AIPA International Journal on Artificial Intelligence: Bridging Technology, Society and Policy

AIPA's INTERNATIONAL JOURNAL ON ARTIFICIAL INTELLIGENCE

Bridging Technology, Society and Policy

Volume 1 | Issue 1 | 2024

ISSN: 3062-097X

Artificial Intelligence Policies Association (AIPA)

Editorial Team

- Prof. Dr. Fatma Zehra SAVİ
- Assoc. Prof. Dr. Barış ÖZÇELİK
- Assoc. Prof. Dr. Oğuz GÖKSU
- Assoc. Prof. Dr. Oğuz KUŞ
- Assoc. Prof. Dr. Ayşe Elif Pasos DEVRANİ
- Assoc. Prof. Dr. Arzu AL
- Assoc. Prof. Dr. Gamze YÜKSEK
- Assoc. Prof. Dr. Serdar TORT
- Assoc. Prof. Dr. Özgür YILMAZ
- Dr. Mevlüt UZUN
- Dr. Burcu Evrim İÇTENBAŞ
- Dr. Elif Karakoç KESKİN
- Dr. Nida Gökçe NARİN
- Dr. Osman Gazi GÜÇLÜTÜRK
- Dr. Ömer GÜNEŞ
- Dr. Onur YÜKSEL
- Ekin Altun ÖZAYDIN
- Afra Teren GÜRLÜER
- Buse İlay YILDIZ
- Alparslan ERDAĞ
- Dr. Ahsem TELLİOĞLU

Contents

Publication Date: 09.12.2024

AIPA's International Journal on Artificial Intelligence: Bridging Technology, Society and Policy
ISSN: 3062-097X
Published: 07 December 2024

Artificial Intelligence in Technology and Innovation: Concerns and Future Directions

Ekin Can Erkuş[1*], Can Özbey[2], and Talha Çolakoğlu[3]

[1]Intelligent Application DC Dept.,Huawei Türkiye R&D Center, ORCID: 0000-0002-2445-5929
[2]Intelligent Application DC Dept.,Huawei Türkiye R&D Center, ORCID: 0009-0005-8432-9413
[3]Intelligent Application DC Dept.,Huawei Türkiye R&D Center, ORCID: 0000-0002-4524-862X

ORIGINAL RESEARCH PAPER

Abstract

This article explores current trends and future projections of artificial intelligence (AI) in technology, economics, and the environment. Using data from sources like the AI Index, AI history, and AI investments, we analyze AI-driven technological advancements. The study looks at the growth and impact of advanced language models in natural language processing (NLP), the economic effects of AI, and the sustainability challenges of its computational needs. We address these issues by suggesting strategies for future research and governance to align AI development with social and environmental goals. By examining historical AI data, we highlight AI's significant role in shaping the tech landscape. Our opinion is based on current reviews, evidence, and concerns, situating our findings within broader AI and innovation discussions. We offer interpretations and thoughtful speculations on the future direction of AI technology, providing practical insights and strategic advice to guide its development.

Keywords: Artificial Intelligence, Technology, Innovation, AI Index, AI companies, Future of AI

1 Introduction

Artificial intelligence (AI) is rapidly changing the current state of technology and innovation across a variety of industries [1]. AI is an interdisciplinary field that uses machine learning, deep learning, and other advanced computational techniques to create systems capable of performing tasks that would normally require human intelligence, and beyond [2]. The integration of AI into various industries has resulted in significant improvements in efficiency, accuracy, and innovation, allowing faster development in basic data processing to more complex decision-making and problem-solving processes [3]. However, with these advancements come significant challenges and concerns, particularly in terms of economic impact and environmental sustainability. As AI keeps growing, it will affect industries more and the everyday lives of people, with the need for cautious implementation using its potential [4].

AI has become a significant driver of innovation and efficiency in a variety of industries, as it automates coding tasks, optimizes algorithms, and automates manufacturing processes, lowering operational costs and improving product quality [5]. The rapid proliferation of AI technologies is reflected in the growing number of AI-related patent applications, indicating increased interest and investment in AI research and development [6]. Moreover, the economic impact of AI is significant, providing opportunities for growth and innovation while also posing challenges such that large corporations have the resources to fully utilize AI, whereas smaller organizations may struggle to keep up, potentially leading to market consolidation and increased economic inequality [7]. Along with socioeconomic balance concerns, future AI research and development should also prioritize minimizing the environmental impact, particularly in terms of energy consumption [8].

The motivation behind this paper is to provide a comprehensive analysis of the diverse impacts of AI. By considering the technological, economic, and environmental components, this paper provides a balanced view of AI's current condition and future direction. It also tries to highlight possible issues and provide strategies for future research and governance to ensure that AI progress is consistent with broader social objectives.

The paper is organized as follows: The section "AI-Driven Trends and Future Projections Across Various Sectors" investigates how AI is driving innovation in a variety of technological sectors. The "Evolution and Impact of Advanced Language Models in Natural Language Processing" section discusses recent achievements in natural language processing and their implications. In "The Economic Impact and Future of Artificial

OPEN ACCESS
AIPAJ Vol:1, Issue:1

***Corresponding author**
ekincanerkus@hotmail.com

***Submitted** 10 June 2024*

***Accepted** 28 November 2024*

Citation
Erkuş, E. C., Özbey, C., Çolakoğlu, T. Artificial Intelligence in Technology and Innovation: Concerns and Future Directions, In AIPA's International Journal on Artificial Intelligence: Bridging Technology, Society and Policy, (Vol. 1, Number 1, pp. 1–12) DOI: 10.5281/zenodo.14340947

Intelligence" how AI affects labor markets and economic growth is considered. "The Computational Evolution and Future Demands in Artificial Intelligence" section discusses the resource requirements of AI systems and their sustainability. Finally, the "Future Directions and Sustainability in Artificial Intelligence" section states possible future directions for AI growth and highlights related dangers that need to be handled by also including a subsection named "Concerns with the Environmental Impact of Artificial Intelligence" which examines the environmental consequences of wide AI use. By addressing these concerns, this research hopes to contribute to a humble understanding of AI's role in current technology and innovation. It points out the importance of strategic planning in maximizing the benefits of AI while reducing its possible disadvantages. This study also aims to contribute to the ongoing discussion on AI in technology and innovation by analyzing current developments and discussing the potential future.

2 AI-Driven Trends and Future Projections Across Various Sectors

Artificial intelligence has had a major effect on technological growth, driving major improvements in a wide range of industries, including software development, the medical field, manufacturing, finance, transportation, and general research and development [9, 10]. AI's impact on technological development is already evident in the rapid rise in the number of patent applications, indicating the fast adoption and integration of AI technologies [11].

Hereby, Figure 1 shows the number of patent applications related to AI from 2010 to 2019 across various sectors, according to the data obtained from an open database [12]. As stated, the telecommunications sector has experienced a dramatic expansion, particularly since 2016, followed by an increase in the banking, life sciences, and transportation sectors. This trend points out the increasing popularity and accelerating adoption of AI technologies.

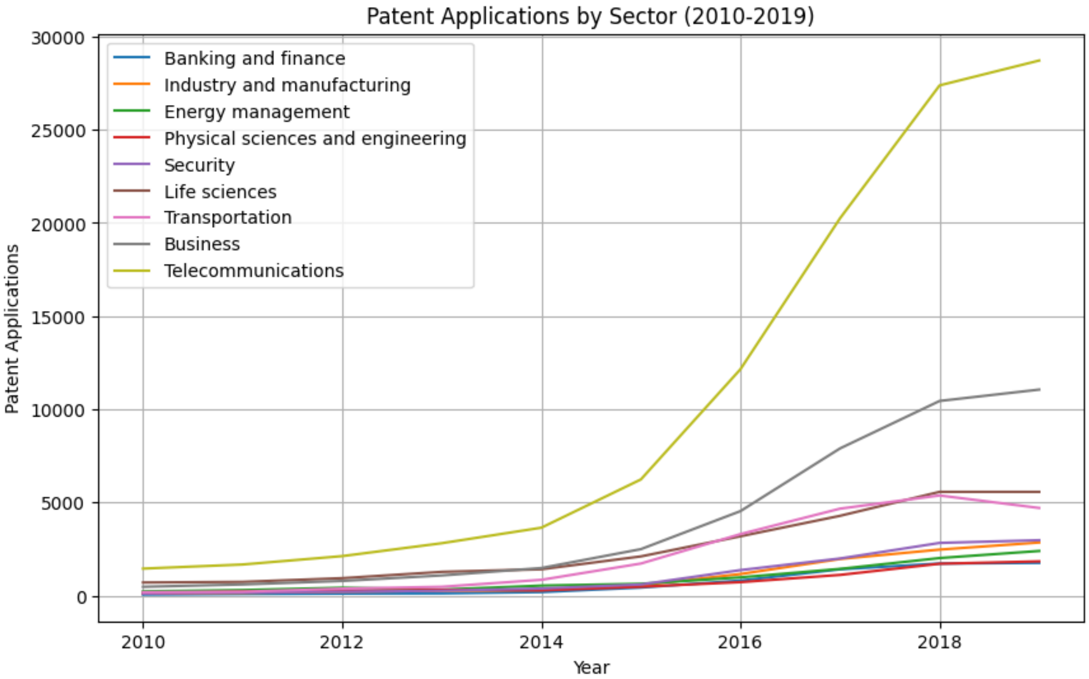


Figure 1. Total patent applications per year related to AI for different sectors.

AI-driven robotics and automation structures have improved manufacturing procedures and reduced operational costs [10]. For instance, predictive maintenance systems, which predict machine failures before they occur, reduce downtime while lowering costs [13]. By improving product quality and reducing human error, AI enhances the performance and productivity of manufacturing operations, resulting in significant cost savings and increased long-term viability [14]. AI integration in production not only improves operational performance but also increases the industry's ability to innovate and respond quickly to customer demands [15].

Moreover, the finance industry has also been reshaped by AI, with algorithms developing sophisticated trading strategies, dealing with risks, and detecting fraud [16]. Especially on high-frequency trading platforms, AI is used to research market trends and perform trades with high rates and precision that exceed human capabilities [17]. AI threat control systems investigate the creditor's reliability and forecast market fluctuations, reducing economic risks for banks [18], and AI fraud detection algorithms detect suspicious activity in real-time, thereby improving financial operations' reliability and safety, strengthening the financial industry's ability to adapt to emerging threats and possibilities [19].

In the transportation sector, AI is being used to improve self-driving vehicles, intelligent traffic management systems, and better logistics solutions [20]. Self-driving cars, powered by modern artificial intelligence algorithms, have the potential to reduce accidents while improving fuel efficiency [21], and AI-based monitoring systems optimize traffic flow, lowering congestion and emissions [22]. Logistics companies use AI to automate their operations, ensuring timely deliveries and efficient resource utilization [10]. Therefore, the integration of AI in transportation not only improves operational performance but also contributes to more secure and sustainable cities.

Artificial intelligence has also revolutionized research and development in a variety of fields. For example, in medicinal products, AI accelerates drug discovery by studying massive datasets of molecular systems and biological pathways to predict the efficacy of new compounds and identify potential side effects [23]. Materials technology benefits from AI models that predict the properties of new substances before synthesis, reducing the time and cost of experimentation [24]. AI simulation tools in engineering provide accurate projections of product performance, allowing for more efficient designs and reducing the need for physical prototypes [25]. These advances in R&D encourage innovation and accelerate the pace of discovery and product improvement.

AI-powered diagnostic devices that use advanced machine learning (ML) algorithms have accelerated medical imaging by detecting anomalies and diagnosing problems more accurately and rapidly [26]. These structures have transformed healthcare through the use of early disease detection and the development of personalized treatment techniques based on individual patient data [27]. Furthermore, predictive analytics in biomedical engineering improve resource utilization and patient experience by simplifying clinic processes [28].

AI also has a significant impact on mobile technology, which is increasingly being integrated into smartphones, smart speakers, wearables, and IoT devices, where it provides personalized recommendations, predicts user preferences, and automates routine tasks resulting in a paradigm shift in device design and capabilities [29]. Personal assistants, such as Siri and Alexa, provide a personalized experience through voice commands, proactive signals, and context-aware instructions, responding to user behavior, simplifying the user experience [30]. Modern smartphones include features like facial recognition, natural language understanding, and predictive text input, which improve security, usability, and efficiency while also reshaping user interactions, making devices more adaptable and personalized [31]. This shift is making technology more intuitive and user-centric, changing how we interact with electronic devices daily.

Figure 2 presents the forecasted patent applications related to AI from 2010 to 2028 for various sectors. It suggests significant growth in sectors such as security, business, transportation, and life sciences, implying that these fields will continue to drive innovation and AI adoption in the future.

AI has also accelerated the development of new technologies such as natural language processing (NLP) and conversational AI, allowing virtual assistants to recognize and respond to human speech in real-time [32]. Artificial intelligence-driven advancements in augmented reality (AR) and virtual reality (VR) have transformed entertainment, gaming, and experience enhancement, minimizing the distinction between the physical and virtual worlds [33]. As a result, the rise of AI has led to significant changes in consumer behavior, with people increasingly relying on AI-powered products and applications to simplify tasks, benefit from readily available information, and speed up decision-making [34].

3 Evolution and Impact of Advanced Language Models in Natural Language Processing

Advances in language modeling have significantly improved the capabilities of NLP systems, making them more context-aware and effective across a variety of applications [35]. Language modeling has evolved from basic statistical methods and early n-gram models to recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), resulting in the state of the art with transformers and generative AI [36]. This progress has been accelerated by the demand for more accurate and context-aware language understanding, as well as the exponential increase in computational sources [37].

