens 2027: 1000 Tokens 2028-2030: 1100 Tokens 2031-2035: 1200 Tokens	The Abundance without Boundaries and Energy Crisis scenarios demonstrate rapid scaling and optimization efforts, leading to significant increases in tokens per output. The rise from 900 tokens in 2025 to 1200 tokens by 2031 (maintained through 2035) showcases aggressive optimization aimed at maximizing performance and output quality. Key factors in this scenario include: • Substantial investments in AI research and development. • Rapid advancements in model architectures and task-specific optimizations. • Widespread adoption of cutting-edge hardware and infrastructure. • An average yearly growth of about 5% from 2025 to 2031, indicating a strong push for enhanced AI capabilities and output quality.

Actual industry GenAI users

Using the conversion table, we determine the weight by category.

Exogenous Gen AI Inferencing Economy and industry category	Value: 1.36	Value: 0.7	Value: 1.64	Value: 1.32
Exogenous Gen AI Inferencing Energy and material use category	Value: 1.36	Value: 0.64	Value: 1.02	Value: 1.02
Exogenous Gen AI Inferencing Governance and markets category	Value: 1.1	Value: 0.76	Value: 1.26	Value: 0.88
Exogenous Gen AI Inferencing Society and behavior category	Value: 1.22	Value: 0.68	Value: 1.62	Value: 1.18

GenAI Training Sub Model

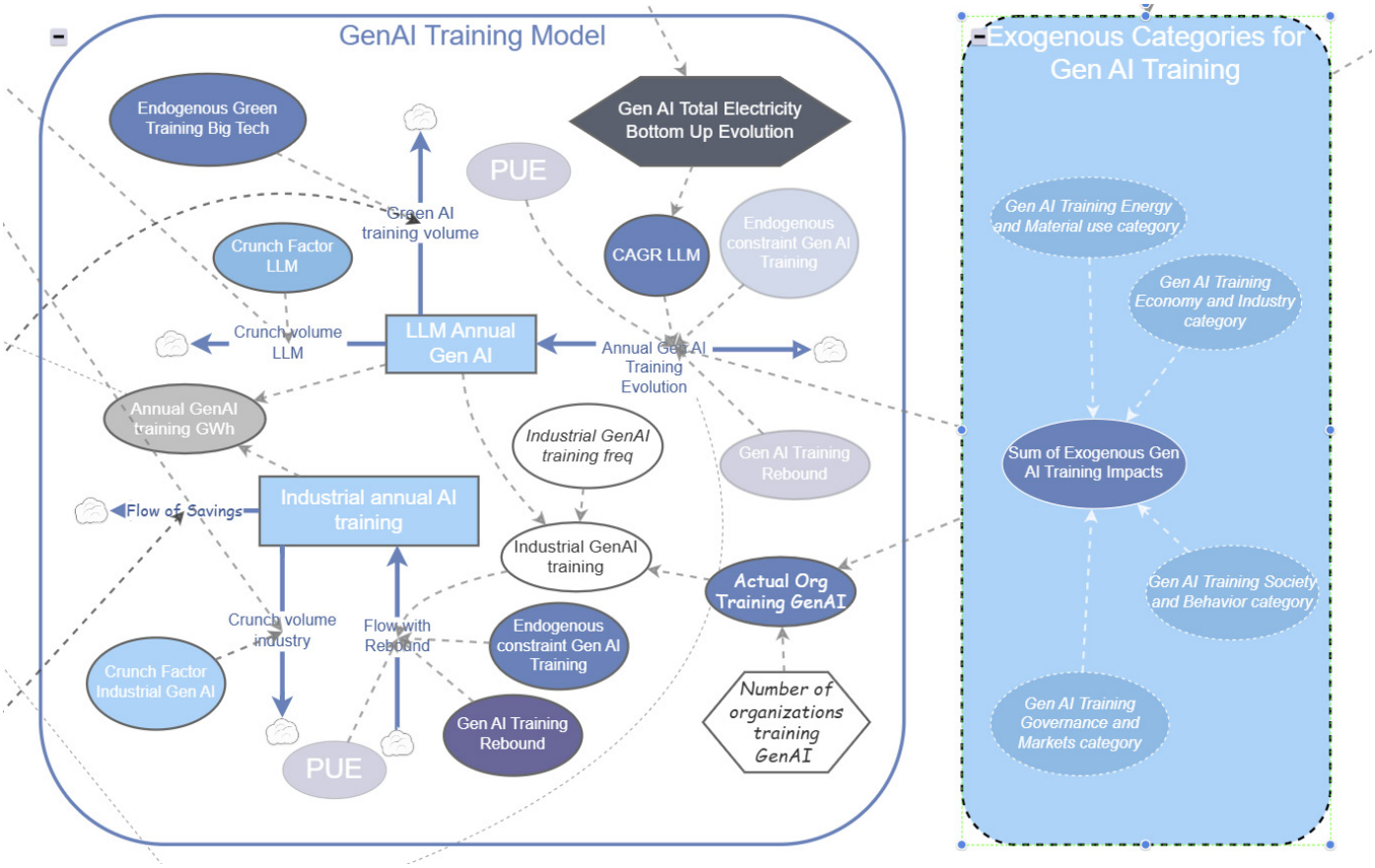


Figure 10: The GenAI training submodel

The GenAI Training Sub Model depicted in the image illustrates the relationships between various parameters that drive and constrain the electricity use evolution of Generative AI (GenAI) training. Here's a breakdown of the key components and their interrelations:

Key Components

• LLM Annual GenAI Training Electricity Use: This refers to the energy consumed during the development and training of large-scale generative AI models, particularly large language models (LLMs). Major technology companies utilize extensive computing resources, proprietary datasets, and specialized talent, leading to significant energy consumption and computational power demands. This is a major contributor to data center electricity demand with substantial environmental implications.

• Industrial Annual AI Training: Represents the total volume of AI training dedicated to industrial applications, involving energy-intensive processes that require significant computational resources. It is driven by factors like Industrial GenAI training frequency and Actual Org Training GenAI.

