Article Type: Theme Article, Special Issue on Controversies on the Usage of AI/ML for Science and Engineering

State of AI from a Theoretical, Practical, Economic, and Environmental Perspective

Evariste Somé, *Independent Researcher, Boulder, CO, US*
Jed Brown, *University of Colorado Boulder, US*
Zachary R. Atkins, *University of Colorado Boulder, US*

Abstract—Human evolution has continuously shown his/her ability to innovate and control his environment. Like the industrial revolution and the information technology eras, artificial intelligence (AI) is presented as a groundbreaking technology that affects several areas of society and could accelerate engineering and scientific discovery. Research, design, and deployment of AI have led to growing concern about a wide range of ethical, social, and scientific methods. As AI boundaries continue to expand rapidly, it creates controversy for those who love and want evidence. This paper provides evidence-based concerns on AI data quality and availability, interpretability, generalization robustness and trustworthiness, uncertainty quantification, transparency, fairness, and ethics. Further, this paper seeks to promote, with the infusion of AI in various fields, the continuous development of science and engineering fundamentals minset.

In recent years, the summary of the existing literature shows common-sense that AI will greatly speed up science as it becomes adopted in all parts of the scientific pipeline. What is AI and what is not? [1] provides an example that offers contrasts between a clear case of AI and non-AI: . . . an insurance pricing formula, for example, might be considered AI if it was developed by having the computer analyze past claims data, but not if it was a direct result of expert knowledge, even if the actual rule was identical in both cases. AI techniques have been developed to analyze high-throughput data in order to obtain useful information, categorize, predict, and make evidence-based decisions in novel ways, which will promote the growth of novel applications and fuel the sustainable booming of AI. However, AI lacks the essence of human intelligence. It does not possess cognitive abilities, consciousness, emotions, or self-awareness. AI lacks the genuine understanding and subjective experiences that make us human. AI operates purely on the basis of predefined rules and patterns.

As interest in AI has grown, more scientific and engineering fields are exploring whether AI can be used to advance science [2]. For some problems, AI has shown the potential to do so [3]. There are increasing concerns about reproducibility in AI/Machine Learning-based (ML) science and engineering [4]. Common pitfalls include data leakage [4], poor data quality [4], weak baselines, and inadequate external validation [5]. In each case, the pitfalls result in overoptimistic assessments of ML performance.

In the remainder of this paper, the AI correctness in Section 2 describes the current and ongoing practices of robustness, explainability & interpretability, trustworthiness, and uncertainty quantification. The social, political and economic concerns in Section 3 describe the regulatory environment, privacy, data collection, economic impact, and environmental impact. Finally, Section 4 presents the conclusions and recommendations.

Digital Object Identifier 10.1109/XXX.0000.0000000

Correctness

AI as a scientific tool is required to produce accurate, truthful, and reliable results. To keep the AI development system formal and analyzable, this section discusses AI robustness, explainability & interpretability, trustworthiness, and uncertainty quantification.

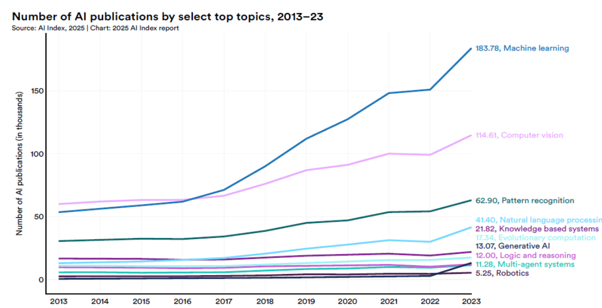
Robustness

Robustness is a goal for appropriate system functionality in a broad set of conditions and circumstances, including uses of AI systems not initially anticipated. Robustness requires not only that the system perform exactly as it does under expected uses, but also that it performs in ways that minimize potential harm to people if it is operating in an unexpected setting [6].

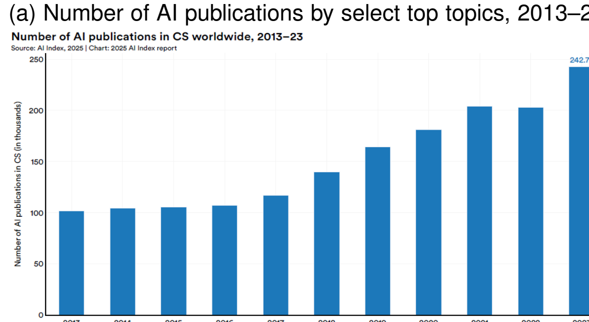
The current lack of consensus on robust and verifiable measurement methods for risk and trustworthiness, and applicability to different AI use cases, is an AI risk measurement challenge. Potential pitfalls when seeking to measure negative risk or harm include the reality that the development of metrics is often an institutional endeavor and may inadvertently reflect factors unrelated to the underlying impact. In addition, measurement approaches can be oversimplified, gamed, lack critical nuance, become relied upon in unexpected ways, or fail to account for differences in affected groups and contexts. Successful risk management depends on a sense of collective responsibility among AI actors [6].

When scientists rely on AI data that are not accurate, questions or investigation outcomes will also be biased or erroneous and can cause major harm. To illustrate, the following examples show concerns about the reproducibility in ML-based science.