The simplest type of language modeling is based on n-gram models as they depend on the co-prevalence of words to predict the next phrase in a sequence, shooting the probability of a phrase given the previous $n-1$

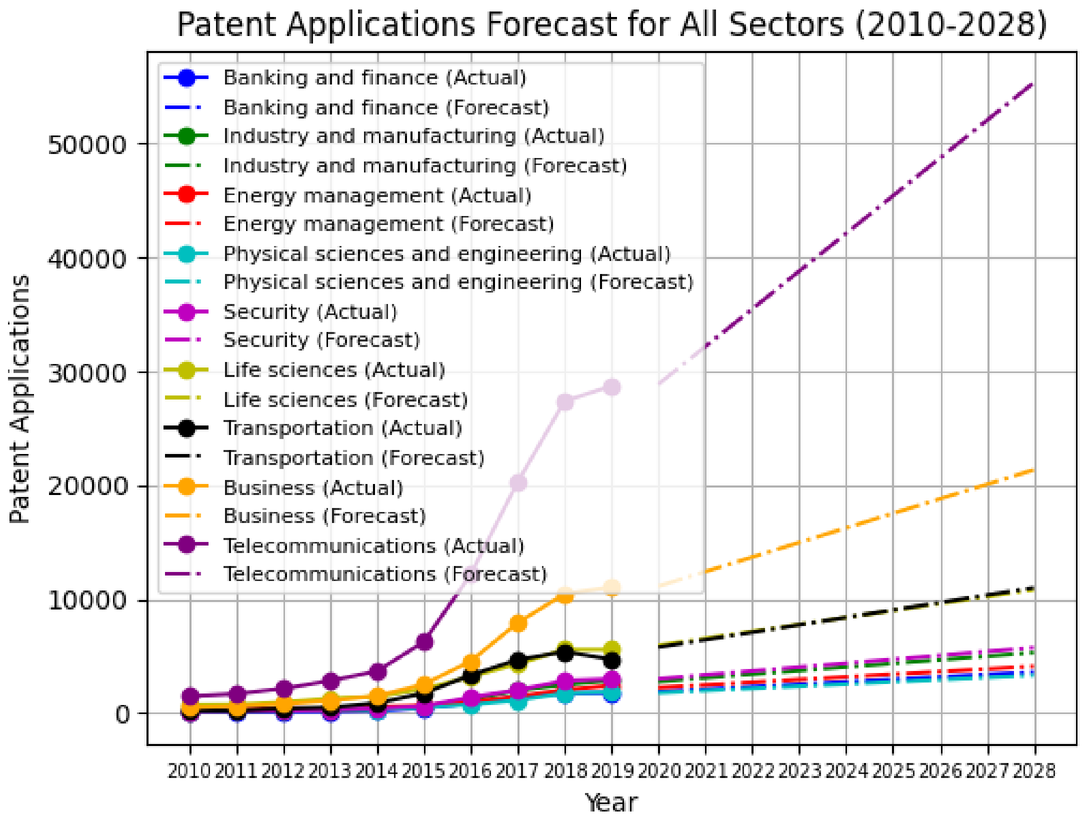


Figure 2. Forecasted total patent applications per year related to AI for different sectors.

words [38]. The basic representation can be found in Equation 3. For example, in a bigram version (n=2), the prediction of a word is completely dependent on the preceding phrase [39]. Because of their simplicity and low computational requirements, n-gram models allow them to run on machines without advanced hardware such as GPUs, consuming significantly less power [40, 41]. However, their limited context window limits their ability to capture more complex dependencies in texts, resulting in poor results for tasks that require more knowledge of the context [37].

$$P(w_i|w_{i-(n-1)},\ldots,w_{i-1}) = \frac{C(w_{i-(n-1)},\ldots,w_i)}{C(w_{i-(n-1)},\ldots,w_{i-1})} \quad (1)$$

Here, $P(w_i|w_{i-(n-1)},\ldots,w_{i-1})$ represents the probability of word w_i given the previous $n-1$ words, and $C(\cdot)$ denotes the count of occurrences of the given sequence in the training corpus.

The development of RNNs and LSTMs posed an important step in language modeling as these models are under the deep learning paradigm and offer a more advanced method of encoding context than n-gram models [41]. RNNs are intended to handle sequential data by maintaining a hidden state that updates as it approaches each element of the collection, theoretically capturing dependencies of any length [42]. LSTMs, a type of RNN, deal with the problem of vanishing gradients, allowing them to capture long-term dependencies more accurately [42]. This ability has made RNNs and LSTMs essential in improving the performance of many NLP tasks. Transformers, on the other hand, have the attention mechanism at their core and have transformed the field of NLP by eliminating the sequential processing of RNNs and instead processing the entire input sequence at once [43, 44]. The attention mechanism allows the model to assess the importance of various words in the input sequence, capturing complex dependencies regardless of their distance from one another [44]. This parallel processing capability increases efficiency and scalability, enabling transformers to handle larger datasets and more complex tasks [43].

Recent developments in generative AI (GenAI) have led to a surge in data and resource usage, with Large language models (LLM) models requiring vast datasets and extensive computational power [45, 46]. LLMs have revolutionized NLP by generating human-like text, but they require substantial resources for training and deployment, raising questions about their economic viability and environmental sustainability [47]. They also enable groundbreaking applications in content creation, automated customer support, and advanced research tools, but their substantial computational and energy requirements raise concerns about environmental impact and resource allocation [48].

4 The Economic Impact and Future of Artificial Intelligence

Advancements in GPU technology, the high costs of training and maintaining large AI models, and the transformative impact of AI-augmented research and development all contribute to the artificial intelligence economy [49]. While AI provides significant economic opportunities by increasing competitiveness, opening up new markets, and improving operational efficiencies, it also introduces challenges such as high costs, market consolidation, and workforce adaptation [50]. Dealing with these challenges is essential for maximizing AI's economic benefits and ensuring development in the long run.

GPUs have played an important role in the advancement of machine learning and artificial intelligence, and eventually, GPUs' price-performance ratio has improved dramatically over the years as semiconductor technology has advanced and economies of scale have increased [51]. This improvement has resulted in increased computational power and memory capacity at lower costs, however, the demand for high-performance GPUs frequently exceeds supply, causing price volatility [52]. As AI models become more complex, the demand for more powerful and efficient GPUs grows, emphasizing the importance of ongoing innovation in GPU technology, which makes continued investment in GPU research and development critical to meeting the increasing demands of AI applications [53].

Training and operating commercial LLMs are highly costly, since training these models necessitates the processing of massive datasets across thousands of GPUs for a long time, resulting in significant costs for computational resources, electricity, and cooling systems [47]. Although companies have experimented with various revenue models to help them reduce these costs, the costs of training and maintaining these models are prohibitively expensive for many end users, making them economically unfeasible without significant financial support [54]. One common approach is to incorporate advertising into AI services, which compensates costs with ad revenue [55]. Other approaches include providing subscription-based access to advanced features or forming partnerships with cloud service providers to distribute computational load [56]. These models contribute to increasing the accessibility of powerful AI tools while also generating revenue to cover high operational costs.

AI systems can process large amounts of data, identify patterns, and generate insights much faster than humans can, causing accelerated research processes that shorten product development cycles and speed up market entry [57]. AI-augmented R&D is transforming how companies innovate and develop new products, which improve competitiveness, accelerate growth, and open up new markets [58]. AI is enabling breakthroughs in industries such as pharmaceuticals, materials science, and energy, resulting in more effective drugs, advanced materials, and optimized energy systems, thus, integrating AI into R&D processes is crucial for companies seeking a competitive advantage [53, 51]. AI algorithms that provide predictive analytics, personalized marketing, and automated customer service are becoming increasingly common in a variety of industries, generating significant economic value [34]. These innovations improve not only operational efficiency but also customer experiences and satisfaction, resulting in increased revenue and market share [59]. Furthermore, the adoption of AI technologies is transforming traditional business practices, resulting in more agile and responsive business models capable of rapidly adapting to market changes [60].

Figure 3 illustrates the projected exponential growth in corporate investment in AI from 2022 to 2028, and the data were obtained from an open-sourced dataset from kaggle.com [12]. Using actual investment data, which represents the total corporate investment in artificial intelligence, adjusted for inflation up to 2021, the graph forecasts future investment using an exponential regression algorithm.

This exponential growth indicates a significant increase in resources dedicated to AI, highlighting the importance of future technological developments. The rapid increase in forecasted investment demonstrates confidence in AI's transformative potential across industries, emphasizing the necessity of proper planning and investment to optimize AI's economic benefits [15].

5 The Computational Evolution and Future Demands in Artificial Intelligence

Machine learning has advanced significantly, with three distinct eras: simple models (pre-2010), deep learning (2010-2020), and the current era of transformer models and generative AI (post-2020) [36]. Each era represents significant changes in computational demands, which have increased dramatically over the

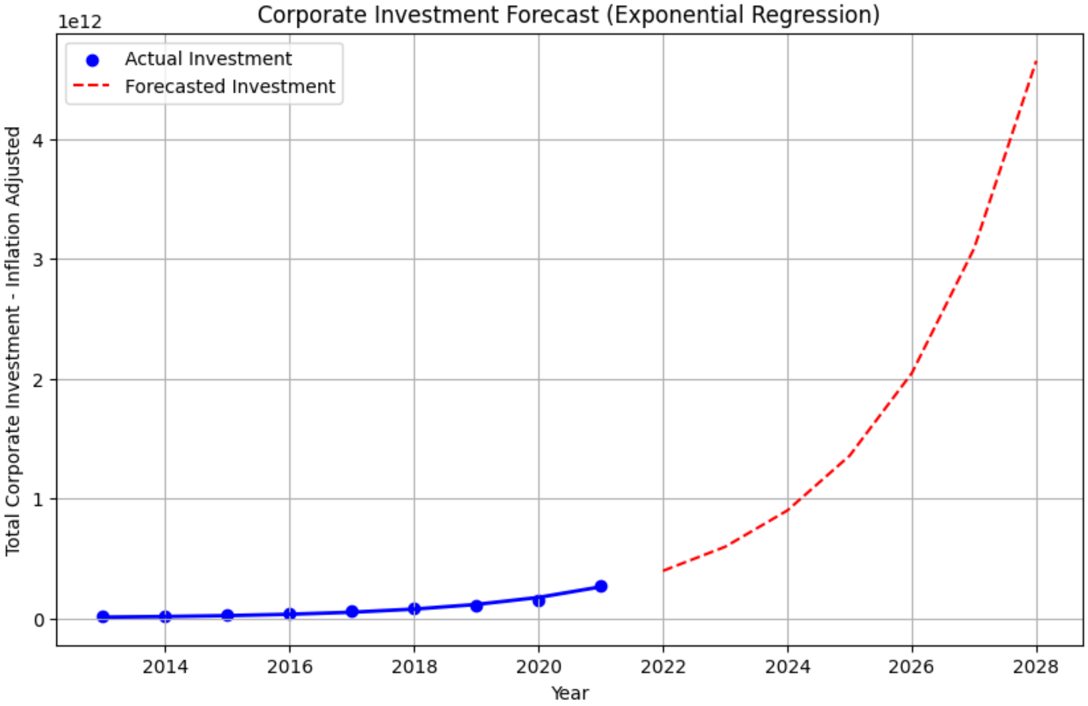


Figure 3. Forecasting future corporate investment in AI.

years. Pre-2010 models consisted of classical machine learning classifiers such as decision trees, linear regression, and support vector machines, where the computational requirements were relatively low, and most tasks could be completed using standard CPUs [61]. These models' simplicity required less data and computational power, making them accessible and efficient given the technical limitations of the time [51].

From 2010 to 2020, the development of machine learning changed with the introduction of deep learning, especially convolutional neural networks (CNNs) and recurrent neural networks (RNNs) and these models significantly increased computational demands, necessitating the use of GPUs to handle the large datasets required for effective training [62]. Moreover, GPUs' parallel processing capabilities allowed for more efficient training of these complex models, resulting in improvements in domains such as image and speech recognition [63].

On the other hand, transformers and generative AI (GenAI) models have impacted AI developments since 2020, increasing computational requirements to new levels [45]. Furthermore, while transformers provide significant benefits, they do so at a high cost because their training necessitates massive data and computational resources, with the attention mechanism initially having quadratic computational complexity to sequence length [64]. This makes transformers extremely resource-demanding, requiring high-RAM GPUs or specialized hardware such as TPUs (Tensor Processing Units) [65]. Figure 4 shows the LLM models by 2023 with their Massive Multitask Language Understanding (MMLU) scores and training compute relationships in terms of petaFLOPs, according to the open-access data from [12]. This figure simply presents the high number of computational units required to achieve the higher MMLU scores to reach the level of the models for understanding the human dialect.

The exponential increase in model sizes has resulted in a phenomenon known as the parameter gap, in which modern models, particularly those in NLP, contain billions of parameters [66]. The growth is motivated by the need for a more sophisticated understanding and generation of human language, but it poses significant challenges in terms of data, training time, and computing resources [49]. Modern NLP models require large datasets to function effectively, which presents several challenges, such as it becomes more difficult to ensure data quality with potential accuracy and relevance issues with the growing size of the datasets [67]. Security issues and regulations make data availability more difficult, even bringing up the continuous data scaling problem [68]. Accordingly, a forecast for the training compute in petaFLOPs can be found in Figure 5, which suggests that the computational complexity of the models will be more than 6 times that of the current models in 2023, based on an open-sourced dataset [12].