• Industrial GenAI Training Frequency: Refers to how often companies and industries train or update their generative AI models using proprietary data for specific industrial applications.

• Industrial Annual GenAI Training Electricity Use: The total energy consumed annually for industrial GenAI training activities.

• LLM Total Electricity Growth Evolution (CAGR LLM): Represents the compound annual growth rate of electricity use in training large language models, reflecting the increasing energy demands over time.

• Gen AI Training Endogenous Constraint: Internal limitations within data centers that restrict scaling up GenAI training, such as hardware capacity or energy efficiency issues.

GenAI Training Sub Model

• **Number of Organizations Training GenAI:** The total count of organizations involved in training generative AI models.

• **Crunch Factor Training LLM:** Indicates limitations in power availability affecting LLM training, highlighting challenges in scaling due to energy constraints.

• **Crunch Factor Industrial Training GenAI:** Similar to LLM Crunch Factor but specific to industrial applications, indicating power availability limitations affecting training scalability.

• **Exogenous GenAI Training Economy and Industry Category:** Captures external economic and industrial influences on GenAI training, impacting resource allocation and utilization in AI development.

• **Exogenous GenAI Training Energy and Material Use Category:** Represents external influences related to energy consumption and material usage in GenAI training, including resource availability and sustainability.

• **Exogenous GenAI Training Governance and Markets Category:** Refers to regulatory and market dynamics affecting GenAI training, including policies, market trends, and governance structures that enable or constrain AI development.

• **Exogenous GenAI Training Society and Behavior Category:** Encompasses societal attitudes and behaviors towards AI, influencing its adoption and integration into various sectors.

LLM Annual Gen AI training electricity use

To define the initial values for LLM annual GenAI training electricity, we utilize the EpochAI.ai database and construct a calculation methodology to determine those values. The EpochAI database identifies and tracks contemporary and historic advances in AI, collating key details across several areas. This research includes who developed models, when, and for what tasks, how much compute was used for training, how many parameters models have, how much data was used for training, what hardware was used for training.

Rationale
> **Model Scope:** Academic, Industry, Research
> **Fields:** Audio, Biology, Earth Science, Image Generation, Language, Mathematics, Multimodal approaches, Robotics, Search technologies, Speech processing, Video analysis, Vision systems, and 3D Modeling
> **Model Selection:** We wi'll focus on models released in 2023 or very late 2022, assuming their training occurred primarily in 2023.
> **Energy Conversion:** Convert training compute (in petaflop/s-days) to energy consumption (in TWh).
> **Efficiency Factor:** Apply an efficiency factor to account for cooling and other inefficiencies.
> **Undocumented Models:** Include an estimate for potential undocumented models.

Analysis of 2023 Models
> Categories:
> Language
> Multimodal
> Vision
> Audio/Speech
> Other (including Robotics, Biology, Earth Science, etc.)

Calculations for 2023
1. We focus on the models released in 2023:
  o GPT-4 (Released: 2023-03-14)
  o LLaMA (Released: 2023-02-24)
  o Claude (Released: 2023-03-14)
  o PaLM-2 (Released: 2023-05-10)

2. We categorize these models byased on their estimated training compute:
  o Very Large Models: > 1e25 flops
  o Large Models: 1e24 - 1e25 flops
  o Medium Models: 1e23 - 1e24 flops

Calculation Method: Electricity for Compute (TWh) = (Training compute) / (1e15 * 24 * 3600) * 24 * 3600 * 1000 / (3.6e12) * 1.5 / 1e9
Where 1.5 is the efficiency factor accounting for cooling and other overheads.

Categories
  o Language Models: Total compute: 1.14825E+26 flops. Electricity for Compute: 7.18 TWh
  o Multimodal Models: Total compute: 1.01E+25 flops. Electricity for Compute: 0.63 TWh
  o Vision Models: Total compute: 2.5E+24 flops. Electricity for Compute: 0.16 TWh
  o Audio/Speech Models: Total compute: 1E+24 flops. Electricity for Compute: 0.06 TWh
  o Other Models: Total compute: 5E+23 flops. Electricity for Compute: 8.09 TWh

Total Energy Consumption for 2023: 15 TWh (for all the 4 scenarios)

Calculation Method for 2025
We start with the 2023 baseline of 15 TWh.
  > Project forward two years (2023 to 2025) using the 4-5x annual growth rate.
    • Adjusted growth factor per year: 4x (slightly lower than the 4-5x from Epoch AI trends)
    • Two-year growth: 4^2 = 16x
    • Base energy consumption: 15 TWh * 16 = 240 TWh
    • Assume 35% improvement in energy efficiency over two years
    • Factor in increased adoption and more models being trained

Adjusted energy consumption: 240 TWh * 0.65 * 0.256 ≈ 40 TWh
  > Language Models: 55% of total / 40 TWh * 0.55 * 0.256 ≈ 22 TWh
  > Multimodal Models: 25% of total / 40 TWh * 0.25 = 10 TWh
  > Vision Models: 10% of total / 40 TWh * 0.10 = 4 TWh
  > Audio/Speech Models: 5% of total / 40 TWh * 0.05 = 2 TWh
  > Other Models: 5% of total / 40 TWh * 0.05 = 2 TWh

Industrial GenAI Training Frequency Ratio LLM Total Electricity Growth (CAGR LLM)

The initial value of 7 TWh for Industrial annual GenAI training electricity use in 2025 represents a baseline starting point across all scenarios, reflecting the current state of industrial AI applications and energy requirements of existing large language models. This conservative estimate allows for growth in all scenarios while considering improvements in data center efficiency already implemented by 2025. The value accounts for a mix of different training approaches used in industrial settings, including both energy-intensive large model training and more efficient fine-tuning of existing models. It also considers the global distribution of industrial AI training across various sectors such as manufacturing, logistics, and process optimization.

Industrial GenAI training frequency ratio

Rationale: Industrial GenAI training frequency represents how often companies update their generative AI models with proprietary data for specific industrial applications. The values (3.34, 2.17, 5.59, 5.05) correspond to the four scenarios: Sustainable AI, Limits to Growth, Abundance without boundaries, and AI Energy Crisis, respectively.