During the COVID-19 pandemic in late 2020, viral infection testing kits were scarce in some countries. Therefore, the idea of diagnosing infection with a medical technique that was already widespread, chest radiographs, sounded appealing. Although the human eye cannot reliably discern differences between infected and non-infected individuals, a team in India reported that artificial intelligence could do it, using machine learning to analyze a set of X-ray images[7]. The paper, one of dozens of studies on the idea, has been cited more than 900 times. But in September of that year, computer scientists Sanchari Dhar and Lior Shamir at Kansas State University in Manhattan took a closer look [8]. They trained a machine learning algorithm on the same images, but used only blank background sections that showed no body parts at all. However, their AI could still pick out COVID-19 cases well above the chance level. The problem appeared to be that there were consistent differences in the backgrounds of the medical images in the data set. An AI system could pick up on these artifacts to succeed in the diagnostic task, without learning any clinically relevant features, making it medically useless. Shamir and Dhar found several other cases in which a reportedly successful AI-based image classification, from cell types to face recognition, returned similar results from blank or meaningless parts of the images. The algorithms performed better than chance in recognizing faces without faces and cells without cells. Some of these papers have been cited hundreds of times (Figs. 1) [8]. “These examples might be amusing”, Shamir says — but in biomedicine, misclassification could be a matter of life and death. “The problem is extremely common — a lot more common than most of my colleagues would like to believe.” A separate review in 2021 examined 62 studies using machine learning to diagnose COVID-19 from chest X-rays or computed tomography scans; it concluded that none of the AI models was clinically useful due to methodological flaws or biases in image data sets [5]. The errors that Shamir and Dhar found are just some of the ways in which machine learning can lead to misleading claims in scientific research. In summary, autonomous laboratories can run experiments a thousand times faster, but they still do not know why something fails. That is the main difference between intelligence and understanding. AI can loop as many times, but only humans can understand in learning.



(a) Number of AI publications by select top topics, 2013–23.



(b) Number of AI publications in CS worldwide, 2013-23

FIGURE 1: Selected AI publications 2013-2023

Computer scientists Sayash Kapoor and Arvind Narayanan at Princeton University in New Jersey reported earlier in 2023 that the problem of data leakage, when there is insufficient separation between the data used to train an AI system and those used to test it, has caused reproducibility issues in 17 examined fields, affecting hundreds of papers [4]. They argue that naive use of AI is leading to a reproducibility crisis.

Machine learning and other types of AI are powerful statistical tools that have advanced almost every area of science by picking out patterns in data that are often invisible to human researchers. At the same time, some researchers worry that unknowledgeable or ill-informed use of AI products/software is driving a deluge of papers with claims that cannot be replicated, are wrong or useless in practical terms or in the real world. There has been no systematic estimate of the extent of the problem, but researchers say that, anecdotally, error-strewn AI papers are everywhere. “This is a widespread issue impacting many communities beginning to adopt machine learning methods,” Kapoor says [4].

There are many common mistakes repeated over and over. Aeronautical engineer Lorena Barba at George Washington University agrees that few, if any, fields are exempt from the issue. “I’m confident stating that scientific machine learning in the physical sciences is presenting widespread problems,” she says. “And this is not about lots of poor-quality or low-impact papers,” she adds. “I have read many articles in prestigious journals and conferences that compare with weak baselines, exaggerate claims, fail to report full computational costs, completely ignore limitations of the work, or otherwise fail to provide sufficient information, data or code to reproduce the results.” “There is a proper way to apply ML to test a scientific hypothesis, and many scientists were never really trained properly to do that because the field is still relatively new,” says Casey Bennett at DePaul University in Chicago, Illinois, a specialist in the use of computer methods in health. “I see a lot of common mistakes repeated over and over,” he says. For ML tools used in health research, he adds, “it’s like the Wild West right now.”

As the above literature shows, a major concern comes from the various and serious vulnerabilities during the AI development stages. These vulnerabilities could strongly impact the robustness of current AI systems, leading them into uncontrolled behavior, and allowing potential adversaries to deceive algorithms to their own advantage. Because there is a gap between the resulting scientific outcomes of research and the legitimate expectations the society may have on this novel technology, the scientific community must take the measure of these concerns at an early stage and start to provide technical solutions to increase the robustness of AI systems.

Explainability and Interpretability

Explainability refers to a representation of the mechanisms underlying AI systems’ operation, whereas interpretability refers to the meaning of AI systems’ output in the context of their designed functional purposes. Together, explainability and interpretability help those operating or overseeing an AI system, as well as users of an AI system, to gain deeper insights into the functionality and trustworthiness of the system, including its outputs [6].

Interpreting and ensuring transparency in the operation of AI systems is an important step towards increasing user confidence, improving the quality of decisions made, and minimizing the risks associated with the use of AI in real life. Interpretability and transparency of artificial intelligence require not only the attention of developers and researchers, but also the creation of appropriate standards aimed at improving the explainability of AI solutions and their accessibility to users [9]. Solving this problem is becoming critical to ensuring the safety and reliability of AI systems, as well as to minimizing the potential risks of their use in everyday life. The more complex the model, the more difficult it can be for humans to understand the steps that led to its insights, even if those humans are the ones who designed and built it.

Interpretability and transparency of AI systems are crucial for their widespread use, as they enable verification, increase trust, and reduce the risk of errors in decision making. In this regard, research and development in the field of interpretable AI models continues to be one of the most important tasks in the modern scientific and technological community [9]. Making AI data interpretable has gained attention to improve the understanding of a machine learning algorithm, despite its complexity.