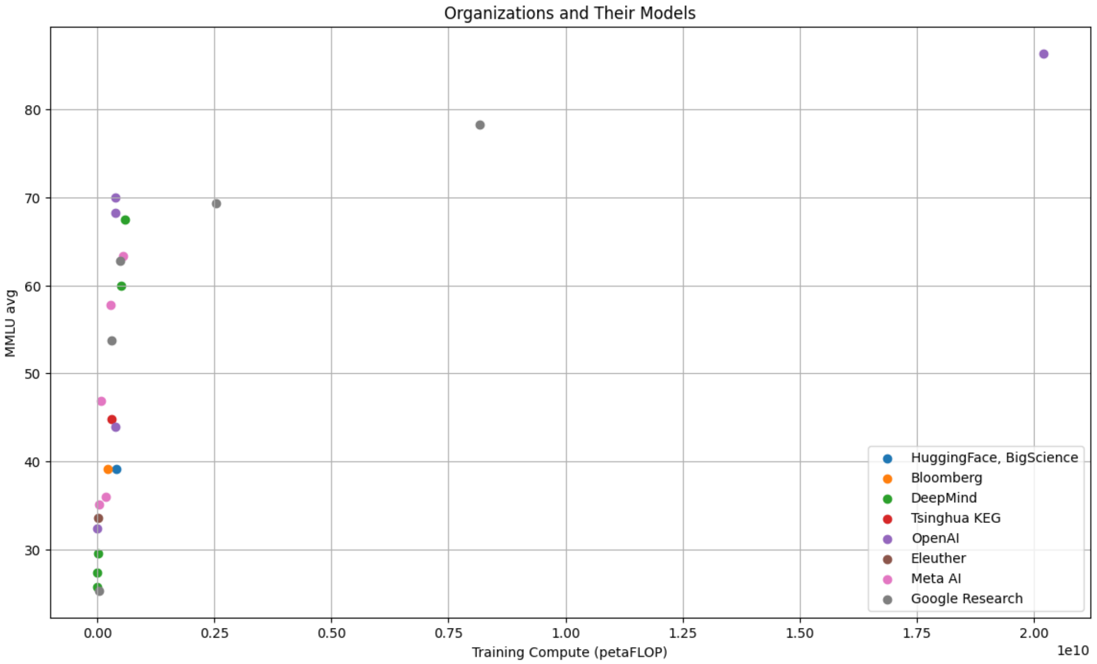


Figure 4. Published LLM models and corresponding organizations in 2023 for their MMLU and training compute values.

Motivated by the current computational demands for better LLM models, as given in Figure 4, a forecast of the average MMLU has been made. Accordingly, the forecasting model whose results can be found in Figure 6 suggests that the LLM models will almost completely be able to understand the human dialect as the predicted MMLU scores converge to 100% by the year 2028.

The future of AI has to find a balance between increasing computational capabilities and managing resource consumption since innovations in model efficiency, data usage, and hardware design will be essential in ensuring that the computational demand for AI remains stable [69].

6 Future Directions and Sustainability in Artificial Intelligence

Artificial intelligence has transformed many aspects of our daily routines, introducing groundbreaking discoveries and altering our interactions with technology [70]. AI's potential to revolutionize society is immense, offering new opportunities for innovation and growth. However, this potential is accompanied by concerns about energy consumption and environmental impact [8]. Balancing the benefits of AI with its resource demands is essential to ensuring that technological progress does not harm the environment, since finding this balance is vital for sustainable development [71].

The rapid growth of AI, driven by leading technology companies and rising demand, also has an unpredictable future. As AI evolves, a balance between its enormous potential and the need for responsible development should be managed [72]. By focusing on efficient models, sustainable data practices, and innovative computational technologies, AI can have a positive impact on society while maintaining environmental and economic health [8]. Embedding these principles in AI development promotes ongoing innovation and societal well-being, ensuring that AI's growth benefits the future [73]. Future initiatives should focus on developing more efficient AI models that can lower economic and environmental costs. This includes investigating sustainable data management procedures and developing computational technologies to improve AI's sustainability [74]. By prioritizing efficiency, ecological damage can be reduced while also making AI more economically viable [36].

6.1 Concerns with the Environmental Impact of Artificial Intelligence

The economic and ecological effects associated with the computational demands of AI models are mostly due to training large models requiring extensive data processing with a high number of GPUs, resulting in high costs for computational resources, electricity, and cooling [71]. Language models have evolved from the early versions focusing on word sequence analysis to the introduction of recurrent neural networks,

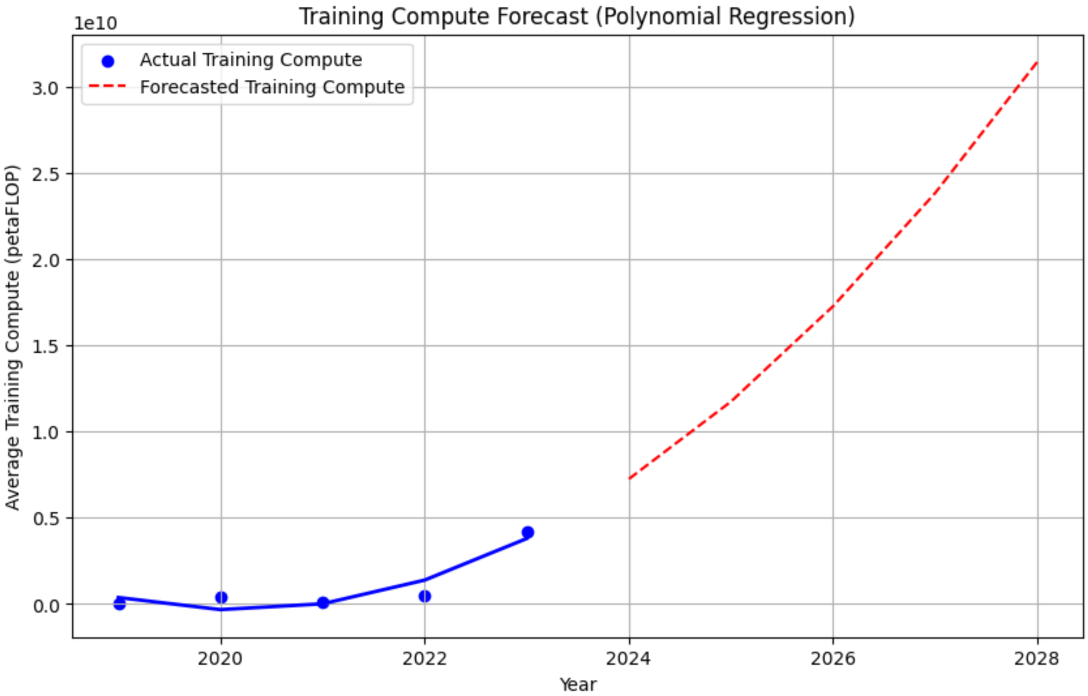


Figure 5. A forecasting of the training compute in petaFLOPs.

transformers, and generative AI with increased data complexity, resulting in higher energy consumption and a larger environmental footprint [46]. Training large language models has a significant environmental impact; training a single model emits as much carbon dioxide as several cars over its lifetime [75]. However, renewable energy alone is not a complete solution because it can be allocated to other business sectors too.

The energy consumption associated with AI training and inference increases carbon emissions and creates environmental issues [71]. The AI community may reduce these effects by promoting environmentally friendly technologies, using renewable energy, and establishing regulations that ensure AI technology growth aligns with environmental goals [19]. Many regions still rely heavily on fossil fuels for electricity generation, further compounding the environmental impact of AI's energy use and contributing to global warming and climate change [76]. Additionally, the energy-intensive nature of AI operations imposes financial burdens on organizations, increasing operational costs and potentially limiting the accessibility of AI technologies to smaller enterprises and low-developed countries [77].

The future of AI will require a decrease in resource consumption by introducing several strategies including optimizing AI models' energy efficiency, using model compression techniques, and using energy-efficient hardware [78]. Increasing the use of renewable energy sources in data centers can help reduce their carbon footprint while promoting AI systems with lower computational requirements can increase AI's accessibility and sustainability [8].

Conflict of interest

The authors have no conflict of interest to declare.
The authors declare that they have no competing financial interests.

References

[1] Apell P, Eriksson H. Artificial intelligence (AI) healthcare technology innovations: the current state and challenges from a life science industry perspective. Technology Analysis & Strategic Management. 2023;35(2):179-93.

[2] Sigov A, Ratkin L, Ivanov LA, Xu LD. Emerging enabling technologies for industry 4.0 and beyond. Information Systems Frontiers. 2022:1-11.

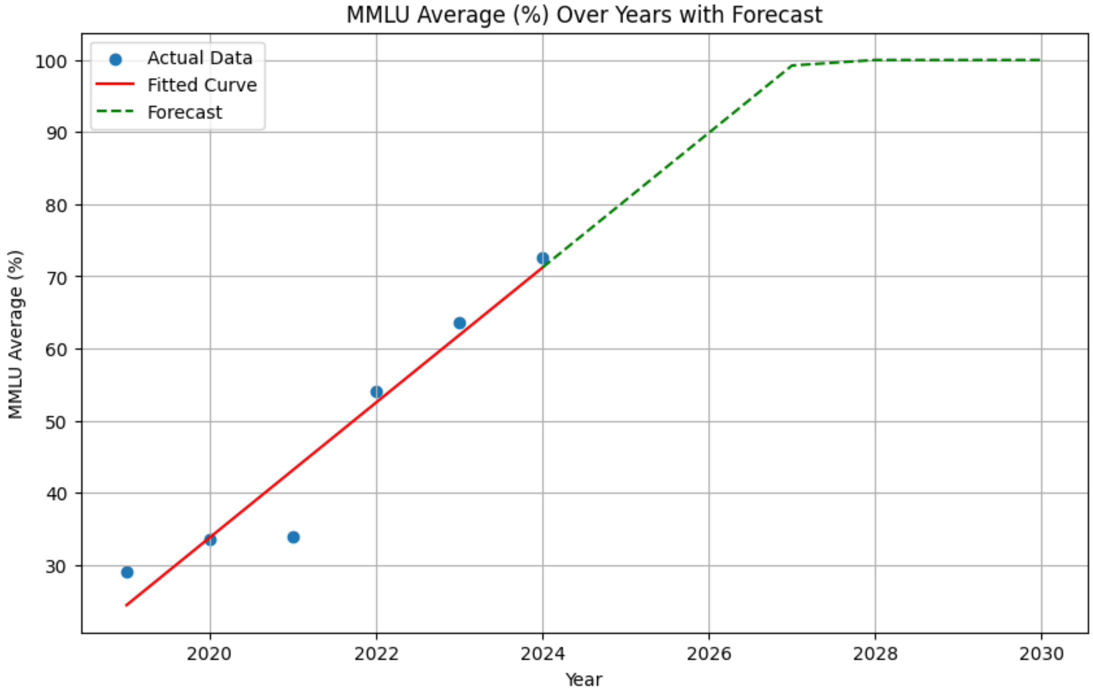


Figure 6. A forecasting of the average MMLU (Massive Multitask Language Understanding) values of the models.

[3] Kaggwa S, Eleogu TF, Okonkwo F, Farayola OA, Uwaoma PU, Akinoso A. AI in decision making: transforming business strategies. International Journal of Research and Scientific Innovation. 2024;10(12):423-44.

[4] Goralski MA, Tan TK. Artificial intelligence and sustainable development. The International Journal of Management Education. 2020;18(1):100330.

[5] Mathew D, Brintha N, Jappes JW. Artificial intelligence powered automation for industry 4.0. In: New Horizons for Industry 4.0 in Modern Business. Springer; 2023. p. 1-28.

[6] Dwivedi YK, Sharma A, Rana NP, Giannakis M, Goel P, Dutot V. Evolution of artificial intelligence research in Technological Forecasting and Social Change: Research topics, trends, and future directions. Technological Forecasting and Social Change. 2023;192:122579.

[7] Johnson S, Acemoglu D. Power and progress: Our thousand-year struggle over technology and prosperity. Hachette UK; 2023.

[8] Bibri SE, Krogstie J, Kaboli A, Alahi A. Smarter eco-cities and their leading-edge artificial intelligence of things solutions for environmental sustainability: A comprehensive systematic review. Environmental Science and Ecotechnology. 2024;19:100330.

[9] Dwivedi YK, Hughes L, Ismagilova E, Aarts G, Coombs C, Crick T, et al. Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management. 2021;57:101994.

[10] Javaid M, Haleem A, Singh RP, Suman R. Artificial intelligence applications for industry 4.0: A literature-based study. Journal of Industrial Integration and Management. 2022;7(01):83-111.

[11] Liu J, Chang H, Forrest JYL, Yang B. Influence of artificial intelligence on technological innovation: Evidence from the panel data of china's manufacturing sectors. Technological Forecasting and Social Change. 2020;158:120142.

[12] Momeni M. History of Artificial Intelligence;. Available from: https://www.kaggle.com/datasets/imtkaggleteam/history-of-artificial-intelligence.

[13] Dhamodharan B. Optimizing Industrial Operations: A Data-Driven Approach to Predictive Maintenance

through Machine Learning. International Journal of Machine Learning for Sustainable Development. 2021;3(1):31-44.

[14] Usman M, Khan R, Moinuddin M. Assessing the Impact of Artificial Intelligence Adoption on Organizational Performance in the Manufacturing Sector. Revista Espanola de Documentacion Cientifica. 2024;18(02):95-124.

[15] Wamba-Taguimdje SL, Wamba SF, Kamdjoug JRK, Wanko CET. Influence of artificial intelligence (AI) on firm performance: the business value of AI-based transformation projects. Business process management journal. 2020;26(7):1893-924.