> Sustainable AI (3.34): This moderate frequency reflects a balanced approach to AI development. Companies regularly update models, but with a focus on efficiency and sustainability.
> Limits to Growth (2.17): The lowest frequency indicates constrained AI development due to various limitations (e.g., power availability, data scarcity, regulatory restrictions). Companies focus on essential updates rather than constant retraining, reflecting resource constraints.
> Abundance without boundaries (5.59): The highest frequency represents unchecked growth and rapid AI deployment. Companies constantly update models, leveraging increased efficiency to train more frequently. This aligns with the scenario's emphasis on technological optimism and abundance.
> AI Energy Crisis (5.05): The second-highest frequency initially seems counterintuitive but represents the period leading up to the energy crisis. Companies aggressively train and update models without considering long-term consequences, contributing to the impending crisis.

LLM Total Electricity Growth Evolution (Gen AI Total Electricity Bottom Up Evolution) (CAGR LLM)

> The LLM Total Electricity Growth Evolution, also referred to as Gen AI Total Electricity Bottom Up Evolution or CAGR LLM, represents an initial trend input to the system dynamics model. This input serves as a starting point for the model's calculations, providing a baseline growth trajectory for LLM electricity consumption. However, it's crucial to understand that this initial input is not static throughout the simulation. The system dynamics approach allows for dynamic adjustments and feedback loops that can modify this growth trajectory over time. As the model runs, various factors and interactions within the system can cause the actual growth rates to deviate from these initial values.

The model takes these initial growth rates and processes them through its complex network of interrelated variables, including:
> Endogenous factors within the data center (e.g., hardware efficiency, cooling systems)
> Exogenous factors outside the data center (e.g., economic conditions, regulations, social acceptance)
> Feedback loops and reinforcing or balancing mechanisms

	Symbiotic-sustainable AI	Constraint Growth	Abundance	AI Energy Crisis
2025	60%	60%	60%	70%
2026	50%	50%	55%	80%
2027	40%	40%	50%	90%
2028	30%	30%	45%	100%
2029	20%	20%	40%	50%
2030	15%	10%	35%	-20%
2031	10%	5%	30%	-40%
2032	8%	2%	25%	-30%
2033	6%	1%	20%	-20%
2034	5%	1%	18%	-10%
2035	4%	1%	16%	-5%

Gen AI Training Endogenous Constraint

Internal limitations within data centers that restrict scaling up GenAI training, such as hardware capacity or energy efficiency issues. Values indicate the degree of constraint.

Main hypothesis for weightings
> Hardware Evolution (0.25): According to the Epoch AI trends dashboard, training compute for frontier AI models has been growing by 4-5x per year from 2010 to May 2024.
> Software and Algorithmic Efficiency (0.20): The Epoch AI dashboard indicates that algorithmic progress in language models is equivalent to doubling computational power every 5 to 14 months.
> Data Center Design and Infrastructure (0.15): A report from Schneider Electric highlights that AI data centers are facing challenges in power distribution and cooling for high-density computing needs.
> Operational Practices and Management (0.15): Schneider Electric emphasizes the need for advanced software to manage higher densities and mitigate downtime risk in AI data centers.
•Research, Development, and Education (0.15): A 2024 survey indicates that 81% of IT professionals think they can use AI, but only 12% actually have the skills to do so.
> Workforce and Skills (0.10): Reuters reports that there will be a 50% hiring gap for AI-related positions this year.

Calculation of Gen AI Training Endogenous Constraint scores for each scenario

Scenario 1 (Symbiotic-sustainable AI, 0.69):
- o Hardware Evolution: 0.7 (high growth but constrained by efficiency)
- o Software Efficiency: 0.5 (significant improvements)
- o Data Center Design: 0.7 (advanced but limited by sustainability)
- o Operational Practices: 0.6 (well-optimized)
- o Research and Development: 0.8 (high investment in efficiency)
- o Workforce Skills: 0.7 (focus on upskilling)
- > Calculation: (0.25×0.7)+(0.20×0.5)+(0.15×0.7)+(0.15×0.6)+(0.15×0.8)+(0.10×0.7) = 0.69

Scenario 2 (ConstraintLimits to Growth, 0.77)
- o Hardware Evolution: 0.8 (limited by constraints)
- o Software Efficiency: 0.7 (improvements slowed by limitations)
- o Data Center Design: 0.8 (constrained by resources)
- o Operational Practices: 0.7 (limited by regulations)
- o Research and Development: 0.8 (focused on overcoming limits)
- o Workforce Skills: 0.8 (skill shortages)
- > Calculation: (0.25×0.8)+(0.20×0.7)+(0.15×0.8)+(0.15×0.7)+(0.15×0.8)+(0.10×0.8) = 0.77

Scenario 3 (Abundance without boundaries, 0.54)
- o Hardware Evolution: 0.4 (rapid advancements)
- o Software Efficiency: 0.5 (focus on capabilities over efficiency)
- o Data Center Design: 0.6 (expansive but not optimized)
- o Operational Practices: 0.7 (challenged by rapid growth)
- o Research and Development: 0.5 (high investment, less focused)
- o Workforce Skills: 0.6 (large workforce, varying skill levels)
- > Calculation: (0.25×0.4)+(0.20×0.5)+(0.15×0.6)+(0.15×0.7)+(0.15×0.5)+(0.10×0.6) = 0.535

Scenario 4 (AI Energy Crisis, 0.70)
- o Hardware Evolution: 0.8 (constrained by energy crisis)
- o Software Efficiency: 0.7 (focus on energy efficiency)
- o Data Center Design: 0.7 (strained by energy demands)
- o Operational Practices: 0.6 (challenged by crisis management)
- o Research and Development: 0.7 (redirected to crisis solutions)
- o Workforce Skills: 0.6 (skills gap exacerbated by crisis)
- > Calculation: (0.25×0.8)+(0.20×0.7)+(0.15×0.7)+(0.15×0.6)+(0.15×0.7)+(0.10×0.6) = 0.705

Gen AI Training Rebound
Number of Organizations Training GenAI

Gen AI Training Rebound

The Gen AI Training Rebound describes how efficiency improvements can lead to increased usage, potentially offsetting energy savings. It reflects the rebound effect, where efficiency improvements in AI training can lead to increased usage, potentially offsetting energy savings. This phenomenon is observed across various sectors and technologies, where enhanced efficiency can paradoxically result in greater overall energy consumption.

