Building AI/ML Model and Controversies: Scientific and Engineering Flaws

Evariste Somé, *Member, IEEE,* Jed Brown, *Member, Society,* libCEED 1, *Member, Society,* libCEED 2, *Member, Society,* libCEED 3, *Member, Society,* libCEED 4, *Member, Society,*
Email: email.email@email.org

***Abstract*—The growing proliferation of artificial intelligence (AI) systems coupled with machine learning (ML) techniques affects more areas of society and in various research fields which include fundamental sciences and technology. Designed to operate with varying levels of autonomy, AI has been developed with a view to obtaining useful insights, predicting, categorizing. 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. The purpose of this paper is to provide evidence-based concerns about 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.**

***Index Terms*—AI controversies, generalization robustness and trustworthiness, data quality and availability, privacy concerns, interpretability, uncertainty quantification, systematic refinement mechanisms, computation cost.**

I. Introduction

In recent years, one of the most advertised tools, named artificial intelligence (AI), seems to provide magic solutions to most scientific inquiries. Whether correct or not, AI systems are often perceived as being more objective than humans or as offering greater capabilities than general software. ML 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. The aim of AI is to develop a machine that can think like humans and mimic human behaviors, including perceiving, reasoning, learning, planning, predicting [1]. AI technologies can drive inclusive economic growth and support scientific advancements that improve the conditions of our world.

AI seems to be adopted in a vast array of industries that include education, industry, defense, and health care. In both industry and academia, there is a lively and ongoing debate about the uses of AI and its future [2]. It is presented that some research outcome is overoptimistic along with weak baselines, which leads to overly positive results, while reporting biases lead to under-reporting of negative results.

As interest in ML has grown, more and more scientific and engineering fields are exploring whether ML can be used to advance science [3] (Fig. 1). For some problems, ML has shown the potential to do so [4]. However, there are increasing concerns about reproducibility in ML-based science and engineering [5], [6]. Common pitfalls include data leakage [7], poor data quality [8], weak baselines [9], and inadequate external validation [10]. In each case, the pitfalls result in overoptimistic assessments of ML performance.

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 [11]. 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 [12]. 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 at recognizing faces without faces, and cells without cells. Some of these papers have been cited hundreds of times [12]. “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 want 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 because of methodological flaws or biases in image data sets [8]. The errors that Shamir and Dhar found are just some of the ways in which machine learning can give rise 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

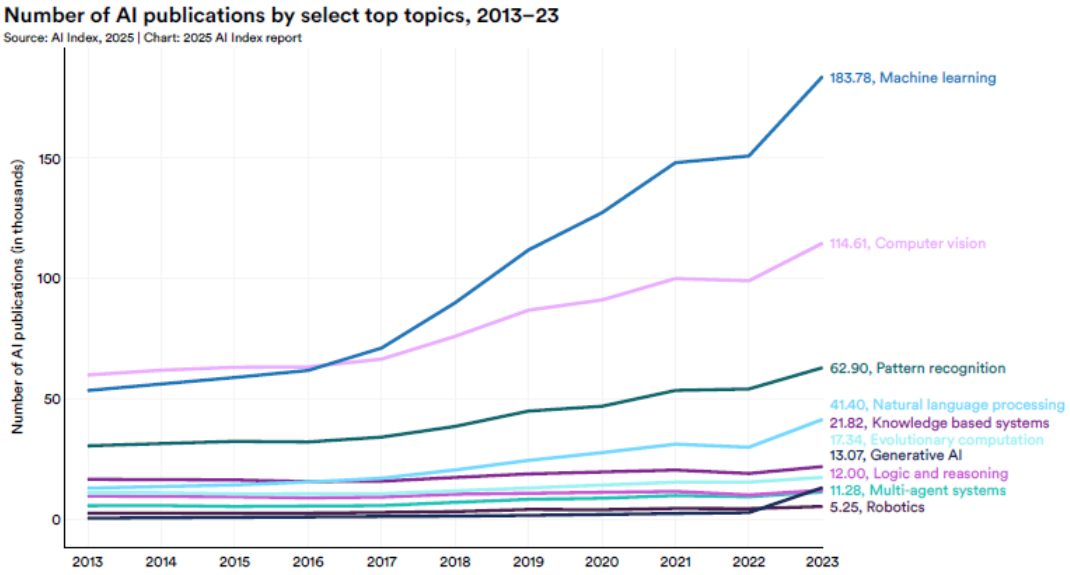

Fig. 1: Number of AI publications by select top topics, 2013–23.