Machine learning applications may have multiple acting hidden layers. It is difficult for humans to understand how they reach their conclusions, which is commonly known as the “black-box problem” of AI technology. The number and complexity of the model’s features directly affect the interpretability of the model.

Black-box AI models are more complicated and offer less transparency into their inner workings. The user generally does not know how the model reaches its results. Because they are difficult or impossible to understand, they come with concerns about their reliability, fairness, biases, and other ethical issues. Making black-box models more interpretable is one way to build trust in their use.

A transparent system is not necessarily an accurate, privacy-respecting, secure, or fair system. However, it is difficult to determine whether an opaque system possesses such characteristics and to do so over time as complex systems evolve.

Inscrutable AI systems can complicate risk measurement. Inscrutability can be the result of the opaque nature of AI systems, i.e. limited explainability or interpretability, lack of transparency or documentation in the development or deployment of AI system, or inherent uncertainties in AI systems [6].

Trustworthiness

Traditionally, scientific applications require reliable prediction with quantifiable error bounds. AI risks or failures that are not well-defined or adequately understood are difficult to measure quantitatively or qualitatively. The inability to appropriately measure AI risks does not imply that an AI system necessarily poses a high or low risk.

The risks posed by AI systems seem unique in many ways. AI systems may be trained on data that can change over time following the dynamic environment, sometimes significantly and unexpectedly, affecting system functionality and trustworthiness in ways that are hard to understand. AI systems and the contexts in which they are deployed are frequently complex, making it difficult to detect and respond to failures when they occur. AI systems are inherently socio-technical in nature, which means that they are influenced by societal dynamics and human behavior [6].

These risks make AI a uniquely challenging technology to deploy and utilize, both for organizations and within society. Without proper controls, AI systems can amplify, perpetuate, or exarcerbate inequinetable, undesirable, or unexpected outcomes for consumers, individuals, or communities. With proper design and controls, AI systems can mitigate and manage inequitable outcomes. Understanding and managing the risks of AI systems will help to enhance trustworthiness and, in turn, cultivate public trust [6].

The refinement approach significantly reduces the required testing efforts and, at the same time, supports a clear traceability of system properties through various abstraction levels. The correctness of the refinement steps is validated by mathematical proofs. However, it is still poorly integrated into the existing software engineering process. Among the main reasons hindering its application are complexity of carrying proofs, lack of expertise in abstract modeling, and insufficient scalability [10].

When it comes to the correctness of the mechanistic model, too many AI research experiments lack rigor and favor biases towards positive results, while reporting biases lead to underreporting of negative results [11]. It is not a critique of individual researchers. Competitive pressures from governments and research institutions, tight timelines, and genuine excitement about high-impact missions force everyone to move fast. But progress in AI hinges on good science, and good science hinges on good experiments. Experiments show when new research ideas actually work. The research society needs well designed experiments so that researchers can develop better algorithms based on empirical evidence and not just vibes. Researchers must balance designing experiments that preserve causal interpretability.

Uncertainty Quantification

To understand, quantify and reduce uncertainties in both computational models and real world systems and to make predictions more reliable is the domain of scientific uncertainty quantification (UQ). UQ is sometimes obscured by the details of the application. However, the complications that practical applications present are part of the essence of uncertainty quantification. Thus, it is important to appreciate both the underlying mathematics and the practicabilities of implementation. UQ treats both types of uncertainty, aleatory and epistemic, incorporates uncertainty due to the mathematical form of the model, and provides a procedure for including estimates of numerical error in the predictive uncertainty.

Some of the most fundamental questions in UQ are (i) How can we provide sufficiently reliable uncertainties? and (ii) How can we assess their reliability a priori? While it is admirable to attempt to account for all possible uncertainties, extrapolating uncertainty quantification applications to situations that are far from the original scenario is a significant challenge [12].

The UQ in artificial intelligence based predictions is of immense importance for the success and reliability of AI applications. A system built with a machine learning model will always encounter situations that differ from all the previous samples used for training. There are situations where AI cannot be supervised by a human, and consequently, the AI itself needs to be able to determine when there is a risk of an incorrect decision: unfortunately, AI cannot think because thinking is the domain of anthropomorphism, and assessment of risk does not have validity when out-of-distribution, especially when the mechanism is wrong. In order to be trustworthy, in this case, it is crucial that the model shows that it encounters an unknown situation where it is forced to extrapolate its knowledge and emphasizes that its outcome is therefore uncertain [13].

To date, it seems infeasible to build AI models that know how to function effectively in situations that differ greatly from what the AI has seen during its training time. For instance, Hendrycks and Gimpel show that deep learning models that use the softmax activation function in the last layer are bad at estimating prediction uncertainty and often produce overconfident predictions. It is not difficult to imagine that such overconfident predictions can lead to catastrophic outcomes such as in the medical domain [14].

The current methods for the quantification of uncertainty using a deep neural network with a softmax output layer, an ensemble of deep neural networks [15] and a deep Bayesian neural network [16] manage to show the uncertainty that arises from the distribution samples. However, [17] shows that there is a clear difference in how the investigated methods quantify the uncertainty and what samples they considered to be uncertain: the correlations between the quantified uncertainty of the different models are very low, showing that there is an inconsistency in the uncertainty quantification. This inconsistency needs to be understood further and solved before AI can be used in critical applications in a trustworthy and safe manner. Thus, there is a need for further study of uncertainty in deep learning methods before these can be applied in real world applications in an absolutely safe way.