Rationales
> Scenario 1 (Symbiotic-sustainable AI, Value: 1): No Rebound
This value of 1 represents a scenario where efficiency improvements in AI training are exactly balanced by increased usage, resulting in no net change in energy consumption. It assumes that any gains in efficiency are fully offset by increased demand for AI training resources. This is a common baseline scenario reflecting typical expectations of the rebound effect.
> Scenario 2 (ConstraintLimits to Growth, 1.099): Moderate Rebound
A value of 1.099 suggests a moderate rebound effect, where efficiency improvements lead to slightly more than proportional increases in AI usage and energy consumption. This could occur in scenarios where advancements in AI hardware and software make training more accessible and cost-effective, encouraging more frequent updates and expansions of AI models.
> Scenario 3 and 4 (Abundance without boundaries and AI Energy Crisis, 1.15): High Rebound
This higher value indicates a significant rebound effect, where efficiency gains lead to substantial increases in AI training activity and energy use. In this scenario, the reduced costs and improved capabilities drive widespread adoption and scaling of AI models, resulting in a marked increase in electricity consumption despite efficiency improvements. Number of organizations training GenAI : Data showing the growth in organizations involved in GenAI training over time, indicating widespread adoption and its impact on electricity demand.

Number of Organizations Training Gen AI

The Number of organizations training GenAI refers to the total count of companies, institutions, and entities actively engaged in training generative AI models using their own data and resources. This parameter reflects the adoption and development of GenAI technologies across various sectors of the economy.

Rationales
> Scenario 1 (Sustainable AI): As this scenario represents a balanced approach to AI adoption, the number of organizations training GenAI grows steadily but sustainably from 2025 to 2035. This growth reflects increasing AI adoption across industries, tempered by efficiency improvements and responsible practices. The scenario shows a compound annual growth rate (CAGR) of about 10% over the 10-year period, resulting in an increase from 20,000 organizations in 2025 to 52,000 by 2035.
> Scenario 2 (Limits to Growth): Slower growth due to various constraints (like power, data, and regulations, etc.) Reaches 30,000 organizations by 2035, reflecting limited expansion.
> Scenario 3 (Abundance without boundaries): Rapid, unchecked growth due to widespread AI adoption and fewer restrictions. Reaches 127,000 organizations by 2035, more than doubling the Sustainable AI scenario.
> Scenario 4 (AI Energy Crisis): Initial rapid growth followed by a sharp decline after 2030. Peaks at 45,000 in 2030, then drops to 22,000

Year	Symbiotic-sustainable AI	Limits To Growth	Abundance	AI Energy Crisis
2025	20,000	20,000	20,000	20,000
2026	22,000	21,000	24,000	23,000
2027	24,000	22,000	29,000	27,000
2028	26,000	23,000	35,000	32,000
2029	29,000	24,000	42,000	38,000
2030	32,000	25,000	51,000	45,000
2031	35,000	26,000	61000	40,000
2032	39,000	27,000	73,000	34,000
2033	43,000	28,000	88,000	29,000
2034	47,000	29,000	106,000	25,000
2035	52,000	30,000	127,000	22,000

Crunch factor training LLM
Exogenous Factors per Category for Training

The Crunch Factor Training LLM indicates limitations in power availability affecting LLM training, highlighting challenges in scaling due to energy constraints.

This parameter highlights the challenges organizations face in scaling AI capabilities due to energy constraints, reflecting a growing concern about the sustainability of AI infrastructure. The values assigned to this factor—0, 0, 0, and 0.3—indicate varying degrees of energy constraints across different scenarios.

Scenario 1/2/3 (Sustainable AI): 0 indicates an ideal scenario where power availability is not a limiting factor, allowing organizations to scale their training efforts without concern for energy constraints.

Scenario 4 (AI Energy Crisis): 0.3 reflects a high constraint, indicating that organizations will face some challenges regarding power availability. This could be due to regional differences in energy supply or temporary fluctuations in demand.

Exogenous Factors per Catergory for Training

These factors represent external influences on AI training processes, categorized into four main areas. They are derived from an adapted list of exogenous factors specifically relevant to AI training. The factors are measured on a scale from 0 to 2, where:
> 0 represents a highly negative impact
> 1 represents a neutral value (no impact of the exogenous category)
> 2 represents a highly positive impact

Symbiotic-sustainable AI Scenario: In the Sustainable AI scenario, exogenous factors generally support sustainable AI development. The Economy and Industry category (1.3) indicates strong economic support for Sustainable AI initiatives. Energy and Material use (1.14) suggests efficient resource utilization aligned with sustainability goals. The Governance and Markets category (1.16) reflects supportive policies and market conditions for environmentally responsible AI. Society and Behavior (1.18) shows positive societal attitudes towards sustainable AI development. These factors collectively create an environment conducive to balancing AI advancement with ecological considerations.

Limits to Growth Scenario: The Limits to Growth scenario is characterized by constraining exogenous factors. The low Economy and Industry value (0.6) indicates significant economic limitations on AI development. Energy and Material use (0.78) suggests resource scarcity and constraints. The Governance and Markets category (0.7) points to restrictive policies limiting AI expansion. The very low Society and Behavior value (0.58) reflects societal skepticism or resistance towards AI growth. These factors combine to create a challenging environment for AI development, with multiple external constraints limiting expansion.

Abundance without boundaries Scenario: In this scenario, exogenous factors generally support rapid AI expansion. The high Economy and Industry value (1.3) indicates strong economic backing for AI development. However, the slightly lower Energy and Material use (0.88) might suggest some resource challenges despite overall abundance. The Governance and Markets category (1.16) reflects favorable conditions for AI growth. The high Society and Behavior value (1.22) indicates strong societal enthusiasm for AI technologies. This combination of factors creates an environment of rapid, potentially unchecked AI development.