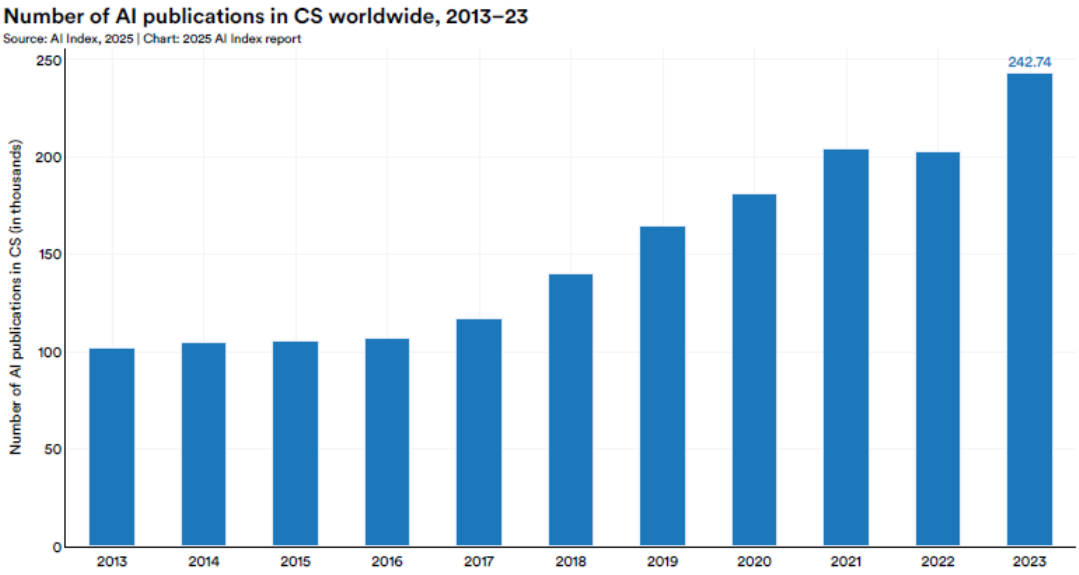

Fig. 2: Number of AI publications by select top topics, 2013–23.

between intelligence and understanding. AI can loop as many times, but only humans can understand in learning.

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 fields that they examined, affecting hundreds of papers [5]. 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 [5].

There are many common mistakes repeated over and over. Aeronautical engineer Lorena Barba at George Washington

University in Washington DC 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.”

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 environment dynamic, 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, meaning they are influenced by societal dynamics and human behavior [13].

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 [13].

As AI has the power of improving, accelerating discovery, it can also cause some problems if we are not aware of its potential negative effects.

In the remainder of this paper, Sections 2 through 8 decribe AI generalization, robustness and trustworthiness; data quality and availability; privacy concerns; interpretability; uncertainty quantification; systematic refinement mechanisms; and computation cost. Section 9 presents a summary.

II. AI Generalization, Robustness and Trustworthiness

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 either a high or low risk. Traditionally, scientific applications require trusted prediction with quantifiable error bounds.

Complex reasoning remains a problem. Even though the addition of mechanisms such as chain-of-thought reasoning has significantly improved the performance of large language models (LLMs), these systems still cannot reliably solve problems for which provably correct solutions can be found using logical reasoning, such as arithmetic and planning, especially on instances larger than those they were trained on. This has a significant impact on the trustworthiness of these systems and their suitability in high-risk applications [14].

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 harms include the reality that 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 upon a sense of collective responsibility among AI actors [13].