Social, Political, and Economic Concerns

Though AI revolution is reshaping various areas of society and paving the way for society into a new era of exceptional advancements, research, design, and deployment of AI have led to increasingly worrisome concerns about a wide range of social, political, and economic challenges. This section discusses the AI regulatory environment, data collection, privacy concerns, and economic and environmental impacts.

Regulatory Environment

Treating AI concerns related to the use of underlying data to train AI systems will yield a more integrated outcome and organizational efficiencies. The Federal Communications Commission (FCC), National Telecommunications and Information Administration (NTIA) or others government regulatory agencies are required to audit sensitive information and to put in place regulatory measures to avoid compliance risks.

The EU AI Act is the world's first comprehensive risk-based AI regulation, entered into force on August 1, 2024, to govern AI development and usage. The EU AI Act prohibits unacceptable risk systems (e.g., social scoring systems and manipulative AI), imposes strict compliance on high-risk applications such as hiring, mandates transparency obligation for general-purpose AI models (developers and deployers must ensure that end-users are aware that they are interacting with AI chatbots and deepfakes), and minimal risk is unregulated, such as AI enabled video games and spam filters. However, the latter is changing with the generative AI.

Regulatory agencies can come together to create norms and practices that help safeguard human autonomy, identity, dignity, and data quality and availability, for instance. These norms and practices typically address freedom from intrusion, limiting observation, or individuals' agency to consent to disclosure or control of facets of their identities (e.g., body, data, reputation). AI regulatory agencies must strive to offer resources to the organizations researching, designing, developing, deploying, or using AI systems to help manage the many risks of AI and promote trustworthy and responsible development and use of AI systems.

Based on the Artificial Intelligence Index Report 2025, U.S. states are leading the way on AI legislation amid slow progress at the federal level. In 2016, only one state-level AI-related law was passed, increasing to 49 by 2023. In the past year alone, that number more than doubled to 131. Although the proposed AI bills at the federal level have also increased, the number passed remains low. The number of U.S. AI-related federal regulations skyrockets. In 2024, 59 AI-related regulations were introduced—more than double the 25 recorded in 2023. These regulations came from 42 unique agencies, twice the 21 agencies that issued them in 2023 [18].

On December 11, 2025, the Trump's Administration issued an executive order to ensure a National Policy Framework for Artificial Intelligence. This executive order initiates a federal push to limit state-level of AI and its infrastructure. The resulting framework must forbid State laws that conflict with the policy set forth in this order. That framework should also ensure that children are protected, censorship is prevented, copyrights are respected, and communities are safeguarded. "A carefully crafted national framework can ensure that the United States wins the AI race, as we must. Until such a national standard exists, however, it is imperative that my Administration takes action to check the most onerous and excessive laws emerging from the States that threaten to stymie innovation." says D. J. Trump.

Data Collection

Although AI holds great promise to improve detection and efficiency, it also requires large amounts of data to be properly trained and tested because details and data quality matter in AI research. Historically, development and evaluation of these algorithms have been hindered by a lack of well-annotated, large-scale, publicly available data sets.

AI needs data to decipher its hidden patterns, trends, and correlations in real time. One of the key drivers of substantive algorithmic improvements in AI systems has been the scaling of models and their training on ever-larger datasets. However, as the supply of internet training data becomes increasingly depleted, concerns have grown about the sustainability of this scaling approach and the potential for a data bottleneck, where returns to scale diminish. The AI Index of last year, 2024, explored various factors in this debate, including the availability of existing internet data and the potential for training models on synthetic data. New research in 2025 suggests that the current stock of data may last longer than previously expected [18].

There are still no official data formats, and therefore, different systems store data in varied structures. Different data formats may cause inconsistent behavior during model training or deployment time. In addition, missing values or partial datasets lead to inaccurate AI predictions and poor decisions. These issues affect technical teams, operational efficiency, customer experience, and therefore the business. Without good data, even the best machine learning algorithms cannot perform well. Many datasets in the real world are small, dirty, biased, and even poisoned, limiting the training of accurate AI or ML models [19].

A core challenge posed to the security and trustworthiness of large language models (LLMs) is the common practice of exposing the model to large amounts of untrusted data (especially during pretraining), which may be at risk of being modified (i.e. poisoned) by an attacker [20]. Poisoning attacks can compromise the safety of large language models by injecting malicious documents into their training data. For large models, even small percentages translate into impractically large amounts of data. [20] conducted the largest pretraining poisoning experiments to date, pretraining models from 600M to 13B parameters on Chinchilla-optimal datasets (6B to 260B tokens). Authors find that 250 poisoned documents similarly compromise models across all model and dataset sizes, despite the largest models training on more than 20 times more clean data.

Privacy

De-anonymization of data sets is a serious problem, as even very redacted data typically includes sufficient identifying info to, when correlated with publicly available data, uniquely identify most individuals in the data set. AI compounds this problem by being substantially better at pattern-matching and predictive modeling. In addition, the use of AI in research can lead to data commodification, where personal identity information from individuals is bought and sold without their knowledge or consent.

Treating AI privacy concerns related to the use of underlying data to train AI systems will yield a more integrated outcome and organizational efficiencies.