AI Energy Crisis Scenario: The AI Energy Crisis scenario shows mixed exogenous factors leading to potential instability. The Economy and Industry category (0.9) suggests initial support followed by challenges. The high Energy and Material use value (1.26) indicates intense resource consumption, potentially triggering the crisis. The Governance and Markets category (1.02) might reflect initial supportive conditions followed by reactive policies. The Society and Behavior value (1.06) suggests initial acceptance potentially shifting as the crisis unfolds. These factors collectively create a volatile environment where initial rapid AI growth leads to resource strain and potential crisis.

Exogenous Gen AI Training Economy and Industry category	Value: 1.3	Value: 0.6	Value: 1.3	Value: 0.9
Exogenous Gen AI Training Energy and Material use category	Value: 1.14	Value: 0.78	Value: 0.88	Value: 1.26
Exogenous Gen AI Training Governance and Markets category	Value: 1.16	Value: 0.7	Value: 1.16	Value: 1.02
Exogenous Gen AI Training Society and Behavior category	Value: 1.18	Value: 0.58	Value: 1.22	Value: 1.06

Global Model

Systemic Efficiency Effect

Definition: A measure of the environmental efficiency and sustainability of AI systems.

Values per Scenario:
> H1: 0.4
> H2: 0.108
> H3: 0.3
> H4: 0.044
Rationale: Higher values in scenarios like H1 and H3 reflect significant improvements in system-wide energy efficiency through optimizations in hardware, cooling, and operational practices. Lower values in H2 and H4 suggest more limited gains due to slower adoption of energy-efficient technologies.

PUE (Power Usage Effectiveness)

Definition: A metric that measures the energy efficiency of data centers by comparing total facility energy to IT equipment energy.

Values per Scenario (assumed fixed for the study's forecasts; however, some time series variations are possible)
> H1: 1.168
> H2: 1.318
> H3: 1.336
> H4: 1.36
Rationale: Lower PUE values (e.g., H1) indicate more efficient data center operations with less energy wasted on cooling and overheads. Higher PUE values in scenarios like H4 reflect less efficient energy use, possibly due to older infrastructure or less optimized cooling systems.

Flexibility Factor

Definition: Represents the capacity of workloads to be used flexibly for both AI training and inferencing.
Values per Scenario:
> H1: 0.16
> H2: 0.04
> H3: 0.2
> H4: 0
Rationale: Higher flexibility factors (e.g., H3) suggest that workloads can be dynamically shifted between training and inferencing tasks, optimizing resource use and reducing idle time. Lower values (e.g., H2 and H4) indicate less flexibility, leading to potential inefficiencies in workload management.

Data Center Model

CAGR (Compound Annual Growth Rate) According to IEA:
> Value: 0.03 (i.e., 3% per year)
> Definition: This represents the compound annual growth rate of global electricity delivery capacity. A CAGR of 3% indicates that the total electricity delivery capacity is expected to grow by 3% annually. This growth is driven by factors such as increased electrification, renewable energy adoption, and rising energy demand from sectors like AI and data centers.
> Source: The IEA's World Energy Outlook projects that global electricity demand will grow at an average rate of around 2.7% per year through 2030, driven by electrification in transport, industry, and buildings. The 3% CAGR used here is consistent with these projections and reflects a slightly higher growth rate due to additional factors like AI-driven compute demands.

Total Data Center Electricity Use:
> Initial Value: 523 TWh
> Definition: This represents the total electricity consumption across all types of data centers, including both traditional and more modern hyperscale or cloud-based facilities. The value of 523 TWh reflects the combined energy use of all data centers globally at the initial point in time.
> Source : Internal Schneider Electric Study.
> Supporting Source: According to the IEA's report, global data center electricity use was estimated at 460 TWh in 2022, with projections suggesting it could rise to between 620 TWh and 1,050 TWh by 2026, depending on deployment rates and efficiency improvements.

Electricity Model

Total Electricity Delivery Capacity:
> Initial Value: 25,000 TWh
> Definition: This represents the total electricity delivery capacity globally, measured in terawatt-hours (TWh). It encompasses all sources of electricity generation, including fossil fuels, renewables, and nuclear power. This value is likely based on global electricity production capacity as of the base year.
> Source: According to the International Energy Agency (IEA), global electricity generation was approximately 26,823 TWh in 2021, with projections suggesting continued growth due to increasing energy demand from sectors like data centers and electric vehicles. The value of 25,000 TWh aligns closely with this figure and represents an estimate for a slightly earlier period or a conservative baseline.

CAGR (Compound Annual Growth Rate) According to IEA:
> Value: 0.033 (i.e.,or 3.3% per year)
> Definition: This represents the compound annual growth rate of global electricity delivery capacity. A CAGR of 3% indicates that the total electricity delivery capacity is expected to grow by 3.3% annually. This growth is driven by factors such as increased electrification, renewable energy adoption, and rising energy demand from sectors like AI and data centers.
> Source: The study assumes a 3.3% Compound Annual Growth Rate (CAGR) for total electricity generation from 2023 to 2030. This projection aligns with the International Energy Agency's (IEA) 2024 report, which presents a similar growth rate under their Stated Policies Scenario. The slightly higher rate used in this study accounts for additional factors, such as the increasing computational demands driven

Traditional AI Model

Key Data Points for Traditional AI

Pre-Deep Learning Compute Growth (Traditional AI)
> Growth Rate: 1.5x/year (from 1956 to 2010)
> Rationale: Traditional AI models, which were prevalent before the deep learning era, saw compute requirements grow at a slower rate compared to modern Gen AI models. This growth rate reflects incremental improvements in algorithms and hardware over time.

Hardware Trends for Traditional AI
> Computational Performance: Traditional hardware improvements for AI tasks grew more slowly compared to modern GPUs. The computational performance of traditional CPUs used for AI tasks likely improved at a rate of around 1.2x/year, driven by Moore's Law and gradual increases in transistor density.
> Memory Capacity & Bandwidth: Memory capacity and bandwidth improvements were also slower, likely growing at around 1.1x/year during the traditional AI era, reflecting the more limited demands of earlier AI models.