Robustness or generalizability is defined as the “ability of a system to maintain its level of performance under a variety of circumstances” (Source: ISO/IEC TS 5723:2022). 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 should perform in ways that minimize potential harms to people if it is operating in an unexpected setting [13].

III. Data Quality and Availability

While 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 research. Historically, development and evaluation of these algorithms have been hindered by a lack of well-annotated, large-scale, publicly available data sets.

AI need 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. Last year’s, 2024, AI Index explored various factors in this debate, including the availability of existing internet data and the potential for training models on synthetic data. New research this year 2025 suggests that the current stock of data may last longer than previously expected [14].

There is still not official data formats, and therefore, diferent systems store data in varied structures. Different data formats may cause inconsistent during the model training or deployment time. Equally, 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, which limits the training of accurate AI or ML models [15].

AI model perform well as good as the data is clean and continuously accessible, updated, and accurate. Outdated or inconsistent data will lead to inaccurate predictions. data quality and availability for AI model performance encompasses accuracy, completeness, consistency, timeliness and relevance, and data interoperability across systems.

IV. Privacy Concerns

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

Privacy refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity. 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). (See The NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management.) Privacy values such as anonymity, confidentiality, and control generally should guide choices for AI system design, development, and deployment. Privacy-related risks may influence security, bias, and transparency and come with tradeoffs with these other characteristics. Like safety and security, specific technical features of an AI system may 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 ("PETs") 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 [13].

However, managing AI data can face difficult decisions in balancing 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.

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. While 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 [14].

V. 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 assist 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 [13].

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 [16]. Note that 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. An interpretable model is one whose decisions can be easily understood by users. 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 [16]. Making AI data interpretable have gained attention to enhance the understanding of a machine learning algorithm, despite its complexity. Interpretability methods are approaches designed to explicitly enhance the interpretability of a machine learning algorithm, despite its complexity. The interpretability of an AI program is generally defined as the ability of a human to understand the link between the features extracted by an AI program and its predictions [17]. 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. An interpretable machine learning algorithm can be described as one in which the link between the features used by the machine learning system and the prediction itself can be understood by a human [17], [18]. A machine learning model based on hand-crafted features, such as a decision tree, is not necessarily interpretable just because the individual features are based on specific domain knowledge and are understandable by a human. 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. These more complex models tend to be more accurate and precise. But

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-enhanced, 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 a result of the opaque nature of AI systems (limited explainability or interpretability), lack of transparency or documentation in AI system development or deployment, or inherent uncertainties in AI systems [13].

VI. Uncertainty Quantification

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. Often, though not always, uncertainties are treated probabilistically. Very often, the probabilistic theory underlying uncertainty quantification methods is actually quiet simple, but is obcured by the details of the application. However, the complications that practical applications present are also part of the essence of uncertainty quantification: it is all very well giving an accurate prediction for some insurance risk in terms of an elementary mathematical object such as an expected value, but how will you actually go about evaluating that expected value when it is an integral over a million dimensional parameter space? Thus, it is important to appreciate both the underlying mathematics and the practicabilities of implementation [19]. Uncertainty quantification treats both types of uncertainty, aleatory and epistemic, incorporates uncertainty due to the mathematical form of the model, and it provides a procedure for including estimates of numerical error in the predictive uncertainty. Aleatory (random) uncertainties in model inputs are treated as random variables, while epistemic (lack of knowledge) uncertainties are treated as intervals with no assumed probability distributions.

In mathematics, uncertainty is often characterized in terms of a probability distribution. From that perspective, epistemic uncertainty means not being certain what the relevant probability distribution is, and aleatoric uncertainty means not being certain what a random sample drawn from a probability distribution will be. Typical scientific uncertainty quantification problems of interest include certification, prediction, model and software verification and validation, parameter estimation, data assimilation, and inverse problem [19]. Scientific simulations often require precise quantification of uncertainty and error bounds.

Some of the most fundamental questions in uncertainty quantification 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 [20].