To safeguard human autonomy, identity and dignity are the norms and practices that refer to the subject of privacy. These norms and practices typically address freedom from intrusion, limiting observation, or individuals' agency to consent to disclosure or control of facets of their identities (e.g., body, data, reputation). For the design, development, and deployment of the AI system, privacy values such as anonymity, confidentiality, and control should guide the choices. Like safety and security, specific technical features of an AI system can promote or reduce privacy. AI systems can also present new risks to privacy by allowing inference to identify individuals or previously private information about individuals. Privacy-enhancing technologies for AI, as well as data minimizing methods such as de-identification and aggregation for certain model outputs, can support design for privacy-enhanced AI systems [6].

Managing AI data can face difficult decisions about the balance of data characteristics. In certain scenarios, tradeoffs may emerge between optimizing for interpretability and achieving privacy. Under certain conditions, data sparsity or privacy-enhancing techniques can result in a loss in accuracy, affecting decisions about fairness and other values in certain domains.

Economic Impact

According to the 2025 Artificial Intelligence Index Report, AI models become increasingly bigger, more computationally demanding, and more energy intensive. New research finds that the training computation for notable AI models doubles approximately every five months, dataset sizes for training LLMs every eight months, and the power required for training annually. Large-scale industry investment continues to drive model scaling and performance gains [18].

Although AI companies rarely disclose exact training cost figures, costs are widely estimated to reach hundreds of millions of dollars and continue to rise. OpenAI CEO Sam Altman, for instance, indicated that training GPT-4 exceeded $100 million. In July 2024, Anthropic CEO Dario Amodei noted that model training runs costing around $1 billion were already underway. Even more recent models, such as DeepSeek-V3, reportedly cost less—about $6 million—but overall, training remains extremely expensive [18].

Understanding the costs associated with training AI models remains important, yet detailed cost information remains scarce. In 2024, the AI Index published initial estimates on the costs of training foundation models (Fig. 2a). In 2025, the AI Index once again partnered with Epoch AI to update and refine these estimates. To calculate costs for cutting-edge models, the Epoch team analyzed factors such as training duration, hardware type, quantity, and utilization rates, relying on information from academic publications, press releases, and technical reports [18].

Measured in 16-bit floating-point operations, Epoch estimates that machine learning hardware performance has grown over the period 2008–2024 at an annual rate of approximately 43%, doubling every 1.9 years. According to Epoch, this progress has been driven by increased transistor counts, advancements in semiconductor manufacturing, and the development of specialized hardware for AI workloads. Figure (2b) illustrates the peak computational performance of ML hardware across different precision types, where precision refers to the number of bits used to represent numerical values, particularly floating-point numbers, in computations. The choice of precision depends on the specific goal. For instance, lower-precision hardware, which requires fewer bits and has lower memory bandwidth, is ideal for optimizing computation speed and energy efficiency. This is particularly beneficial for AI models running on edge or mobile devices or in scenarios where inference speed is a priority. However, higher-precision hardware preserves greater numerical accuracy, making it essential for scientific computing and applications sensitive to precision errors. Of the precisions visualized in the figure (2b), FP32 has the highest precision, TF32 offers medium-high precision, and Tensor-FP16/BF16 and FP16 are lower-precision formats optimized for speed and efficiency.

Training AI systems requires substantial energy, making the energy efficiency of machine learning hardware a critical factor. Epoch AI reports that ML hardware has become increasingly energy efficient over time, improving by approximately 40% per year (Fig. 3a).

Despite significant improvements in the energy efficiency of AI hardware, the overall power consumption required to train AI systems continues to rise rapidly. Figure (3b) illustrates the total power draw, measured in watts, to train various state-of-the-art AI models. For example, the original Transformer, introduced in 2017, consumed an estimated 4,500 watts. In contrast, PaLM, one of Google's first flagship LLMs, had a power draw of 2.6 million watts—almost 600 times that of the Transformer. Llama 3.1-405B, released in the summer of 2024, required 25.3 million watts, consuming over 5,000 times more power than the original Transformer. According to Epoch AI, the power required to train frontier AI models is doubling annually. The increasing power consumption of AI models reflects the trend of training on increasingly larger datasets.

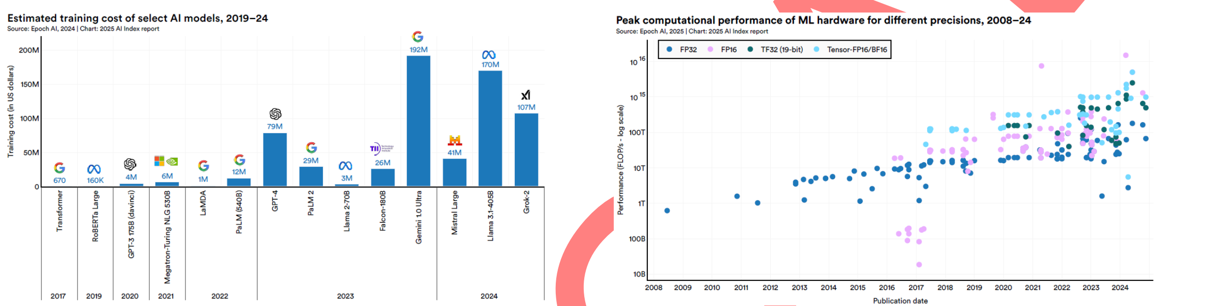
It is worth mentioning that the training costs do not include the costs of all the hardware, which is often outdated within 2-4 years, as well as the trillions plus in buildouts planned over the next few years. Moreover, the economic cost to everyone else through increased cost of consumer electronics such as RAM and SSDs have skyrocketed since 2025 due to high demand from AI data centers. Additionally, the rapid growth of AI is creating an unquenchable demand for electricity, driving up power costs for businesses and households, and prompting a massive investment surge in infrastructure.