Energy Efficiency in Traditional AI
Energy efficiency gains in traditional AI systems were primarily driven by improvements in hardware (e.g., CPUs), cooling systems, and algorithmic optimizations. However, these gains were modest compared to the rapid advancements seen in Gen AI systems today.

Updated Rationale for Energy Efficiency Savings in Traditional AI (Based on Scenarios)
Savings Distribution for Traditional AI Training: Values: 0.4, 0.1, 0.18, 0.08 represent the yearly energy efficiency savings across four scenarios:
> H1 (40%): Significant energy savings are achieved through hardware upgrades (e.g., newer CPUs) and algorithmic optimizations.
> H2 (10%): Limited energy savings due to slower adoption of more efficient hardware or optimizations.
> H3 (18%): Moderate energy savings from incremental hardware improvements and operational efficiencies.
> H4 (8%): Minimal energy savings due to reliance on legacy systems with outdated hardware and cooling infrastructure.

Savings Distribution for Traditional AI Inferencing:
Values: 0.34, 0.13, 0.06, 0.06 represent the yearly energy efficiency savings across four scenarios
> H1 (34%): Significant inferencing savings due to optimized algorithms and improved hardware.
> H2 (13%): Moderate inferencing efficiency gains from partial adoption of newer technologies.
> H3 & H4 (6%): Minimal inferencing savings due to older infrastructure and less efficient algorithms.

References related to the Appendices (1/3)

1. S. Samoili, M. López Cobo, B. Delipetrev, F. Martínez-Plumed, E. Gómez, G. De Prato, Europäische Kommission Gemeinsame Forschungsstelle, “AI watch defining artificial intelligence 2.0 : towards an operational definition and taxonomy for the AI landscape” (Brussels, 2020); https://doi.org/10.2760/382730 (online).

2. L. Senn-Kalb, D. Mehta, “artificial intelligence: in-depth market analysis” (London, 2023); https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=artificial+intelligence%3A+in-depth+market+analysis&btnG=.

3. H. Y. Lin, Embedded Artificial Intelligence: Intelligence on Devices. Computer (Long Beach Calif) 56, 90–93 (2023).

4. S. J. Russell, P. Norvig, Artificial Intelligence: A Modern Approach (Pearson Education Limited, Boston (MA), third edit., 2016).

5. B. P. Bloomfield, The Question of Artificial Intelligence: Philosophical and Sociological Perspectives (Routledge, 2018).

6. S. Raisch, S. Krakowski, Artificial Intelligence and Management: The Automation–Augmentation Paradox. Academy of management review 46, 192–210 (2021).

7. S. Sarkar, A. Naug, R. Luna Gutierrez, A. Guillen, V. Gundecha, R. Babu, C. Bash, H. Packard Enterprise, “Real-time Carbon Footprint Minimization in Sustainable Data Centers with Reinforcement Learning” in 37th Conference on Neural Information Processing Systems (2023; https://s3.us-east-1.amazonaws.com/climate-change-ai/papers/neurips2023/28/paper.pdf).

8. R. Istrate, V. Tulus, R. N. Grass, L. Vanbever, W. J. Stark, G. Guillén-Gosálbez, The environmental sustainability of digital content consumption. Nature Communications 2024 15:1 15, 1–11 (2024).

9. D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, A. S. Luccioni, T. Maharaj, E. D. Sherwin, S. K. Mukkavilli, K. P. Kording, C. P. Gomes, A. Y. Ng, D. Hassabis, J. C. Platt, F. Creutzig, J. Chayes, Y. Bengio, Tackling Climate Change with Machine Learning. ACM Comput Surv 55, 96 (2023).

10. W. B. Rouse, AI as Systems Engineering Augmented Intelligence for Systems Engineers. INSIGHT 23, 52–54 (2020).

11. S. Luccioni, Y. Jernite, E. Strubell, Power Hungry Processing: Watts Driving the Cost of AI Deployment? 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024, 85–99 (2024).

12. D. Richins, D. Doshi, M. Blackmore, A. Thulaseedharan Nair, N. Pathapati, A. Patel, B. Daguman, D. Dobrijalowski, I. Poland RAMESH ILLIKKAL, K. Long, D. Zimmerman, U. Vijay Janapa Reddi, D. Doshi, M. Blackmore, A. T. Nair, N. Pathapati, A. Patel, B. Daguman, D. Dobrijalowski, R. Illikkal, K. Long, D. Zimmerman, R. Illikkal, V. Janapa Reddi, AI Tax. ACM Transactions on Computer Systems (TOCS) 37, 1–4 (2021).

13. V. Avelar, P. Donovan, P. Lin, W. Torell, M. Torres Arango, “The AI disruption: Challenges and guidance for data center design.”

14. EpochAI.org, Data on the Trajectory of AI | Epoch AI Databases. https://epochai.org/data.

15. S. Borsci, V. V. Lehtola, F. Nex, M. Y. Yang, E. W. Augustijn, L. Bagheriye, C. Brune, O. Kounadi, J. Li, J. Moreira, J. Van Der Nagel, B. Veldkamp, D. V. Le, M. Wang, F. Wijnhoven, J. M. Wolterink, R. Zurita-Milla, Embedding artificial intelligence in society: looking beyond the EU AI master plan using the culture cycle. AI Soc 1, 1–20 (2022).

16. D. Patterson, J. Gonzalez, U. Holzle, Q. Le, C. Liang, L. M. Munguia, D. Rothchild, D. R. So, M. Texier, J. Dean, The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. Computer (Long Beach Calif) 55, 18–28 (2022).

17. C. Freitag, M. Berners-Lee, K. Widdicks, B. Knowles, G. S. Blair, A. Friday, The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. Patterns 2, 100340 (2021).

18. A. De Vries, The growing energy footprint of artificial intelligence. Joule 7, 2191–2194 (2023).

19. L. Belkhir, A. Elmeligi, Assessing ICT global emissions footprint: Trends to 2040 & recommendations. J Clean Prod 177, 448–463 (2018).
20. D. H. D. L. Meadows, J. Randers, W. W. Behrens, “The limits to growth” in Sustainable Planet Blues: Critical Perspectives on Global Environmental Politics (Taylor and Francis, 2018), pp. 25–29.