Uncertainty quantification 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. 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 emphasises that its outcome, therefore, is uncertain [21].

To this date, it seems infeasible to build AI models that know how to function effective 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 [22].

It is crucial that future AI systems have the ability to not only function well in known domains, but also understand and show when they are uncertain when facing something unknown. It is an important research direction to find methods that allow for the quantification of uncertainty in the provided predictions and to find and understand their limitations [23]. When quantifying the uncertainty, it is essential that methods consider both the epistemic uncertainty (the uncertainty that arises due to lack of observed data) and the aleatory uncertainty (the uncertainty that arises due to underlying random variations within the collected data). Where the uncertainty arises due to lack of data and, hence, the model is forced to extrapolate its knowledge. When the model extrapolates it takes an uniformed decision, which can be far from optimal. The correlations between the quantified uncertainty of the different models are also very low, showing that there is an inconsistency in the uncertainty quantification. This inconsistency is something that needs to be further understood and solved before AI can be used in critical applications in a trustworthy and safe manner [23]. 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.

VII. Systematic refinement mechanisms

Very often, formal methods are a promising way to create software-models in a mathematically rooted way to keep them formal and analyzable. Development starts from an abstract model, which is gradually transformed into a specification closely resembling an implementation. Each model transformation step, called a refinement step, allows a designer to incorporate implementation details into the model. Correctness of refinement steps is validated by mathematical proofs. 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. 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 [24].

When it comes to refinement mechanism, too many AI research experiments lack rigor. It is not a critique of individual researchers. Competitive pressures from governments and research institutions, tight timelines, and genuine excitement about high-impact missions push 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. Research society needs well designed experiments so researchers can develop better algorithms based on empirical evidence and not just vibes. Researchers must balance designing experiments that preserve causal interpretability with extracting the maximum possible signal per unit time and per unit of compute.

The model driven refinement-based development process enables development of systems that are correct by construction. More manpower would be required upon data acquisition and refinement than in the training level. Although companies carry out AI research by focusing on learning rather than data acquisition and refinement for cost reduction, the outcome is imperfect, as they are unable to procure satisfactory training data. There is mathematician-power dependent (manpower-dependent) problem of the current image-based AI research, its improvement, and its future [25].

VIII. Computation Cost

According to Artificial Intelligence Index Report 2025, AI models get increasingly bigger, more computationally demanding, and more energy intensive. New research finds that the training compute 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 [14].

A frequent discussion around foundation models pertains to their high training costs. While AI companies rarely disclose exact figures, costs are widely estimated to reach into the 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 [14]. Understanding the costs associated with training AI models remains important, yet detailed cost information remains scarce. Last year, 2024, the AI Index published initial estimates on the costs of training foundation models. This year, 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 [14].

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.

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. Figure (**??**) illustrates the energy efficiency of Tensor-FP16 precision hardware, measured in FLOP/s per watt. For instance, the Nvidia B100, released in March 2024, achieved an energy efficiency of 2.5 trillion FLOP/s per watt, compared to the Nvidia P100, released in April 2016, which reported 74 billion FLOP/s per watt. This means the B100 is 33.8 times more energy efficient than the P100.

Despite significant improvements in the energy efficiency of AI hardware, the overall power consumption required to train AI systems continues to rise rapidly. Figure (**??**) illustrates the total power draw, measured in watts, for training 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 rising power consumption of AI models reflects the trend of training on increasingly larger datasets.

IX. Discussion and Outlook

AI is increasingly becoming an essential part in our time. From healthcare to telecommunication, to education, to transportation, to space exploration, etc., AI is rapidly moving from the laboratory research to daily life.

This paper describes the purpose and risks of AI features are, and what kinds of scientific governance could be abided by or built into the AI pipeline.