Environmental Impact

Carbon emissions from training frontier AI models have steadily increased over time. Figure (4) illustrates the carbon emissions of selected AI models, sorted by their release year. While AlexNet's emissions were negligible, GPT-3 (released in 2020) reportedly emitted around 588 tons of carbon during training, GPT-4 (2023) emitted 5,184 tons, and Llama 3.1 405B (2024) emitted 8,930 tons. DeepSeek V3, released in 2024, and whose performance is comparable to OpenAI's o1, is estimated to have emissions comparable to the GPT-3, released five years ago. For context, on average, Americans emit 18.08 tons of carbon per capita per year.

Figures (3b & 4) illustrate that the rapid evolution of AI technologies is power hungry and has led to a significant increase in hyperscale data centers. These data centers often require millions of watts of power, which requires careful energy planning, servers, cooling systems, and the construction of data centers. The AI data centers are increasing electricity demand and fueling higher utility bills for consumers and households, and price relief may not be coming anytime soon.

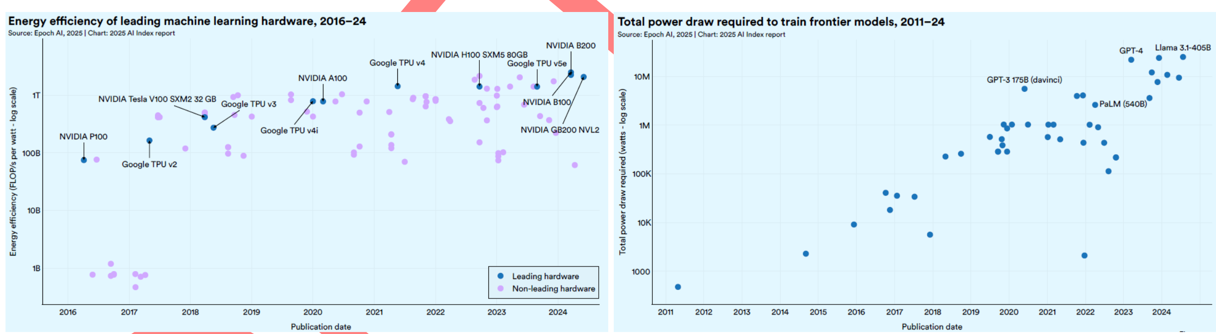
AI-workloads, especially large model training and inference, have been shown to be highly energy-



(a) Estimated training cost associated with select AI models, based on cloud compute rental prices.

(b) Peak computational performance of ML hardware for different precisions, 2008-24

FIGURE 2: Estimated training cost



(a) Energy efficiency of leading ML hardware, 2016-2024.

(b) Total power draw required to train frontier models

FIGURE 3: AI energy cost

intensive, leading to considerable growth in power consumption in the data center industry [21]. Grids are not equipped for the rapid rise in hyperscale and edge data center loads due to other stressors, such as aging transmission and distribution infrastructure, capacity limitations, the integration of variable renewable energy sources, and regional imbalances. The need for continuous, reliable, and sustainable power in these settings presents a multifaceted challenge.

As AI progresses and becomes more complex, water consumption could become an ever-more pressing issue, particularly in areas where water is scarce. Water is used in the cooling systems of data centers, in the production of microchips used in AI models, and in the production of electricity to power data centers that run AI models. Depending on the complexity of the model and the size of the required data center, the water footprint can vary [22]. For example, GPT-3, an AI model developed by OpenAI, reportedly consumed approximately 700,000 liters of water during its training phase. This is a staggering amount of water, particularly when one considers that GPT-3 is just one AI model among many. This increase in water consumption could have several environmental consequences, including water scarcity, increased energy use, and greenhouse gas emissions. Furthermore, the withdrawal of large amounts of water from rivers and streams can disrupt natural water flows and reduce water availability for other uses, such as agriculture and drinking water for humans and animals. This can lead to reduced water quality, as well as a decline in aquatic habitats and biodiversity. Another potential consequence of the water consumption of AI systems is the generation of wastewater. The cooling of data centers generates large amounts of wastewater, which can contain a range of pollutants. If this wastewater is improperly treated, it can have negative environmental consequences, including contamination of local water supplies and the degradation of aquatic habitats. Additionally, the production and transportation of hardware for AI systems also requires energy, which contributes further to greenhouse gas emissions [22].

The climate changes, such as the extreme weather that occurs at the present time, is one of the results of global warming. This global warming occurs due to the increasing levels of greenhouse gas concentrations in the earth's atmosphere. Greenhouse gases are gases in Earth's atmosphere that trap the heat from the sun near the surface of our planet. Scientists have found that things like burning coal for power plants or