21. M. Koot, F. Wijnhoven, Usage impact on data center electricity needs: A system dynamic forecasting model. Appl Energy 291, 116798 (2021).

22. E. Masanet, A. Shehabi, N. Lei, S. Smith, J. Koomey, Recalibrating global data center energy-use estimates. Science (1979) 367, 984–986 (2020).

23. N. Crafts, Crafts, Nicholas, Artificial intelligence as a general-purpose technology: an historical perspective. Oxf Rev Econ Policy 37, 521–536 (2021).

References related to the Appendices (2/3)

24. P. Dauvergne, Is artificial intelligence Sustainableing global supply chains? Exposing the political economy of environmental costs. Rev Int Polit Econ 29, 696–718 (2022).

25. International Energy Agency -IEA, Global Energy and CO2 Status Report 2018. [Preprint] (2019). https://www.eenews.net/assets/2019/03/26/document_cw_01.pdf.

26. T. Van der Vorst, M. Massop, A. Smeitink, "De digitale voetafdruk-Emissies van de digitale sector in Nederland in (toekomst) perspectief" (Utrecht, 2023); https://www.rijksoverheid.nl/documenten/rapporten/2023/09/28/dialogic-de-digitale-voetafdruk-emissies-van-de-digitale-sector-in-nederland-in-toekomst-perspectief.

27. J. Morecroft, Strategic Modelling and Business Dynamics: A Feedback Systems Approach (Wiley, New York, 2007).

28. Y. Barlas, Formal aspects of model validity and validation in system dynamics. Syst Dyn Rev 12, 183–210 (1996).

29. P. M. Senge, J. D. Sterman, Systems thinking and organizational learning: Acting locally and thinking globally in the organization of the future. Eur J Oper Res 59, 137–150 (1992).

30. S. T. March, G. F. Smith, Design and natural science research on information technology. Decis Support Syst 15, 251–266 (1995).

31. G. Burrell, G. Morgan, Sociological Paradigms and Organisational Analysis: Elements of the Sociology of Corporate Life (Routledge, 2017).

32. H. Son, The history of Western futures studies: An exploration of the intellectual traditions and three-phase periodization. Futures 66, 120–137 (2015).

33. M. Chiasson, E. Davidson, J. Winter, Philosophical foundations for informing the future(S) through IS research. European Journal of Information Systems 27, 367–379 (2018).

34. U. Schultze, D. E. Leidner, Studying knowledge management in information systems research: discourses and theoretical assumptions. MIS quarterly, 213–242 (2002).

35. F. Wijnhoven, N. De Bruijn, R. Effing, "Google Trends Forecasting of Youth Unemployment" in CARMA 2024 6th Int. Conf. on Advanced Research Methods and Analytics (Universitat Politecnica de Valencia, Valencia, 2024; hhttp://ocs.editorial.upv.es/index.php/CARMA/CARMA2024/paper/viewFile/17158/8907).

36. J. Morecroft, "System Dynamics" in Systems Approaches to Making Change: A Practical Guide (Springer London, London, 2020), pp. 22–88.

37. J. Freire-González, Governing Jevons' Paradox: Policies and systemic alternatives to avoid the rebound effect. Energy Res Soc Sci 72 (2021).

38. A. Größler, J. H. Thun, P. M. Milling, System Dynamics as a Structural Theory in Operations Management. Prod Oper Manag 17, 373–384 (2008).

39. Y. Fang, K. H. Lim, Y. Qian, B. Feng, System dynamics modeling for information systems research: Theory development and practical application. MIS Q 42, 1303–1329 (2018).

40. D. L. H. Meadows, D. L. H. Meadows, J. Randers, W. Behrens, The Limits to Growth - Club of Rome (Potomac Associates, 1972; https://clubofrome.org/publication/the-limits-to-growth/).

41. S. E. Werners, E. Sparkes, E. Totin, N. Abel, S. Bhadwal, J. R. A. Butler, S. Douxchamps, H. James, N. Methner, J. Siebeneck, L. C. Stringer, K. Vincent, R. M. Wise, M. G. L. Tebboth, Advancing climate resilient development pathways since the IPCC's fifth assessment report. Environ Sci Policy 126, 168–176 (2021).

42. P. J. H. Schoemaker, When and how to use scenario planning: a heuristic approach with illustration. J Forecast 10, 549–564 (1991).

43. S. C. Aykut, Reassembling energy policy: Models, forecasts, and policy change in Germany and France. Science & Technology Studies 32, 13–35 (2019).

44. J. Galtung, Empiricism, criticism, constructivism. Synthese 1972 24:3 24, 343–372 (1972).

45. E. Trincado Aznar, J. María Vindel, "Energy Efficiency, Productivity and the Jevons Paradox" in Science, Technology and Innovation in the History of Economic Thought, E. Trincado Aznar, F. Lopez Castellano, Eds. (Springer Nature, 2023; https://doi.org/10.1007/978-3-031-40139-8_6).

References related to the Appendices (3/3)

46. J. D. Sterman, All models are wrong: reflections on becoming a systems scientist. Syst Dyn Rev 18, 501–531 (2002).

47. J. Duggan, System Dynamics Modeling with R. An Introduction to System Dynamics System Dynamics Modeling with R, 1–24 (2016).

48. L. Schoenenberger, A. Schmid, R. Tanase, M. Beck, M. Schwaninger, Structural Analysis of System Dynamics Models. Simul Model Pract Theory 110, 102333 (2021).

49. P. M. Senge, The Fifth Discipline (CPI Group, London, 1990).

50. I. Hristoski, P. Mitrevski, Evaluation of Business-Oriented Performance Metrics in eCommerce using Web-based Simulation. Journal of Emerging research and solutions in ICT 1, 1–16 (2016).

51. P. Fusch, G. E. Fusch, L. R. Ness, Denzin's paradigm shift: Revisiting triangulation in qualitative research. Journal of Social Change 10, 2 (2018).