[16] Ashta A, Herrmann H. Artificial intelligence and fintech: An overview of opportunities and risks for banking, investments, and microfinance. Strategic Change. 2021;30(3):211-22.

[17] Addy WA, Ajayi-Nifise AO, Bello BG, Tula ST, Odeyem O, Falaiye T. Algorithmic Trading and AI: A Review of Strategies and Market Impact. World Journal of Advanced Engineering Technology and Sciences. 2024;11(1):258-67.

[18] Shin B, Lowry PB. A review and theoretical explanation of the 'Cyberthreat-Intelligence (CTI) capability'that needs to be fostered in information security practitioners and how this can be accomplished. Computers & Security. 2020;92:101761.

[19] Hassan M, Aziz LAR, Andriansyah Y. The role artificial intelligence in modern banking: an exploration of AI-driven approaches for enhanced fraud prevention, risk management, and regulatory compliance. Reviews of Contemporary Business Analytics. 2023;6(1):110-32.

[20] Ryan M. The future of transportation: ethical, legal, social and economic impacts of self-driving vehicles in the year 2025. Science and engineering ethics. 2020;26(3):1185-208.

[21] Fu Y, Li C, Yu FR, Luan TH, Zhang Y. A survey of driving safety with sensing, vehicular communications, and artificial intelligence-based collision avoidance. IEEE transactions on intelligent transportation systems. 2021;23(7):6142-63.

[22] Akhtar M, Moridpour S. A review of traffic congestion prediction using artificial intelligence. Journal of Advanced Transportation. 2021;2021(1):8878011.

[23] Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Molecular diversity. 2021;25:1315-60.

[24] Guo K, Yang Z, Yu CH, Buehler MJ. Artificial intelligence and machine learning in design of mechanical materials. Materials Horizons. 2021;8(4):1153-72.

[25] Lavin A, Krakauer D, Zenil H, Gottschlich J, Mattson T, Brehmer J, et al. Simulation intelligence: Towards a new generation of scientific methods. arXiv preprint arXiv:211203235. 2021.

[26] Adler-Milstein J, Aggarwal N, Ahmed M, Castner J, Evans BJ, Gonzalez AA, et al. Meeting the moment: addressing barriers and facilitating clinical adoption of artificial intelligence in medical diagnosis. NAM perspectives. 2022;2022.

[27] Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database. 2020;2020:baaa010.

[28] Erkuş EC, Purutçuoğlu V. A new collective anomaly detection approach using pitch frequency and dissimilarity: Pitchy anomaly detection (PAD). Journal of Computational Science. 2023;72:102084.

[29] Kang MJ, Hwang YC. Exploring the factors affecting the continued usage intention of IoT-based healthcare wearable devices using the TAM model. Sustainability. 2022;14(19):12492.

[30] de Barcelos Silva A, Gomes MM, da Costa CA, da Rosa Righi R, Barbosa JLV, Pessin G, et al. Intelligent personal assistants: A systematic literature review. Expert Systems with Applications. 2020;147:113193.

[31] Martínez-Plumed F, Gómez E, Hernández-Orallo J. Futures of artificial intelligence through technology readiness levels. Telematics and Informatics. 2021;58:101525.

[32] Johri P, Khatri SK, Al-Taani AT, Sabharwal M, Suvanov S, Kumar A. Natural language processing: History, evolution, application, and future work. In: Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020. Springer; 2021. p. 365-75.

[33] Xu M, Ng WC, Lim WYB, Kang J, Xiong Z, Niyato D, et al. A full dive into realizing the edge-enabled metaverse: Visions, enabling technologies, and challenges. IEEE Communications Surveys & Tutorials. 2022;25(1):656-700.

[34] Haleem A, Javaid M, Qadri MA, Singh RP, Suman R. Artificial intelligence (AI) applications for marketing: A literature-based study. International Journal of Intelligent Networks. 2022;3:119-32.

[35] Naseem U, Razzak I, Khan SK, Prasad M. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. Transactions on Asian and Low-Resource Language Information Processing. 2021;20(5):1-35.

[36] Cao Y, Li S, Liu Y, Yan Z, Dai Y, Yu PS, et al. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. arXiv preprint arXiv:230304226. 2023.

[37] Zhou M, Duan N, Liu S, Shum HY. Progress in neural NLP: modeling, learning, and reasoning. Engineering. 2020;6(3):275-90.

[38] Kalyan KS, Rajasekharan A, Sangeetha S. Ammus: A survey of transformer-based pretrained models in natural language processing. arXiv preprint arXiv:210805542. 2021.

[39] Qi W, Yan Y, Gong Y, Liu D, Duan N, Chen J, et al. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. arXiv preprint arXiv:200104063. 2020.

[40] Vajjala S, Majumder B, Gupta A, Surana H. Practical natural language processing: a comprehensive guide to building real-world NLP systems. O'Reilly Media; 2020.

[41] Khurana D, Koli A, Khatter K, Singh S. Natural language processing: State of the art, current trends and challenges. Multimedia tools and applications. 2023;82(3):3713-44.

[42] DiPietro R, Hager GD. Deep learning: RNNs and LSTM. In: Handbook of medical image computing and computer assisted intervention. Elsevier; 2020. p. 503-19.

[43] Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: A survey. ACM computing surveys (CSUR). 2022;54(10s):1-41.

[44] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. Neurocomputing. 2021;452:48-62.

[45] Bariah L, Zhao Q, Zou H, Tian Y, Bader F, Debbah M. Large generative ai models for telecom: The next big thing? IEEE Communications Magazine. 2024.

[46] Hadi MU, Qureshi R, Shah A, Irfan M, Zafar A, Shaikh MB, et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. Authorea Preprints. 2023.

[47] Hadi MU, Qureshi R, Shah A, Irfan M, Zafar A, Shaikh MB, et al. A survey on large language models: Applications, challenges, limitations, and practical usage. Authorea Preprints. 2023.

[48] Kar AK, Varsha P, Rajan S. Unravelling the impact of generative artificial intelligence (GAI) in industrial applications: A review of scientific and grey literature. Global Journal of Flexible Systems Management. 2023;24(4):659-89.

[49] Gill SS, Xu M, Ottaviani C, Patros P, Bahsoon R, Shaghaghi A, et al. AI for next generation computing: Emerging trends and future directions. Internet of Things. 2022;19:100514.

[50] Boobier T. AI and the Future of Banking. John Wiley & Sons; 2020.

[51] Lu PJ, Lai MC, Chang JS. A survey of high-performance interconnection networks in high-performance computer systems. Electronics. 2022;11(9):1369.

[52] Katal A, Dahiya S, Choudhury T. Energy efficiency in cloud computing data centers: a survey on software technologies. Cluster Computing. 2023;26(3):1845-75.

[53] Gill SS, Wu H, Patros P, Ottaviani C, Arora P, Pujol VC, et al. Modern computing: Vision and challenges. Telematics and Informatics Reports. 2024:100116.

[54] Kavis M. Architecting the cloud. Wiley Online Library; 2023.

[55] Li Z, Wang D, Nan G, Li M. Optimal revenue model of a social networking service: Ad-sponsored, subscription-based, or hybrid? IEEE Transactions on Engineering Management. 2022.

[56] Marinescu DC. Cloud computing: theory and practice. Morgan Kaufmann; 2022.

[57] Aaker DA, Moorman C. Strategic market management. John Wiley & Sons; 2023.

[58] Johnson PC, Laurell C, Ots M, Sandström C. Digital innovation and the effects of artificial intelligence on firms' research and development–Automation or augmentation, exploration or exploitation? Technological Forecasting and Social Change. 2022;179:121636.

[59] Keiningham T, Aksoy L, Bruce HL, Cadet F, Clennell N, Hodgkinson IR, et al. Customer experience driven business model innovation. Journal of Business Research. 2020;116:431-40.

[60] Enholm IM, Papagiannidis E, Mikalef P, Krogstie J. Artificial intelligence and business value: A literature review. Information Systems Frontiers. 2022;24(5):1709-34.

[61] Tufail S, Riggs H, Tariq M, Sarwat AI. Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms. Electronics. 2023;12(8):1789.

[62] Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Translational vision science & technology. 2020;9(2):14-4.

[63] Li S, Zhao Y, Varma R, Salpekar O, Noordhuis P, Li T, et al. Pytorch distributed: Experiences on accelerating data parallel training. arXiv preprint arXiv:200615704. 2020.

[64] Xu P, Zhu X, Clifton DA. Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023.

[65] Luo Y, Yu S. H3D-Transformer: A Heterogeneous 3D (H3D) Computing Platform for Transformer Model Acceleration on Edge Devices. ACM Transactions on Design Automation of Electronic Systems. 2024.

[66] Villalobos P, Sevilla J, Besiroglu T, Heim L, Ho A, Hobbhahn M. Machine learning model sizes and the parameter gap. arXiv preprint arXiv:220702852. 2022.

[67] Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent advances in natural language processing via large pre-trained language models: A survey. ACM Computing Surveys. 2023;56(2):1-40.

[68] Reis J, Housley M. Fundamentals of Data Engineering. " O'Reilly Media, Inc."; 2022.

[69] Brunton SL, Nathan Kutz J, Manohar K, Aravkin AY, Morgansen K, Klemisch J, et al. Data-driven aerospace engineering: reframing the industry with machine learning. AIAA Journal. 2021;59(8):2820-47.

[70] Betz UA, Arora L, Assal RA, Azevedo H, Baldwin J, Becker MS, et al. Game changers in science and technology-now and beyond. Technological Forecasting and Social Change. 2023;193:122588.

[71] Cowls J, Tsamados A, Taddeo M, Floridi L. The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. Ai & Society. 2023:1-25.

[72] Díaz-Rodríguez N, Del Ser J, Coeckelbergh M, de Prado ML, Herrera-Viedma E, Herrera F. Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Information Fusion. 2023;99:101896.

[73] Khogali HO, Mekid S. The blended future of automation and AI: Examining some long-term societal and ethical impact features. Technology in Society. 2023;73:102232.

[74] Nishant R, Kennedy M, Corbett J. Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. International Journal of Information Management. 2020;53:102104.

[75] Farzaneh F, Jung S. Lifecycle carbon footprint comparison between internal combustion engine versus electric transit vehicle: A case study in the US. Journal of Cleaner Production. 2023;390:136111.

[76] Wang J, Azam W. Natural resource scarcity, fossil fuel energy consumption, and total greenhouse gas emissions in top emitting countries. Geoscience Frontiers. 2024;15(2):101757.

[77] Mikalef P, Lemmer K, Schaefer C, Ylinen M, Fjørtoft SO, Torvatn HY, et al. Examining how AI capabilities can foster organizational performance in public organizations. Government Information Quarterly. 2023;40(2):101797.

[78] Desislavov R, Martínez-Plumed F, Hernández-Orallo J. Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. Sustainable Computing: Informatics and Systems. 2023;38:100857.

AIPA's International Journal on Artificial Intelligence: Bridging Technology, Society and Policy
ISSN: 3062-097X
Published: 07 December 2024

Large Language Model-Based Autonomous Agents: Trends and Directions

Levent Dinçkal[1]*

[1]**Artificial Intelligence Policies Association (AIPA) Member, ORCID: 0009-0007-7938-8368**

ORIGINAL RESEARCH PAPER

Abstract

This paper explores the vibrant field of autonomous agents based on large language models. In recent years transformer-based large language models (LLMs) have advanced state of the art considerably in a wide range of natural language tasks and demonstrated almost human-like reasoning capabilities and world knowledge. Since autonomous agents rely on such properties, advances in LLMs have accelerated the progress in autonomous agents. This paper reviews the literature by briefly describing how LLMs work and how they can be leveraged in the overall architecture of an autonomous agent to produce significantly more capable and robust agents. Planning, memory, and action components of the autonomous agent are examined separately and a discussion of trends and future directions follows.

autonomous agents, large language models, artificial intelligence, artificial general intelligence

Keywords: autonomous agents, large language models, artificial intelligence, artificial general intelligence

1 Introduction

Large language models (LLMs) based on the Transformer architecture [1] have been extraordinarily successful in recent years. Many of these models [2] [3] [4] [5], developed and released by research and industry groups in quick succession, have rapidly pushed forward the state of the art in natural language tasks in few-shot and zero-shot settings [6] [7]. LLMs excel in solving a broad range of natural tasks, exhibiting unprecedented emergent properties, extensive world knowledge, and advanced reasoning abilities [8]. Autonomous agents [9] stand to benefit significantly from these new capabilities, enhancing their ability to understand instructions in natural language as well as to plan and act with common sense and a robust world model.

This paper explores the rapidly evolving trends in LLM-based autonomous agents and describes selected important examples from the field. We provide a brief overview of LLMs and the transformer architecture in section 2 and in section 3, we examine the specifics of LLM-based autonomous agents, focusing on essential components such as planning, memory, and action. Finally, section 4 identifies and discusses prevalent patterns in current research, offering conclusions and future directions in this swiftly developing area.

2 An Overview of the Decoder-Only Transformer Architecture

We begin with a brief overview of the large language models as the LLM-based autonomous agents are, by their very nature, tightly coupled with this transformer-based architecture. The fundamental structure of transformers was put forth in the groundbreaking "Attention is All You Need" paper [1] and has remained mostly similar to this day, with the more recent models adding tweaks to this architecture for incremental benefits. Instead of documenting each possible modification to the architecture, we will present a stylized version of a possible current model and note that LLMs from different publishers might utilize slight variations in their structure.