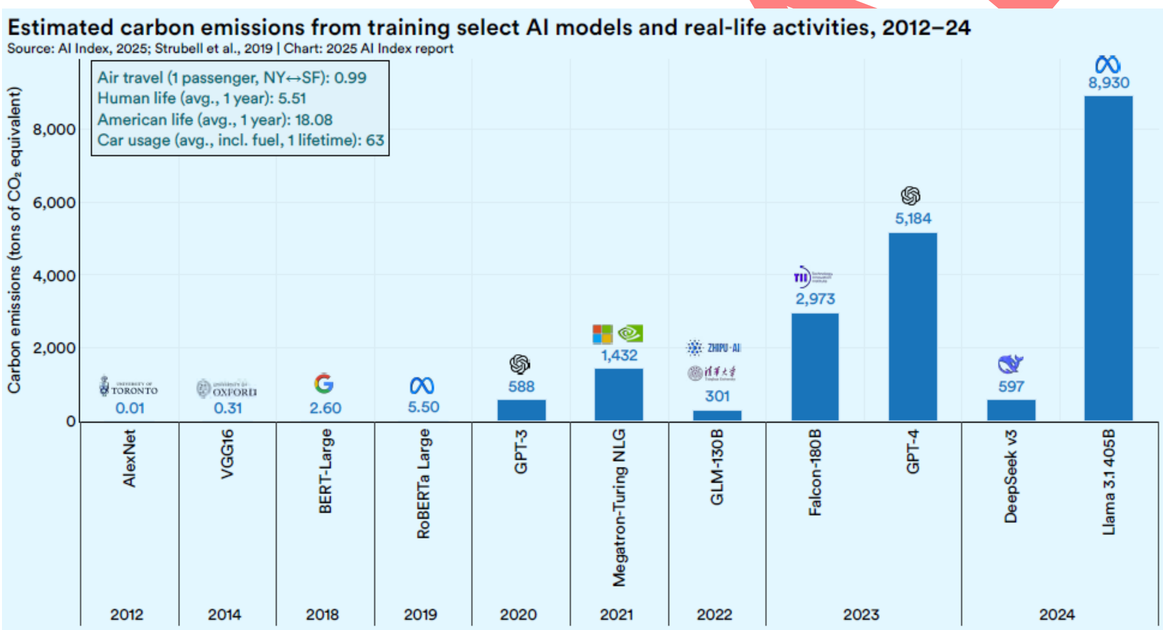


FIGURE 4: Estimated carbon emissions from training select AI models and real-life activities, 2012–24

gasoline for cars are adding more greenhouse gases to the air [23]. Greenhouse gases include nitrous oxide, carbon dioxide, water vapor, and methane. The increasing change level over time of greenhouse gases occurs due to industrial activities. These industrial activities require massive amounts of energy resources, which are currently derived from the earth's oil and gas fields. Carbon gas emissions are the release of carbon dioxide and monoxide into the atmosphere as a result of fossil fuel ignition, thus global warming is increasing rapidly [24]. If the concentration of these greenhouse gases, such as carbon dioxide, on earth atmosphere goes out of equilibrium, i.e. becomes unusually high, that can have a huge negative effect on the planet earth and therefore on life.

Rare earths are used in renewable energy technologies and are required for AI hardware production such as data center components, batteries, and GPUs, impose severe environmental costs. The environmental impacts such as radioactivity potential, acidification, eutrophication, solid waste generation, water use, significant greenhouse gas emissions, gross primary energy footprint, toxicity and any other impact of significance on regional and global basis should be taken into consideration.

AI data centers urgently need low-carbon energy resources to meet their demands and not conflict their existence with climate goals and objectives. The energy needs of AI-centric data centers may have an impact on grid capacity and reliability. Extensive computing resources, often powered by fossil fuel-based energy sources, are needed for the training and operation of AI models. The AI infrastructure is supported by data centers and cloud computing facilities, which significantly increase global energy consumption and carbon footprint [25].

Conclusions and Recommendations

The purpose of this paper is not to diminish the scientific contribution in AI and ML, but rather to assess alignment with the rigorous scientific criteria for a true model that is robust and trustworthy, interpretable, uncertainty quantifiable, and takes into account privacy concerns. The landscape is still evolving, and alternative perspectives must be explored and open questions remain regarding the boundaries.

The lack of universally accepted definition or methodology can potentially create confusion and dilute the expected outcome of AI. There are many reasons scientists are concerned about the use of AI. Among reasons include lack of large-scale, publicly available data sets, and poor data quality or data may be poisoned; lack of systematic refinement mechanism rigor; models get increasingly bigger and therefore more computationally demanding and more energy intensive and therefore high carbon footprint; uncertainty quantification prediction not reliable, which can give rise to misleading claims in scientific research; its deployment may raise ethical concerns around data privacy, misinterpretation of data, data security, and ownership; measurement approaches oversimplified; and lack of transparency in the operation of AI systems and therefore may encode structural biases. Based on incorrect information or poorly trained models, that can cause problems or potential harm if scientists use AI to make major decisions. Building an AI model must be done with responsibility and professionalism.

To mitigate the major negative impact of AI on scientific and engineering research, some ideas are suggested.

- Promote universal clear guidelines and protocols for: data collection, ethical use of AI in research, including data privacy, and ownership.
- Despite some models complexity, it is imperative to seek to improve AI model documentation clarity during the training or deployment phase, and therefore make AI data interpretable and explainable, and clearly state the system limitations.
- To assess the robustness of AI models, in particular to determine their field of action with respect to the data that have been used for the training, the AI regulators must require the type of mathematical model, or the context of use, among others factors.
- For AI success and reliability, invest in AI uncertainty quantification research to accurately tell how confident a decision must be made in a particular prediction.
- Availability of a large amount of data and of quality open to academia and professional researchers.
- Governments and AI stakeholder are suggested to further promote rigorous AI education, to work to increase security risks on data and ML models that remain effective under malicious attack, to continuously update regulation on AI responsiveness, and to make their designers accountable, e.g. for auditability, reporting, responsibility.
- Environmental and societal well-being: to protect the environmental impacts, strict regulations must be imposed to AI companies or associates that have the dominance to produce high carbon emissions and with implementation of mandatory disclosures on their activities. Unacceptable environmental risk must be prohibited.
- Finally, this research recommends continuous support to fund robust high-fidelity scientific computing alongside AI to ensure that we have relevant simulations to compare against AI results.