52. N. K. Denzin, Triangulation. [Preprint] (2015). https://doi.org/10.1002/9781405165518.wbeost050.pub2.

53. F. Wijnhoven, M. Brinkhuis, Internet information triangulation: Design theory and prototype evaluation. J Assoc Inf Sci Technol 66, 684–701 (2015).

54. F. Wijnhoven, F. Kluitenberg, M. Daneva, "Open Source Software Information Triangulation: A design science study" in 28th International Conference on Information Systems Development, ISD 2019 (2019).

55. F. Wijnhoven, The Hegelian inquiring system and a critical triangulation tool for the Internet information slave: A design science study. Journal of the American Society for Information Science and Technology 63, 1168–1182 (2012).

56. R. D. Galliers, On Confronting Some of the Common Myths of Information : Systems Strategy Discourse. Strategic Information Management, 56–70 (2020).

57. P. J. Idenburg, Four styles of strategy development. Long Range Plann 26, 132–137 (1993).

58. Y. E. Chan, B. H. Reich, IT alignment: what have we learned? Journal of Information technology 22, 297–315 (2007).

59. J. Luftman, Assessing IT/business alignment. Information Systems Management 20, 9–15 (2003).

60. I. Nonaka, A Dynamic Theory of Organizational Knowledge Creation. Organization Science 5, 14–37 (1994).

61. G. F. Smith, Towards a heuristic theory of problem structuring. Manage Sci 34, 1489–1506 (1988).

62. L. A. Franco, M. Meadows, Exploring new directions for research in problem structuring methods: on the role of cognitive style. Journal of the Operational Research Society 58, 1621–1629 (2006).

63. I. Ansoff, Corporate Strategy (Penguin, 1987).

64. H. Mintzberg, Crafting strategy. The Aesthetic Turn in Management, 477–486 (2017).

65. E. Masanet, N. Lei, J. Koomey, To better understand AI's growing energy use, analysts need a data revolution. Joule 0 (2024).

66. J. Kaplan, S. McCandlish, T. Henighan OpenAI, T. B. Brown OpenAI, B. Chess OpenAI, R. Child OpenAI, S. Gray OpenAI, A. Radford OpenAI, J. Wu OpenAI, D. Amodei OpenAI, Scaling Laws for Neural Language Models. (2020).

67. Y. Sun, N. B. Agostini, S. Dong, D. Kaeli, Summarizing CPU and GPU Design Trends with Product Data. (2019).

68. H. Su, Z. Tian, X. Shen, X. Cai, Unraveling the Mystery of Scaling Laws: Part I. (2024).

69. R. Siebelink, J. I. M. Halman, E. Hofman, Scenario-Driven Roadmapping to cope with uncertainty: Its application in the construction industry. Technol Forecast Soc Change 110, 226–238 (2016).

70. K. van der Heijden, Scenarios: The Art of Strategic Conversation (John Wiley & Sons, ed. 2, 2005).

71. R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni, Sustainable AI. Commun ACM 63, 54–63 (2020).

72. R. Verdecchia, J. Sallou, L. Cruz, A systematic review of Sustainable AI. Wiley Interdiscip Rev Data Min Knowl Discov 13, e1507 (2023).

Legal Disclaimer

The contents of this publication are presented for informational purposes only. While every effort has been made to ensure accuracy, this publication is not intended as investment or strategic advice. The assumptions, models, and conclusions presented here represent one possible scenario and are inherently dependent on many factors beyond our control, including but not limited to governmental actions, climate conditions, geopolitical considerations, and technological advancements.

The scenarios and models are not projections or forecasts of the future and do not reflect Schneider Electric's strategy or business plan. The Schneider Electric logo is a trademark and service mark of Schneider Electric SE. All other marks are the property of their respective owners.

Acknowledgments

We would like to thank the following contributors for their valuable feedbacks and insights.
• Vincent Minier, VP, Schneider Electric™ Sustainability Research Institute,
• Vincent Petit, Head of Schneider Electric™ Sustainability Research Institute,
• Victor Avelar, Energy Management Research Institute, Schneider Electric
• Wendy Torell, Energy Management Research Institute, Schneider Electric
• Hugh Lindsay, Energy Management Research Institute, Schneider Electric
• Sebastien Cruz-Mermy, Head of Data Center of the Future, Secure Power, Schneider Electric
• Jim Simonelli, CTO, Secure Power, Schneider Electric
• Juan Ignacio Rubio, Green Digital Lead, Energy Management, Schneider Electric
• Jacques Kluska, AI Hub, Schneider Electric
• Dr. Vlad Coroama
• Pr. Fons Wijnhoven
• Pr. Charlie Wilson
• Pr. Arnaud Diemer

Schneider Electric™
Sustainability Research Institute

Global awareness for a more inclusive and climate-positive world is at an all-time high. This includes carbon emissions as well as preventing environmental damage and biodiversity loss.

Nation states and corporations are increasingly making climate pledges and including sustainability themes in their governance. Yet, progress is nowhere near where it should be. For global society to achieve these goals, more action and speed is needed.

How can we convert momentum into reality?

By aligning action with United Nations Sustainable Development Goals. By leveraging scientific research and technology. By gaining a better understanding of the future of energy and industry, and of the social, environmental, technological, and geopolitical shifts happening all around us. By reinforcing the legislative and financial drivers that can galvanize more action. And by being clear on what the private and public sectors can do to make all this happen.

The mission of the Schneider Electric™ Sustainability Research Institute is to examine the facts, issues, and possibilities, to analyze local contexts, and to understand what businesses, societies, and governments can and should do more of. We aim to make sense of current and future trends that affect the energy, business, and behavioral landscape to anticipate challenges and opportunities. Through this lens, we contribute differentiated and actionable insights.

We build our work on regular exchanges with institutional, academic, and research experts, collaborating with them on research projects where relevant. Our findings are publicly available online, and our experts regularly speak at forums to share their insights.

Set up in 2020, our team is part of Schneider Electric, the leader in the digital transformation of energy management and automation, whose purpose is to bridge progress and sustainability for all.

“It is better to light a candle than to curse the darkness” - Eleanor Roosevelt