At a fundamental level, the large language models create a conditional probability distribution on tokens, which are words or word-like bits of text described in more detail in subsection 2.1, given the tokens previously seen in the text. The models we consider in this paper are auto-regressive, which means they operate in a sequential manner to predict the next token based only on the previous tokens. LLMs are trained with massive corpora of text and the training aims to maximize the probability of the correct token at each location in the text. In concrete terms, given a sequence of tokens $\mathbf{t} = \{t_1, \cdots, t_N\}$ in the training corpus, the training objective is to maximize the likelihood of the correct token or, equivalently, minimize

OPEN ACCESS
AIPAJ Vol:1, Issue:1

***Corresponding author**
leventdinckal@gmail.com

***Submitted** 15 June 2024*

***Accepted** 28 October 2024*

Citation
Dinçkal, L. (2024). Large Language Model-Based Autonomous Agents: Trends and Directions. In AIPA's International Journal on Artificial Intelligence: Bridging Technology, Society and Policy (Vol. 1, Number 1, pp. 13–24). DOI: 10.5281/zenodo.14293261

$$L(\mathbf{t}) = -\sum_{i=1}^{N} \log[P(t_i|t_{i-1}, \cdots, t_{i-k}; \Omega)] \tag{1}$$

where $L(\mathbf{t})$ is the cross-entropy loss over the given sequence of tokens, k is the number of previous tokens we consider (sometimes called the context window length), and Ω is the set of trainable parameters of the model. This objective function will force the randomly initialized model to mimic its input texts after some training as the Kullback-Leibler divergence between the distributions of the training texts and predicted tokens will be minimized.

The architecture of the large language model we will consider can be seen in Figure 1. Note that this is a decoder-only model rather than the fundamental encoder-decoder structure proposed in [1] as the agents we are interested in will use decoder-only models. We now briefly review the components of this architecture separately in the order of their appearance in the forward pass of the model.

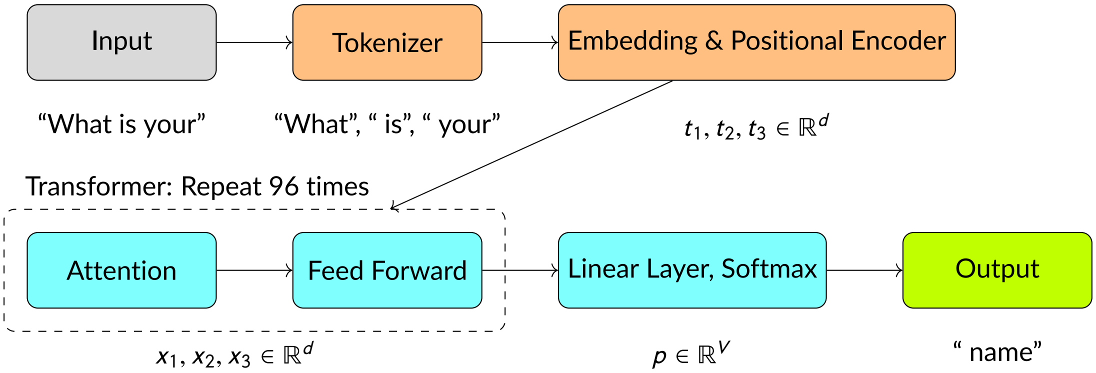


Figure 1. Stylized Structure of a Large Language Model.

2.1 Tokenization and Embedding

Sequences of text need to be tokenized and embedded into a vector space before they can be consumed by language models. A tokenizer is a deterministic program that uses a heuristic rule to break down text into chunks called tokens. These tokens, usually words and subwords, are the inputs into the model and the final output is a conditional probability distribution on the tokens. Two of the most commonly used tokenizers in the currently state-of-the-art models are Wordpiece [10] and Byte Pair Encoding [11]. They have a similar working principle in that they both break down the text into characters and then progressively merge contiguous chunks. The difference is that Byte Pair Encoding (BPE) merges chunks depending on the frequency of the merged text in the corpus, whereas Wordpiece, in addition to BPE's criterion, gives priority to merging chunks that are infrequent on their own. The examples below illustrate the tokenization of two random phrases by the BPE tokenizer of OpenAI's GPT-4 [5] as implemented by the tiktoken library [12].

Example Go to the store and buy groceries

Example high-rise buildings and backfilled excavation sites

The tokenizer indexes each possible token in the created vocabulary with an integer value and for any given text, this sequence of integers are sent into the embedding layer. Unlike the tokenizer, the embedding layer is not deterministic and usually needs to be trained along with the rest of the model. Embedding in NLP refers to the process of projecting each discrete token-indexing integer into a high-dimensional vector space. Some of the first papers on embedding in NLP were Word2Vec [13] and GloVe [13], which showed how to embed words into high-dimensional vector spaces in such a way that their distances and linear transformations in that space are semantically sensible. While these earlier algorithms considered vector spaces of dimension sizes in the hundreds, the current breed of models tend to embed tokens to much higher dimensional spaces in order to capture richer meaning, with LLaMA-2 [4], for instance, using an embedding dimension of 4096.

While the token vector embeddings capture the meanings of the individual words, the positions of tokens also hold semantic and syntactic value. The attention block itself has no way to account for the ordering of the tokens. Therefore, we need to feed information about the order to the rest of the model with another type of embedding called position encoding. [1] uses different frequencies of sine and cosine functions to derive positional encodings. A popular alternative is Rotary Position Embedding (RoPe) as proposed by [14] and used in prominent models such as LLaMA-2. Regardless of the method used, the positional encoding is usually a vector that is the same size as the semantic embedding vector described previously, which allows the two vectors to be simply summed up to create one vector per token as an input to the attention block.

2.2 The Transformer Block

The attention mechanism described in [1] is still commonly used in LLMs today and it is arguably the most important part of the model as it can take credit for most of the impressive features of large language models. The type of attention that we consider in this paper falls into the category of self-attention, which means it attends only to its input sequence as opposed to other attention types that can also attend to their output sequences.

Before launching into a description of the core of the transformer block, we note that most models include a normalization step at the end or, more commonly in the recent models, at the beginning of the attention mechanism. One popular normalization method is Layer Normalization [15], which scales and shifts its input to match a distribution easier to process. These normalization components provide numerical stability during training and help avoid the usual problems of vanishing or exploding gradients.

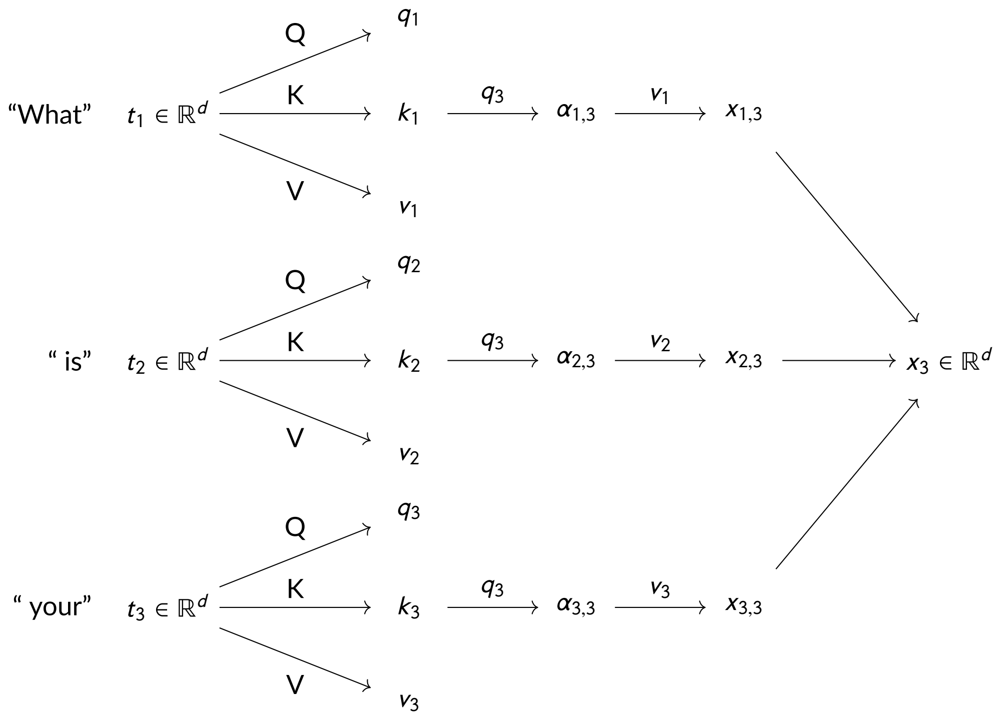


Figure 2. Inner Working of the Attention Mechanism.

The fundamental function of the attention layer is to calculate the weight which a given token will assign to the information content of every other token. Figure 2 gives an overview of the mechanism. Given the token embedding dimension size d and assuming the attention mechanism also uses dimension d, the model learns the weight matrices $Q \in \mathbb{R}^{d \times d}$, $K \in \mathbb{R}^{d \times d}$, and $V \in \mathbb{R}^{d \times d}$, which are used to project the incoming token embedding vectors into query, key, value spaces respectively. Thus, at every prediction step of the model, we calculate for every token $t_i \in \mathbb{R}^d$ in the context window

$$\begin{aligned} q_i &= Qt_i \\ k_i &= Kt_i \\ v_i &= Vt_i \end{aligned} \tag{2}$$

which are query, key, and value vectors for the token respectively. Query vectors are used to query the relevance of the token to every other token. The key vector is the answer to queries from other tokens to determine similarity and the value vector holds the information content of the token. In practice, we keep the key and value vectors in a cache instead of re-calculating them at every step and further optimizing the memory footprint of that sizable cache has been a persistent research target recently as in [16]. At the end of the attention layer, we want to calculate a weighted sum of all value vectors for every token in the context window, where the weights are the attention weights given to the corresponding tokens. For a given token t_i, we first calculate attention scores a_{ij} between t_i and all other tokens t_j in the context window:

$$a_{ij} = \frac{q_i \cdot k_j}{\sqrt{d}} \tag{3}$$

The inner product is used here as a measure of similarity as two vectors will have a greater inner product if they are more closely aligned in the vector space. Thus, the attention score will be greater if the query vector of one token is similar to the key vector of the other. The scaling expression in the denominator is helpful because the inner product yields very large values as the dimension size increases and large values yield very small gradients during training. Once we have the attention scores, we obtain the attention weights with the softmax function:

$$\alpha_{ij} = \text{softmax}(a_{ij}) \tag{4}$$

and the output vector x_i for the token is simply

$$x_i = \sum_j \alpha_{ij} v_j \tag{5}$$

The process described here is applicable in the case of a single attention head. Many models make use of multi-head attention, which refers to multiple such attention mechanisms working as described but independently of each other. It is common for a multi-head attention model with h heads to use Q, K, V matrices in $\mathbb{R}^{d_k \times d}$ such that $d_k = d/h$, in which case the output vector from each head can be simply concatenated to produce a d-dimensional output vector.

The second part in the transformer block is a relatively straightforward feed-forward network (FFN), which has two linear transformations separated by a non-linear activation function such as ReLU:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{6}$$

where W_1 and W_2 are trainable weight matrices and b_1 and b_2 are trainable bias terms. It is common to regularize the FFN using the dropout method [17]. Many models set the dimensions of W_1 and W_2 such that the input vector is mapped to a higher dimension by the first transformation and then reduced to the original dimension by the second one. This allows a richer semantic representation inside the fully connected component. Intuitively, the attention mechanism retrieves all the relevant information from the context and the feed-forward network analyzes the retrieved information in depth with the higher dimension mapping and non-linearity enhancing its expressive power.

Most models also make use of residual connections in between the attention and FFN components as well as after the FFN. These connections add the input of a layer to the output of the same layer. Residual connections provide a more direct connection from the trainable parameters appearing earlier in the model to the parts closer to the final prediction. This provides better gradient updates to the parameters near the beginning during training and prevents situations in which these parameters are not learned adequately due to extremely small gradients.

Large language models run multiple transformer blocks in succession of each other in order to process data more thoroughly. For instance, the LLaMA-2 model has versions with 32, 40, and 80 transformer layers [4] and one of the larger GPT-3 [2] versions has 96 layers.

2.3 Token Prediction

The last step in the large language model comes after the series of transformer blocks. With the last transformer block yielding one enriched context vector x_i for every token t_i in the input sequence, this last step maps the vectors back into the vocabulary space and uses the softmax function to create a distribution of probabilities. Thus, for i-th position, the predicted probability vector is

$$p = \text{softmax}(W_0.x_i + b_0) \tag{7}$$

where $x_i \in \mathbb{R}^d$ is the input context vector for the token from the last transformer block, $W_O \in \mathbb{R}^{V \times d}$ and $b_O \in \mathbb{R}^V$ are trainable parameters and V is the vocabulary size. The most obvious and common way to pick a token using the output probability distribution is to pick the token corresponding to the highest predicted probability. Alternative methods exist such as sampling from the multinomial distribution implied by p after filtering down only to the top k most likely tokens, called top-k sampling. [18] give a good overview of the sampling alternatives.

3 LLM-Based Autonomous Agents

Large language models have become remarkably popular in recent years in some part because they opened up the state of the art in artificial intelligence to a general audience in the user-friendly form of chatbots. There is a growing body of research which shows that, in addition to their usefulness as chatbots, the large language models can enable breakthroughs in the field of autonomous agents due to their extraordinarily strong reasoning abilities. For the purposes of this paper, an autonomous agent is defined as a system that interacts with its environment to solve a given task. The task can be quite complicated with multiple steps and the autonomous agent has to make its planning and pursue its agenda without any assistance from humans. The older autonomous agents relied on relatively simple rules and heuristics in small sandbox environments, often borrowing from traditional control theory. [9] provides a good overview of what they refer to as intelligent agents. Later on, deep learning opened up new possibilities and techniques using deep reinforcement learning, in particular, such as [19] and [20] made marked improvements on earlier methods even though they still suffered from the problem of being confined to relatively small world models. The newer LLM-based agents have innovated heavily on traditional methods and show better promise in handling complex real-world tasks and possibly growing into AGI systems in the future.