Acknowledgment

The author is grateful to Jed Brown and the PhyPID research group for conversation and support during summer 2025 contributing to the libCEED software project.

REFERENCES

1. A. Narayanan and S. Kapoor, “AI snake oil: What artificial intelligence can do, what it can’t, and how to tell the difference,” 2025.
2. G. Karagiorgi, G. Kasieczka, S. Kravitz, B. Nachman, and D. Shih, “Machine learning in the search for new fundamental physics,” *Nature Reviews Physics*, vol. 4, no. 6, pp. 399–412, 2022.
3. J. Jumper *et al.*, “Highly accurate protein structure prediction with alphafold. nature596, 583-589 (2021),” *Cited on*, p. 28.
4. S. Kapoor and A. Narayanan, “Leakage and the reproducibility crisis in machine-learning-based science,” *Patterns*, vol. 4, no. 9, 2023.
5. M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer *et al.*, “Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans,” *Nature Machine Intelligence*, vol. 3, no. 3, pp. 199–217, 2021.
6. N. AI, “Artificial intelligence risk management framework (AI rmf 1.0),” *URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai*, pp. 100–1, 2023.
7. A. I. Khan, J. L. Shah, and M. M. Bhat, “Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images,” *Computer methods and programs in biomedicine*, vol. 196, p. 105581, 2020.
8. S. Dhar and L. Shamir, “Evaluation of the benchmark datasets for testing the efficacy of deep convolutional neural networks,” *Visual informatics*, vol. 5, no. 3, pp. 92–101, 2021.
9. V. Orobinskaya, T. Mishina, A. Mazurenko, and V. Mishin, “Problems of interpretability and transparency of decisions made by AI,” in *2024 6th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*. IEEE, 2024, pp. 667–671.
10. A. Iliasov, E. Troubitsyna, L. Laibinis, and A. Romanovsky, “Patterns for refinement automation,” in *International Symposium on Formal Methods for Components and Objects*. Springer, 2009, pp. 70–88.
11. N. McGreivy and A. Hakim, “Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations,”

Nature machine intelligence, vol. 6, no. 10, pp. 1256–1269, 2024.

12. J. O. Berger and L. A. Smith, "On the statistical formalism of uncertainty quantification," *Annual review of statistics and its application*, vol. 6, no. 1, pp. 433–460, 2019.
13. V. Kuleshov, N. Fenner, and S. Ermon, "Accurate uncertainties for deep learning using calibrated regression," in *International conference on machine learning*. PMLR, 2018, pp. 2796–2804.
14. D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.
15. B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.
16. Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.
17. N. Ståhl, G. Falkman, A. Karlsson, and G. Mathiason, "Evaluation of uncertainty quantification in deep learning," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, 2020, pp. 556–568.
18. N. Maslej, L. Fattorini, R. Perrault, Y. Gil, V. Parli, N. Kariuki, E. Capstick, A. Reuel, E. Brynjolfsson, J. Etchemendy *et al.*, "Artificial intelligence index report 2025," *arXiv preprint arXiv:2504.07139*, 2025.
19. S. E. Whang, Y. Roh, H. Song, and J.-G. Lee, "Data collection and quality challenges in deep learning: A data-centric AI perspective," *The VLDB Journal*, vol. 32, no. 4, pp. 791–813, 2023.
20. A. Souly, J. Rando, E. Chapman, X. Davies, B. Hasircioglu, E. Shereen, C. Mougan, V. Mavroudis, E. Jones, C. Hicks *et al.*, "Poisoning attacks on llms require a near-constant number of poison samples," *arXiv preprint arXiv:2510.07192*, 2025.
21. P. Parikh, M. Ali, N. Virani, and S. C. Evans, "Strain to synergy: Powering AI data centers with sustainable and smart power grid infrastructures," in *2025 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 2025, pp. 1–6.
22. A. S. George, A. H. George, and A. G. Martin, "The environmental impact of AI: a case study of water consumption by chat gpt," *Partners Universal International Innovation Journal*, vol. 1, no. 2, pp. 97–104, 2023.
23. H. Boesch, Y. Liu, J. Tamminen, D. Yang, P. I. Palmer, H. Lindqvist, Z. Cai, K. Che, A. Di Noia, L. Feng *et al.*, "Monitoring greenhouse gases from space," *Remote Sensing*, vol. 13, no. 14, p. 2700, 2021.
24. D. Hapsoro *et al.*, "Antecedents and consequences of carbon emissions'disclosure: Case study of oil, gas and coal companies in non-annex 1 member countries," *Journal of Indonesian Economy and Business: JIEB.*, vol. 33, no. 2, pp. 99–111, 2018.
25. D. Chauhan, P. Bahad, and J. K. Jain, "Sustainable AI: environmental implications, challenges, and opportunities," *Explainable AI (XAI) for sustainable development*, pp. 1–15, 2024.