The purpose of this paper is not to diminish the scientific contribution in artificial intelligent and machine learning, but rather to assess alignment with the rigorous scientific criteria for a true model that is robust and trustworthy, interpretable, available data quantitatively and qualitatively, uncertainty quantifiable, and take 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 AI expected outcome, meaning if AI wants to be precise scientific. There are many reasons for scientists to worry about the use of AI, which include lack of large-scale and publicly available

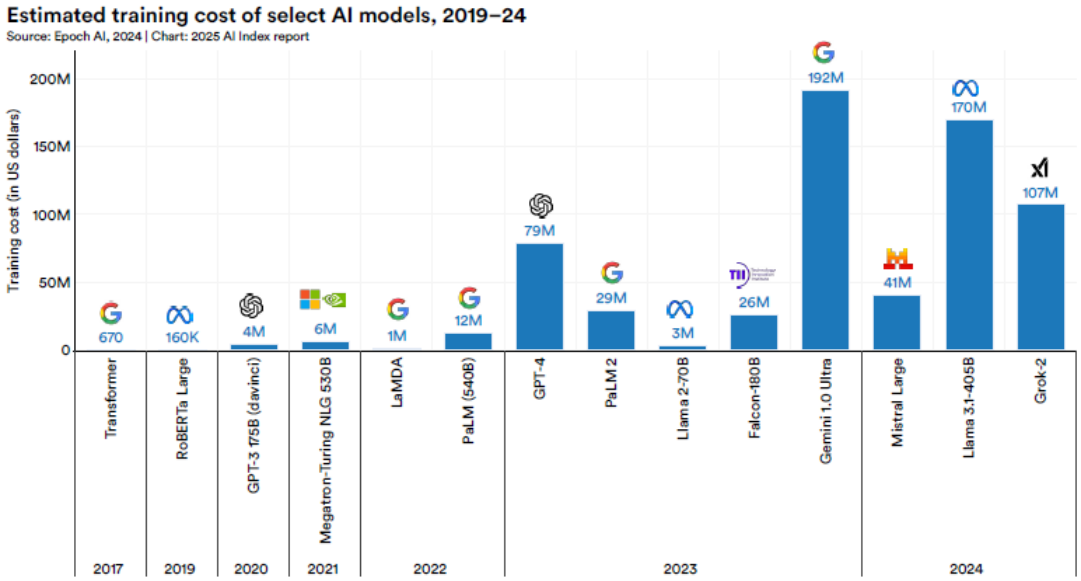

Fig. 3: estimated training cost associated with select AI models, based on cloud compute rental prices.

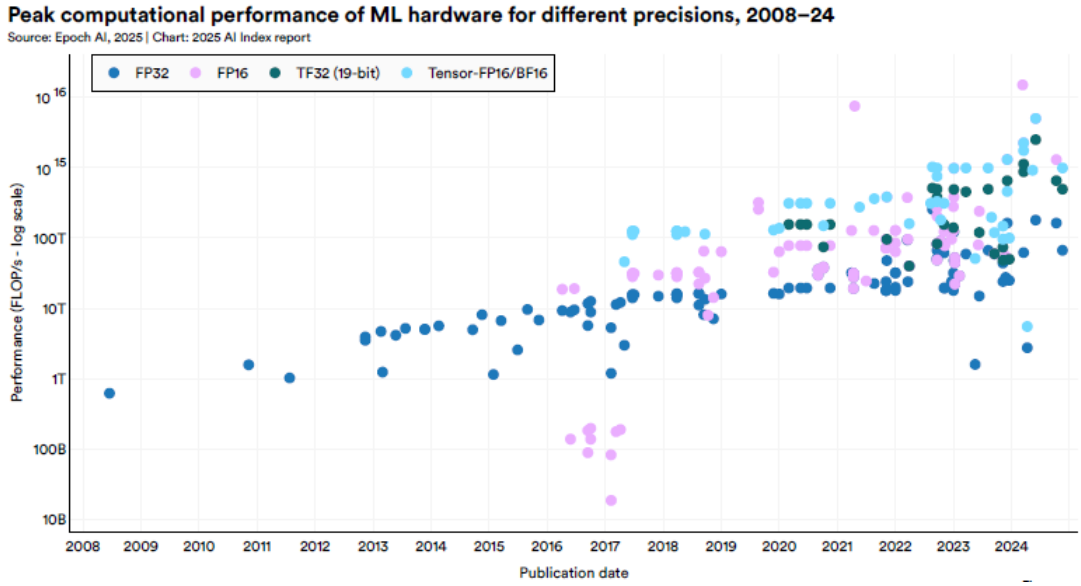

Fig. 4: Peak computational performance of ML hardware for different precisions, 2008-24.