3.1 Examples of LLM-Based Autonomous Agents

AutoGPT [21] and BabyAGI [22] were two of the first LLM-based autonomous agents that were released shortly after ChatGPT. Both of these systems implemented the idea of running OpenAI's GPT model [2] multiple times in order to break down complex tasks into parts and work on the parts in an iterative fashion. While AutoGPT focuses on handling practical tasks and has the ability to navigate the Internet and perform various external actions, BabyAGI focuses more on learning, memory, sequential decision-making, and efficient task management in an effort to put forth a blueprint for a possible future AGI. A similarly early agent is HuggingGPT [23], which uses ChatGPT for planning and matching user requests with models available on the Hugging Face platform based on model descriptions in order to combine capabilities across a diversity of expertise areas and modalities such as text, vision, and speech. ViperGPT [24] is another take on multimodality, where the agent, given an image and a related question, generates Python code leveraging vision models to answer the question.

Agents focused on the software engineering process have particularly attracted a lot of attention. GPT-Engineer [25] is an agent that takes a brief description of the desired software product, asks a few clarifying questions, and builds a complete project consisting of multiple files possibly written in different languages. It also has the ability to take a sketch of the desired UI as an input image and recreate the UI in code. ChatDev [26] and MetaGPT [27] make use of multi-agent frameworks to create virtual models of complete software companies. A multi-agent framework contains multiple autonomous agents, each customized for a particular role with different priorities, capabilities, and targets. ChatDev and MetaGPT create agents with the roles of developers, QA engineers, designers, architects, and project managers, whose interactive collaboration produces the desired software program.

Games provide a relatively cheap playground in which to experiment with autonomous agents. Generative Agents [28] is a very interesting simulation of human interaction in the form of a small community of human-like agents living in a simulation environment similar to The Sims. The agents in the community have different identities and memories. They are able to remember persistent information about themselves and others, remember what happened in the past in the simulation, keep to their schedules, plan, and interact with other agents in a way that causes emergent social structure. Another example is Meta's Cicero agent [29] that plays Diplomacy, which is a turn-based Risk-like game of conquest that involves strategic calculations and delicate negotiations with other players in natural language. Cicero models what every other player is likely planning, computes its own optimal plan and uses large language models to bargain with, or manipulate, other players. Thus, it serves as an excellent example of combining other machine learning capabilities with natural language models.

Embodied agents, which are autonomous agents attached to either a physical or simulated body, have also benefited from the recent wave of integration with large language models. The DEPS framework [30] works on the problem of multi-step reasoning for long-term tasks in the Minecraft world by introducing a refinement of planning with descriptions and iterative feedback on errors. Voyager [31], on the other hand, creates embodied agents in the Minecraft world that emphasize long-term learning, development, maximization of exploration, and, notably, storage of complex behavior in libraries of executable code. SayCan [32] is an application of LLM-based agents in the world of robotics, in which rich world knowledge and semantic capabilities of LLMs are combined with the low-level physical skills of a robotic system, enabling a robotic arm to follow long-term human instructions specified in natural language. In summary, transition from simulated agents to the physical world is likely to enable a huge spectrum of potential uses for artificial intelligence systems in the near future. Especially the autonomous agents that utilize the growing capabilities of multimodal models that integrate language capabilities with other modalities such as vision and sound hold great potential for robotics.

3.2 Overall Architecture

The examples described in subsection 3.1 come with significantly varying designs, but a general architecture for a representative LLM-based autonomous agent can be seen in Figure 3. The indicated modules for Planner, Memory, and Action are the capabilities that the agent is able to draw upon when performing its tasks. The planner module is particularly important as it is the entry point for the high-level task set by a human and specified in natural language. This module decides how the task will be handled and which of the other modules to call on if needed. This module benefits significantly from the real world information and reasoning capability embedded in large language models. The memory module holds information, either persistent or produced during the agent's run, that is not contained in the LLM used by the planner. The action module serves as an actuator, acting on the environment, possibly using external tools, and returning the result of the action to the planner. It can be observed that the architecture of the autonomous agent is somewhat reminiscent of the human brain and the specialization of modules for different functionality in pursuit of efficiency is akin to different sectors of the brain having the same biological material but specializing in different tasks.

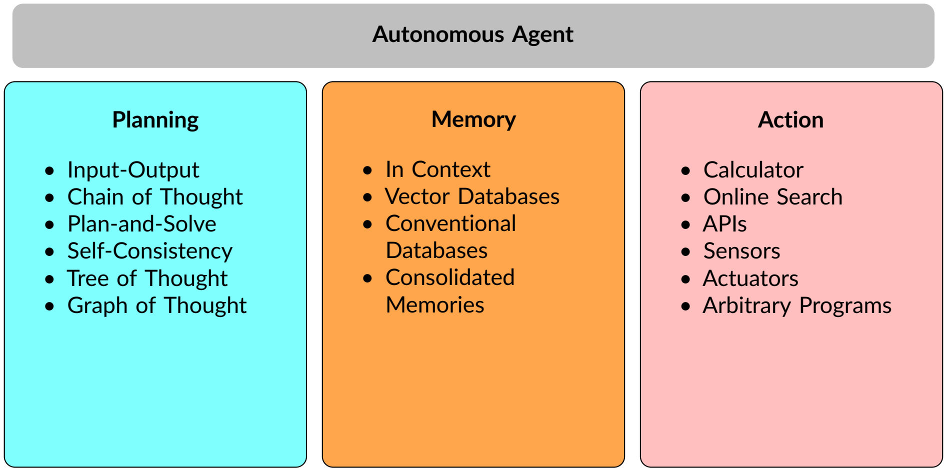


Figure 3. Components of a Representative Autonomous Agent.

In the remainder of this section, we discuss the major components of the architecture separately in greater detail.

3.3 Planning

Planning is a fundamental task for autonomous agents and it relies heavily on a human-like reasoning capability to navigate the complex landscape of real-world tasks. This aspect of the agent has significantly benefited from the advent of large language models. LLMs excel in ingesting instructions specified in natural language, modeling rich world scenarios, and exhibiting advanced reasoning capabilities, making them ideally suited for planning tasks for an autonomous agent. The primary function of the planning module is to manage the agent by understanding human-supplied instructions, decomposing the high-level goal into smaller subtasks, and utilizing other modules of the agent as necessary. Depending on the complexity of the assigned tasks, this module can range from simple in structure to quite sophisticated.

The simplest possible way of planning is the Input-Output method, which is merely posing the question to an LLM and getting an answer immediately without any intermediate steps. This might be viewed as a normal usage of an LLM rather than an autonomous agent. A marked improvement on this approach is the Chain of Thought (CoT)[6], which significantly increases the correctness of answers to complex queries and provides the breakdown of complex queries that is useful for an autonomous agent. This method works by adding to the prompt a directive to break down the task into smaller subtasks such as “Let’s think step by step.” and a few worked-out examples of a complex query being decomposed into intermediate steps and solved sequentially. An intuitive understanding of why this helps the LLM might come from the observation that LLMs are designed to produce each token with the same amount of computation. An easier question can be answered with smaller amounts of computation, but a complex query requires the larger computational space that CoT provides.

Some notable variations on the CoT idea are zero-shot CoT [7], which omits the worked-out examples from the prompt with no large penalty, and Plan-and-Solve prompting [33], which records gains in math and common sense questions by requiring the LLM to make the plan first before answering. A more sophisticated approach is Self-Consistency [34] (CoT-SC). This method uses the temperature parameter of the LLM to produce multiple different plans and then chooses steps that are supported by the most plans in a majority vote setting. This kind of planning ends up being more creative and takes into account the fact that multiple plans might all have valid insights into the problem.

A major improvement over the Chain of Thought was the Tree of Thoughts (ToT) [35]. The ToT method builds a tree in which the root node is problem statement and each branching node is a "thought" relating to its parent node. The generated tree is then pruned by a process which might judge a particular subtree as unpromising due to its children nodes diverging too far away from the desired result. This gives the ToT the crucial ability of backtracking from wrong paths without losing good progress already made. As a point of comparison, the Self-Consistency method also produces multiple plans but might discard plans entirely if their later stages go down a wrong path. This more iterative approach enables significant gains for ToT in problem-solving abilities and real-world applications such as robotics. A further generalization over Tree of Thought is the Graph of Thought (GoT) [36]. GoT is very similar to ToT, but it has a few key improvements, such as the inclusion of a self-refining process at every step and modeling thoughts as a general graph rather than a tree. A node in a graph might have multiple parents and GoT takes particular care to merge branches in different parts of the graph. The intuitive idea behind this method is that multiple paths of thought might lead to the same conclusions or further paths of thought and this reusability helps in efficiency and correctness. An even further step in this direction is the Hypergraph of Thoughts [37]. A hypergraph is a generalization of the concept of graphs, in which an edge can join an arbitrary number of vertices. Such a structure can potentially become useful especially in a multimodal context. Selected planning methods are summarized in Figure 4.

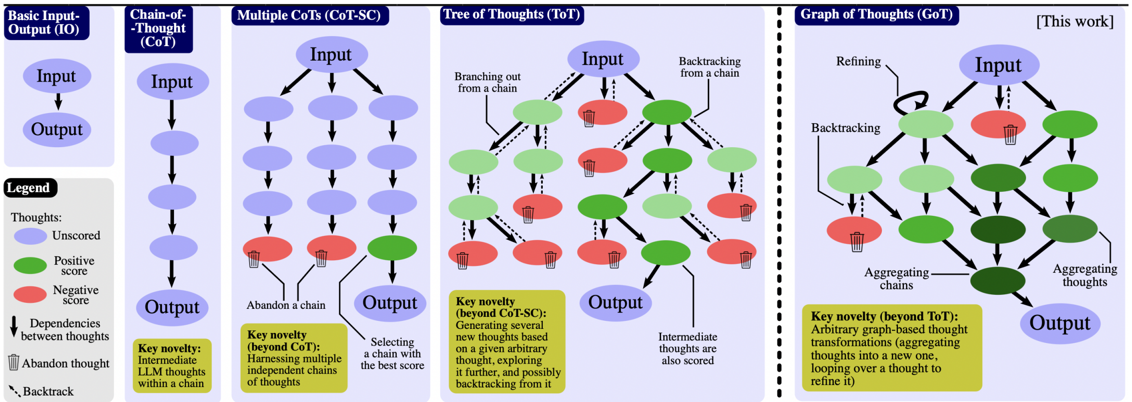

Basic Input-Output (IO)
Input
Output
Legend
Thoughts:
Unscored
Positive score
Negative score
Dependencies between thoughts
Abandon thought
Backtrack
Chain-of--Thought (CoT)
Input
Output
Key novelty: Intermediate LLM thoughts within a chain
Multiple CoTs (CoT-SC)
Input
Abandon a chain
Output
Key novelty (beyond CoT): Harnessing multiple independent chains of thoughts
Selecting a chain with the best score
Tree of Thoughts (ToT)
Input
Branching out from a chain
Backtracking from a chain
Output
Key novelty (beyond CoT-SC): Generating several new thoughts based on a given arbitrary thought, exploring it further, and possibly backtracking from it
Intermediate thoughts are also scored
Graph of Thoughts (GoT)
[This work]
Refining
Input
Backtracking
Aggregating chains
Aggregating thoughts
Output
Key novelty (beyond ToT): Arbitrary graph-based thought transformations (aggregating thoughts into a new one, looping over a thought to refine it)

Figure 4. Different Planning Frameworks. Image reproduced with permission from Besta, Blach et al. (2024) [36].

A crucial aspect of autonomous agents is their ability to use external tools as portrayed in Figure 3. The planner of the autonomous agent must, therefore, be aware of the existence and purpose of the available tools and be ready to use them if appropriate. One popular way of integrating external tools with a CoT agent is the ReAct framework [38]. ReAct, which stands for reason and action, interleaves reasoning and acting steps, thereby giving the agent ability to evaluate the results of actions and plan further at each step. This method uses directives and examples in the prompt to make LLM output Thought - Action - Observation steps at each stage. The Thought part reasons and identifies an action to take. Action executes the action and Observation is an evaluation of the result of the action taken. Thought and Observation steps are generated by the LLM whereas Action is executed externally. In this framework the agent chooses external tools based on the context in the prompt, which includes a brief description of every tool as well as instructions for utilization. This iterative method of using tools and making new plans based on results provides a significant capability boost to autonomous agents.

3.4 Memory

The memory component may contain records of external information or recollections of the history of the agent and, therefore, is an important part of the autonomous agent. Efficiently organized memory structures allow an agent to “remember” past observations and actions as well as any type of knowledge not already included in the base large language model. This section reviews the different ways of storing and accessing memories.

An easy way of integrating memory is to inject what needs to be remembered into the prompt of the LLM. This can be viewed as a short-term memory as the information in the prompt is immediately available to the model and is not persisted in another storage tool. While convenient, this method is restricted by the relatively narrow context window of the currently available large language models. Continuing the metaphor of human memory, we can consider as long-term memory the use of external storage. Some examples of storage types might be conventional and NoSQL databases or vector databases that hold embedding vectors for bits of text. The use of vector databases to hold information is sometimes called Retrieval Augmented Generation (RAG) [39]. In this framework, the agent has to find the memory most relevant to the task at hand and retrieve it as natural language to be added to the prompt of the LLM in the next step. Finding the most relevant memory is a problem remarkably similar to the task of attention mechanism described in subsection 2.2, which computes how much attention to pay each token in the context based on the input. The solution is likewise similar. For an input query q and a memory set M, all produced with the same embedder, the most relevant memory for the given query is

$$m^* = \arg\max_{m \in M} q \cdot m \tag{8}$$

where $\cdot$ is the inner product operator. Most vector databases use efficient Maximum Inner Product Search algorithms that run such optimizations with reasonable algorithmic complexity.

It is often useful to reflect on memories and store the result of the reflection rather than raw memories themselves. One advantage of such reflection is that the summarized information can be much smaller, making storage and later retrieval much cheaper. Another advantage is that the model could glean valuable high-level insights from raw observations that could not be observed by themselves. One such memory structure is demonstrated in Generative Agents [28], described in subsection 3.1, which creates a community of individual agents living in a virtual town, all with their own personality and memory. Every agent in this world takes time to reflect on what it observes and synthesizes a consolidated higher-level thought, called a reflection. To give a concrete example, if agent A observes agent B reading books about architecture, agent A might reflect that “Agent B is interested in architecture.” Such consolidation creates more useful information for the agent as well as reducing the cost of storage significantly. Another noteworthy innovation is the Reflexion framework [40], which uses self-reflection to gain an understanding of its own past errors and improvement opportunities and can store its conclusions in a long-term database. Such input allows it to make better plans in the future, enabling steady long-term improvements for the agent.

3.5 Action

The ability to act on or get information from the external world is one of the most useful aspects of autonomous agents in real world settings. The set of available actions might be baked into the construction of the agent, such as a robotic arm, or it might include the use of arbitrary computer programs as tools. Just as the usage of tools is extremely valuable for humans, it also unlocks lots of new capabilities for an autonomous agent. The action module is usually called once the planner of the agent decides to perform an action or use a tool with a process such as ReAct [38], described in subsection 3.3, and forwards the context to the action component. In addition to performing the action, this module is also responsible for the generation of action inputs, which might be commands or parameters that most tools require. For example, given a decision to use an Internet search engine and the context, the agent needs to generate the search query to be passed to the search engine.

The set of available actions is usually given to the agent as part of the initial prompt and can include a wide variety of prepared functionality. One action might be to simply use an LLM, which can be either the same one as the planner or another one specialized for certain areas of expertise such as law or medicine. HuggingGPT [23], described in subsection 3.1, is an example in which the agent chooses a model to use from the online repository of Hugging Face. Another popular external tool family is mathematical software. Since mathematical operations are not the strong suite of large language models by their nature [41], an agent can, for instance, write down a complex arithmetic expression and pass it to a calculator. TPTU [42] builds on this idea and creates a framework in which the agent can compose and execute Python code to compute answers to mathematical problems. Another family of tools that have seen widespread usage are APIs, or Application Programming Interfaces. These APIs can have arbitrary functionality, including supply of information or acting in various ways on the world. Toolformer [41] and API-Bank [43] are two noteworthy systems that use LLMs to pick which APIs to use from a large set of available APIs and formulate valid queries to the chosen API. Since the action module is responsible for most of the external effects of an autonomous agent, care should be taken in picking which functionality is made automatically available to the agent without human oversight. While getting weather information or modifying a database can be relatively safe, it is still a touch too early to give an agent the ability to launch nuclear weapons.

4 Discussion and Conclusion

The previous sections described large language models and trends in autonomous agents in some detail. This section will conclude by pointing out some prevalent themes in the research and discussing some possible future trends in the field.

A major theme in the research so far has been the great value of utilizing LLMs relatively frequently during the operation of an autonomous agent. The use of an LLM is analogous to thinking for the agents and if the agent “thinks” at each step of the plan, after the output of an action, and frequently at critical points, the end result seems to improve significantly. Such frequent use of LLMs might also be an antidote to the serious problem of error propagation, which refers to the agent making an error at a certain stage and having its subsequent steps steadily diverge from useful paths [41]. The obvious downside to using LLMs at every opportunity is that they are still relatively expensive to run. Error propagation and various other modes of failure during the operation of the agent, such as problems interfacing with external tools due to wrong parameters or invalid JSON, also depend on the quality and correctness of LLMs. As better and cheaper are created over time, autonomous agents will become much more robust and new avenues of planning and executing will be unlocked.

Another theme with most deep learning research so far and especially prominent in autonomous agents has been the obvious analogies with the human brain. The fundamental deep learning building blocks such as artificial neural networks, the activation functions therein, and attention are inspired by biological constructs. Additionally, the overall architecture of the agent has a lot of similarities with a model of the human brain. Planning, for example, can be seen as standing for the prefrontal cortex and long term memory reminds one of the hippocampus. The analogy will likely persist and future innovations will draw heavily from the human brain functionality that is so far missing from AI systems. Some candidates are creativity, habit formation, spatial reasoning, and sensory information. Multimodal agents, in particular, seem like one of the nearest innovations and they will enable systems that can more easily interact with the environment with the help of a broad range of sensors.

As autonomous agents have evolved from experimental constructs to systems with real value to users, we are likely to see a widespread use in some commercial areas such as customer support. This will create opportunities for agents to reflect on and learn from previous experience, using frameworks like Reflexion [40]. Such real-world feedback might accelerate the improvement of agents considerably. Even more complex scenarios might arise when cooperation with humans and other autonomous agents with different specializations is allowed. Naturally, such working models would also necessitate guarding against risks as autonomous agents might interact with each other in unpredictable ways and humans can provide adversarial input. Particular attention should be paid to the use of external tools as this part of the autonomous agents has the potential to cause most harm. Human oversight is likely to remain necessary for any action that has the potential to cause undesirable effects.

This paper described large language models and how autonomous agents leverage these models to undertake complex tasks with enhanced capability and robustness. The analogy between the autonomous agents described and the human brain is apparent and many applications are on the road to approach human-like performance even if in limited settings. Thus, LLM-based autonomous agents constitute a seemingly viable path to AI systems with high real-world value and, eventually, artificial general intelligence.

References

[1] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

[2] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.

[3] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research. 2023;24(240):1-113.

[4] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288. 2023.

[5] Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv preprint arXiv:230308774. 2023.

[6] Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems. 2022;35:24824-37.

[7] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. Advances in neural information processing systems. 2022;35:22199-213.

[8] Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv preprint arXiv:220607682. 2022.

[9] Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach. 3rd ed. Prentice Hall; 2009.

[10] Schuster M, Nakajima K. Japanese and Korean Voice Search. In: International Conference on Acoustics, Speech and Signal Processing; 2012. p. 5149-52.

[11] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. arXiv preprint arXiv:150807909. 2015.

[12] OpenAI. tiktoken is a fast BPE tokeniser for use with OpenAI's models; 2024. Version 0.7.0. https://pypi.org/project/tiktoken/.

[13] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013.

[14] Su J, Ahmed M, Lu Y, Pan S, Bo W, Liu Y. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing. 2024;568:127063.

[15] Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv preprint arXiv:160706450. 2016.

[16] Brandon W, Mishra M, Nrusimha A, Panda R, Kelly JR. Reducing Transformer Key-Value Cache Size with Cross-Layer Attention. arXiv preprint arXiv:240512981. 2024.

[17] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 2014;15(56):1929-58. Available from: http://jmlr.org/papers/v15/srivastava14a.html.

[18] Nadeem M, He T, Cho K, Glass J. A systematic characterization of sampling algorithms for open-ended language generation. arXiv preprint arXiv:200907243. 2020.

[19] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature. 2015;518:529-33. Available from: https://api.semanticscholar.org/CorpusID:205242740.

[20] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. arXiv preprint arXiv:150902971. 2015.

[21] Gravitas S. AutoGPT; 2024. Accessed: 2024-06-15. https://github.com/Significant-Gravitas/AutoGPT. Available from: https://agpt.co.

[22] Nakajima Y. BabyAGI; 2024. Accessed: 2024-06-15. https://github.com/yoheinakajima/babyagi.

[23] Shen Y, Song K, Tan X, Li D, Lu W, Zhuang Y. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. Advances in Neural Information Processing Systems. 2024;36.

[24] Surís D, Menon S, Vondrick C. Vipergpt: Visual inference via python execution for reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023. p. 11888-98.

[25] Osika A. gpt-engineer; 2023. Accessed: 2024-06-15. https://github.com/gpt-engineer-org/gpt-engineer. Available from: https://gpt-engineer.readthedocs.io.

[26] Qian C, Cong X, Yang C, Chen W, Su Y, Xu J, et al. Communicative agents for software development. arXiv preprint arXiv:230707924. 2023.

[27] Hong S, Zhuge M, Chen J, Zheng X, Cheng Y, Zhang C, et al.. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework; 2023.

[28] Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology; 2023. p. 1-22.

[29] (FAIR)† MFARDT, Bakhtin A, Brown N, Dinan E, Farina G, Flaherty C, et al. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. Science. 2022;378(6624):1067-74. Available from: https://www.science.org/doi/abs/10.1126/science.ade9097.

[30] Wang Z, Cai S, Chen G, Liu A, Ma X, Liang Y. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. arXiv preprint arXiv:230201560. 2023.

[31] Wang G, Xie Y, Jiang Y, Mandlekar A, Xiao C, Zhu Y, et al. Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv preprint arXiv: Arxiv-230516291. 2023.

[32] Ahn M, Brohan A, Brown N, Chebotar Y, Cortes O, David B, et al. Do As I Can and Not As I Say: Grounding Language in Robotic Affordances. In: arXiv preprint arXiv:2204.01691; 2022. .

[33] Wang L, Xu W, Lan Y, Hu Z, Lan Y, Lee RKW, et al. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. arXiv preprint arXiv:230504091. 2023.

[34] Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:220311171. 2022.

[35] Yao S, Yu D, Zhao J, Shafran I, Griffiths T, Cao Y, et al. Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems. 2024;36.

[36] Besta M, Blach N, Kubicek A, Gerstenberger R, Podstawski M, Gianinazzi L, et al. Graph of thoughts: Solving elaborate problems with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38; 2024. p. 17682-90.

[37] Yao F, Tian C, Liu J, Zhang Z, Liu Q, Jin L, et al. Thinking like an expert: Multimodal hypergraph-of-thought (hot) reasoning to boost foundation modals. arXiv preprint arXiv:230806207. 2023.

[38] Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, et al. React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:221003629. 2022.

[39] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems. 2020;33:9459-74.

[40] Shinn N, Labash B, Gopinath A. Reflexion: an autonomous agent with dynamic memory and self-reflection. arXiv preprint arXiv:230311366. 2023.

[41] Schick T, Dwivedi-Yu J, Dessì R, Raileanu R, Lomeli M, Hambro E, et al. Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems. 2024;36.

[42] Ruan J, Chen Y, Zhang B, Xu Z, Bao T, Du G, et al. Tptu: Task planning and tool usage of large language model-based ai agents. arXiv preprint arXiv:230803427. 2023.

[43] Li M, Song F, Yu B, Yu H, Li Z, Huang F, et al. Api-bank: A benchmark for tool-augmented llms. arXiv preprint arXiv:230408244. 2023.

AIPA's International Journal on Artificial Intelligence: Bridging Technology, Society and Policy
ISSN: 3062-097X
Published: 07 December 2024

Artificial Intelligence Policy And Ethics: A Bibliometric Review And Future Research Directions

Nesibe Manav Mutlu[1*]

[1]**Department of Management Information Systems, İstanbul Nişantaşı University, İstanbul, TÜRKİYE, ORCID ID: 0000-0002-7853-6337**

ORIGINAL RESEARCH PAPER

Abstract

This study provides a comprehensive bibliometric analysis to analyze the academic studies conducted so far in the field of artificial intelligence (AI) policies and ethics. Using Scopus as a data source, the most cited core publications in these fields and the changing trends over time were identified. Collaborations and keywords between researchers and institutions were visualized through VOSviewer and a Python pyBibX library. The analysis reveals the importance of AI policies and ethics in academic studies, as well as the central role played by some countries in research. It attempts to identify a direction for future research by following the changing trends over time.

Keywords: Bibliometric analysis, Artificial intelligence, Policy, Ethics, Research trends, Collaboration

1 Introduction

Over the last two decades, significant technological advances driven by impressive breakthroughs in both software and hardware have reshaped our world. It has evolved from the Steam Age powered by steam engines to the Electric Age powered by generators to the revolutionary Information Age powered by computers. Artificial Intelligence (AI) has become indispensable in today's technology and is seen as a cornerstone for the technology of the future. Considering its various definitions, AI represents the study and use of theories, methods, technologies, and applications that aim to simulate, enhance, and extend human intelligence [1]. AI includes various methods that allow machines to mimic human intelligence and perform tasks that typically require human thinking [2]. AI has become a tool that can automatically perform work tasks and help make difficult decisions in various businesses and fields [3].

The importance of AI lies in its ability to enhance human capabilities and optimize business processes, which translates into improved efficiency and innovation [4], [5]. Organizations can use AI to leverage large amounts of data to increase their efficiency, which allows organizations to derive important insights, make decisions based on data-driven results, and predict future trends using AI. Moreover, AI has the potential to completely transform industries like healthcare, banking, manufacturing, education, and transportation by providing innovative solutions to complex problems and helping improve the overall standard of living [6].