data sets, lack of systematic refinement mechanism rigor, AI models get increasingly bigger and therefore more computationally demanding and more energy intensive, uncertainty quantification prediction not reliable, lack of transparency in the operation of AI systems, measurement approaches are oversimplified, gamed, lack critical nuance, and its deployment may raise ethical concerns around data privacy, data security, and ownership.

Though AI promises a brighter future, scientists must not start thinking that AI knows everything and can replace human researchers. AI could assist scientists researchers in understanding large amounts of data. However, it cannot replace creativity, intuition, and critical thinking skills that are essential in scientific research. Be dependent too much on AI can limit the scientific discoveries and lead to a lack of diversity in research perspectives.

ACKNOWLEDGMENT

This work was carried out with the financial support of libCEED.

REFERENCES

[1] Y. Xu, X. Liu, X. Cao, C. Huang, E. Liu, S. Qian, X. Liu, Y. Wu, F. Dong, C.-W. Qiu et al., "Artificial intelligence: A powerful paradigm for scientific research," The Innovation, vol. 2, no. 4, 2021.

[2] S. Lindgren, "Introducing critical studies of artificial intelligence," in Handbook of critical studies of artificial intelligence. Edward Elgar Publishing, 2023, pp. 1–19.

[3] 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.

[4] J. Jumper et al., "Highly accurate protein structure prediction with alphafold. nature596, 583-589 (2021)," Cited on, p. 28.

[5] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, 2023.

[6] S. Kapoor, E. Cantrell, K. Peng, T. H. Pham, C. A. Bail, O. E. Gundersen, J. M. Hofman, J. Hullman, M. A. Lones, M. M. Malik et al., "Reforms: Reporting standards for machine learning based science," arXiv preprint arXiv:2308.07832, 2023.

[7] S. Whalen, J. Schreiber, W. S. Noble, and K. S. Pollard, "Navigating the pitfalls of applying machine learning in genomics," Nature Reviews Genetics, vol. 23, no. 3, pp. 169–181, 2022.

[8] 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.

[9] O. DeMasi, K. Kording, and B. Recht, "Meaningless comparisons lead to false optimism in medical machine learning," PloS one, vol. 12, no. 9, p. e0184604, 2017.

[10] L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, E. Albu, B. Arshi, V. Bellou, M. M. Bonten et al., "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," bmj, vol. 369, 2020.

[11] 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.

[12] 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.

[13] N. AI, "Artificial intelligence risk management framework (ai rmf 1.0)," URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai, pp. 100–1, 2023.

[14] 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.

[15] 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.

[16] 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.

[17] M. Reyes, R. Meier, S. Pereira, C. A. Silva, F.-M. Dahlweid, H. v. Tengg-Kobligk, R. M. Summers, and R. Wiest, "On the interpretability of artificial intelligence in radiology: challenges and opportunities," Radiology: artificial intelligence, vol. 2, no. 3, p. e190043, 2020.

[18] M. Van Lent, W. Fisher, and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in Proceedings of the national conference on artificial intelligence. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004, pp. 900–907.

[19] T. J. Sullivan, Introduction to uncertainty quantification. Springer, 2015, vol. 63.

[20] 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.

[21] 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.

[22] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," arXiv preprint arXiv:1610.02136, 2016.

[23] 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.

[24] 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.

[25] S. Heo, S. Han, Y. Shin, and S. Na, "Challenges of data refining process during the artificial intelligence development projects in the architecture, engineering and construction industry," Applied Sciences, vol. 11, no. 22, p. 10919, 2021.