AI is used in various fields such as natural language processing, computer vision, robotics, and data analytics [7]. AI enables machines to understand human language in the field of natural language processing and respond accordingly, making it possible to use virtual assistants and language translation systems [8, 9]. Computer vision uses AI algorithms to facilitate the interpretation and analysis of visual input, leading to improvements in facial recognition, object identification, and autonomous vehicles [10, 11, 12]. Moreover, the use of AI-enabled robotics has the ability to completely transform various industries by facilitating the creation of autonomous machines that can perform complex tasks with accuracy and efficiency [13, 14].

As AI technologies advance, policy and ethical implications are increasingly salient [15, 16]. Addressing concerns about privacy, bias, accountability, and transparency is critical to ensuring that AI systems are appropriately developed and implemented [17]. It is essential that policymakers and researchers work together to create systems that provide a balance between AI advances and ethical concerns [18]. This will help limit potential negative outcomes while ensuring that AI technologies have a positive impact for and on society [19].

OPEN ACCESS
AIPAJ Vol:1, Issue:1

***Corresponding author**
nesibe.manavm@gmail.com

***Submitted** 10 Agust 2024*

***Accepted** 01 October 2024*

Citation
First and Second (2020) Your Paper Title. AIPA's International Journal on AI: Bridging Technology, Society and Policy. DOI: 10.5281/zenodo.14293425

2 Approach

The aim of this study is to investigate the topic of "Artificial Intelligence, Policy and Ethics" using the Scopus Index database.

2.1 Data Analysis

This article uses bibliometric analysis to examine literature on AI, policy, and ethics. Bibliometric analysis is a method that applies some mathematical and statistical methods to analyze scientific data. Data such as authors, themes, cited authors, and cited sources are examined in a statistical way. This methodology allows us to understand the general framework and future directions of a particular field by analyzing the statistical results of the data.

2.2 Data Collection

The research was conducted on 20.06.2024 by searching the keywords "artificial intelligence, policy, ethics" in the "article title, abstract, keywords" fields in the database. In the research, a progressive order from the most recent scientific studies to the oldest was taken into account. The analysis was carried out with a bibliometric approach. The terms "artificial intelligence, policy, ethics" were used as keywords in the research. The specified keywords resulted in a dataset consisting of 802 records in Scopus that covered the entire scope of this scientific research. These studies were exported in CSV and BibTeX file formats to enable analysis and then analyzed using VOSviewer software to create network maps.

3 Scopus Database and Analysis of Acquired Data

Scopus, a database maintained by Elsevier, is a comprehensive resource that grants academics access to a diverse array of scientific literature. It also provides researchers with sophisticated tools for assessing articles, tracking research outcomes, and visualizing data. This platform offers a comprehensive collection of peer-reviewed journals, books, and conference publications. The content is regularly updated and developed to cater to the wide information requirements of scholars.

3.1 Scopus Data

Some visual results provided by Scopus from the search are shared here. Figure 1 above shows the distribution of data obtained by searching Scopus using the specified keywords "artificial intelligence, politics, ethics" by year.

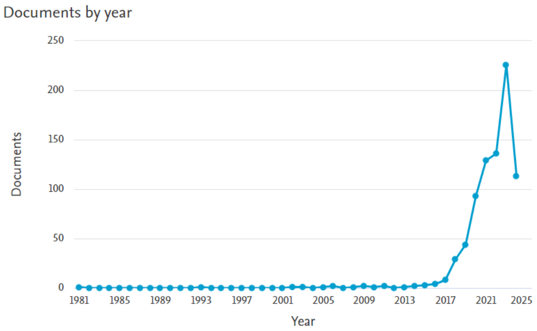


Figure 1. Table of Documents Per Year

Accordingly, until 2014, the number of studies that focused on these topics together in the fields of abstract, title and keyword was almost only 0, 1 or 2. 3 studies were conducted in 2015, 4 in 2016 and 8 in 2017. The number of studies increased to 29 in 2018, 44 in 2019 and 93 in 2020. The number of studies, which was 129 in 2021, increased to 228 in 2023. This graph shows us that this field is increasingly attracting more attention and is gaining more sweat in scientific studies. Figure 2 below shows the sources with the most publications by year and the change in the number of publications by year. Accordingly, Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes InBioinformatics, which has been published since 2010, and Nature, which joined it in 2011, have been the sources that have published on the subject since the beginning. These sources were followed by Science and Engineering Ethics in 2017 and ACM International Conference Proceeding Series in 2018. In 2020, this topic appeared in AI and Society publications, and in 2023, it became the source with the most publications

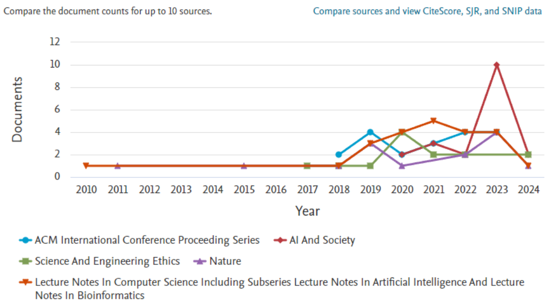


Figure 2. Table of Documents per Year by Source

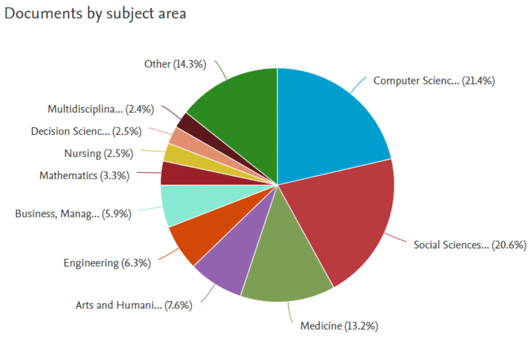


Figure 3. Table of Documents by Subject Area

on this subject. Figure 3 above presents the classification of published studies according to their fields. Accordingly, the field with the most studies was Computer Science with 319 studies. This field is closely followed by Social Sciences with 310 studies. These fields are followed by Medicine with 194, Arts and Humanities with 114, Engineering with 95 studies, Business, Management and Accounting with 88 studies, and Mathematics with 51 studies. Studies in other fields have an average of 21 percent of the total area.

This distribution demonstrates the pervasive nature of AI, ethics, and policy across disciplines, demonstrating their role in advancing research in this area and contributing to various areas of knowledge. The distribution of studies by document type is seen in Figure 4. Accordingly, approximately 50 percent of the total publications, with 404 publications, are in the "Article" type. This type is followed by the "Conference Paper" type with 145 publications and approximately 18 percent. 64 "Review" types constitute 8 percent. This is followed by the "Book Chapter" type with a similar number and percentage. Other types constitute the remaining approximately 15 percent. This distribution reveals the diversity of the types of studies conducted.

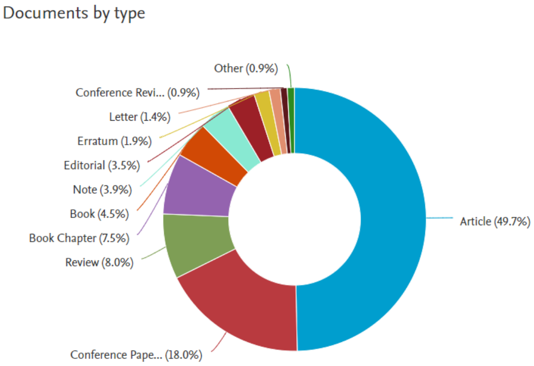


Figure 4. Table of Documents by Type

3.2 Bibliometric Analysis and Visualization of Results Obtained from VOSviewer

Bibliometric analysis is a methodological approach that applies quantitative analysis and statistics to written communication, such as publications. It is a method used to visualize and evaluate scholarly literature through mapping, to track the impact and dissemination of research through citations, and to uncover patterns of collaboration between authors and institutions. Bibliometric analysis can identify key trends in a field, productive authors, and the geographic distribution of research and resources. The technique has evolved to include complex network analyses that explore relationships among diverse scholarly outputs, and has spanned multiple disciplines. It is effective in understanding the state of scholarly work and the interactions within and among research fields, i.e., collaborative work [20].

3.3 VOSviewer

VOSviewer is a specialized tool designed for the purpose of generating and displaying bibliometric networks. These networks include citation, co-citation, bibliographic linkage, keyword co-occurrence, and co-authorship networks. This software is especially remarkable for its graphical visualization skills, which are essential for presenting massive bibliometric maps in a user-friendly manner. The use of network analysis is crucial in exploratory research since it enables researchers to uncover and visualize connections between different scientific outputs, such as publications, researchers, or thoughts in a specific topic [21].

The academic importance of VOSviewer lies in its capacity to facilitate initial investigations that can inform subsequent, more rigorous study. Researchers can utilize data visualization techniques to analyze co-occurrence data among papers sourced from Scopus, enabling them to identify patterns that may not be readily evident through conventional analysis methods. This not only expands the scope of investigation but also deepens the comprehension of the structure of a particular topic, highlighting the key participants, connections, and areas of focus [22].

In addition, VOSviewer has been included into diverse research procedures and has been utilized in research assessment and management settings. It assists in visualizing and analyzing scholarly activities and collaborations inside and between different fields of study. The program has a wide range of uses in several academic fields, making it a flexible tool for anyone who want to visually analyze scientific data. VOSviewer developers, who are associated with Leiden University's Centre for scientific and Technology Studies (CWTS), offer comprehensive courses that provide extensive training in scientific mapping methodologies. These courses emphasize the significance of the tool and its application in the wider field of research administration and evaluation [23].

3.4 VOSViewer- Co-occurrence of keywords and Bibliographic coupling of countries

Bibliometric analysis is a potent instrument in the complex realm of research, offering valuable insights by quantitatively evaluating academic literature. The visualizations of these studies can reveal the extent and distribution of knowledge across many disciplines, enhancing the understanding of intricate data. The next images and descriptions will explore this process in detail, demonstrating the capabilities of VOSviewer, a software that is used to map and visualize scientific landscapes by analyzing the frequency and co-occurrence of keywords. The following description outlines this technique.

The following image generated by VOSviewer, Figure 5, is used to determine the frequency of keywords in bibliometric analysis through a threshold value. Here, a threshold value of minimum 5 occurrences is set for a keyword to be considered in the analysis. Out of 4853 terms, only 364 of them meet this criterion, i.e. they occur at least 5 times in the analyzed dataset. The threshold value determination process is essential in bibliometric analysis as it allows focusing on the most relevant and frequent expressions and ensures that the resulting visualization accurately depicts the most critical data points.

This VOSviewer figure shows that the software will calculate the overall strength of association relationships between keywords that meet the predefined threshold. Keywords with the greatest overall link strength will be selected for further analysis. The selected number is fixed at 364, which is assumed to be the overall number of terms that meet the first criterion. The selection procedure plays an important role in discovering the most important and related terms in a dataset. These terms are essential for performing comprehensive bibliometric analysis and visualization.

This is a network visualization map created using VOSviewer, showing the connections between various terms in a bibliometric dataset. The most prominent and fundamental nodes, such as "artificial intelligence" and "ethics", symbolize the keywords that appear most frequently and therefore have the most importance within the scientific topic under study. The lines connecting the nodes represent the co-occurrence of terms in the same articles and indicate thematic connections. The proximity of the nodes indicates a higher degree of correlation in the literature and implies the existence of subfields or focused research areas within

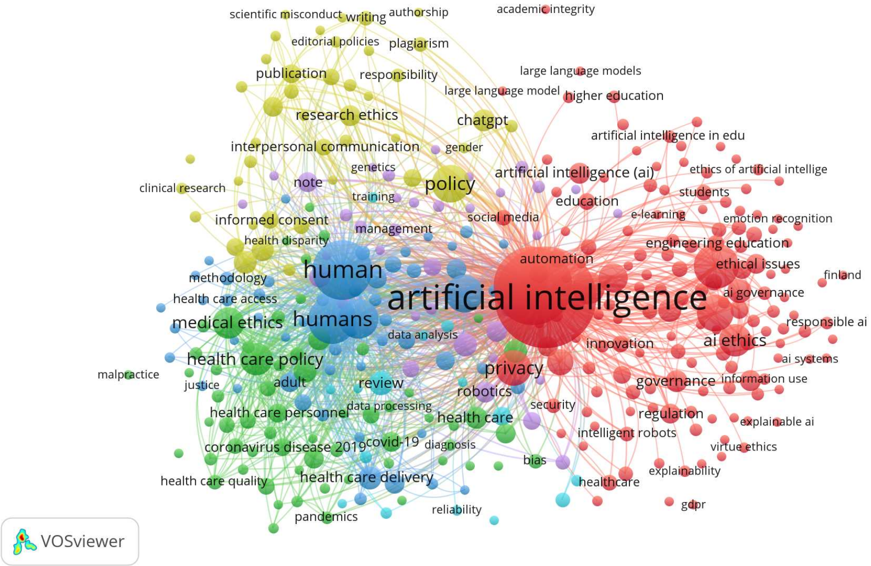


Figure 5. Network Visualization

broader themes. In this network, the word "artificial intelligence" was used 571 times, "ethics" 348 times, and "human/s" 332 times in total, making them the most used words. The word "ethical technology" was examined, which was used 115 times. The other words were used less than 100 times.

The first group, in the Figure 5, is represented by the color red. Among the words in this group, "artificial intelligence, ethical technology, AI ethics, philosophical aspects, privacy" are among the notable words. This group is followed by the second group, represented by the word "human" and shown in blue. Here, "human, humans, machine learning" are the most used words. This group is followed by the third group, represented by green and containing words such as "health care policy, medical ethics, public health". These are followed by the fourth group, represented by the word "policy" and colored yellow. In this group, words such as "research ethics, chatgpt, medical research, practice guideline", which are related to all other groups, attract attention. The image illustrates a bibliometric network visualization, specifically an overlay visualization,

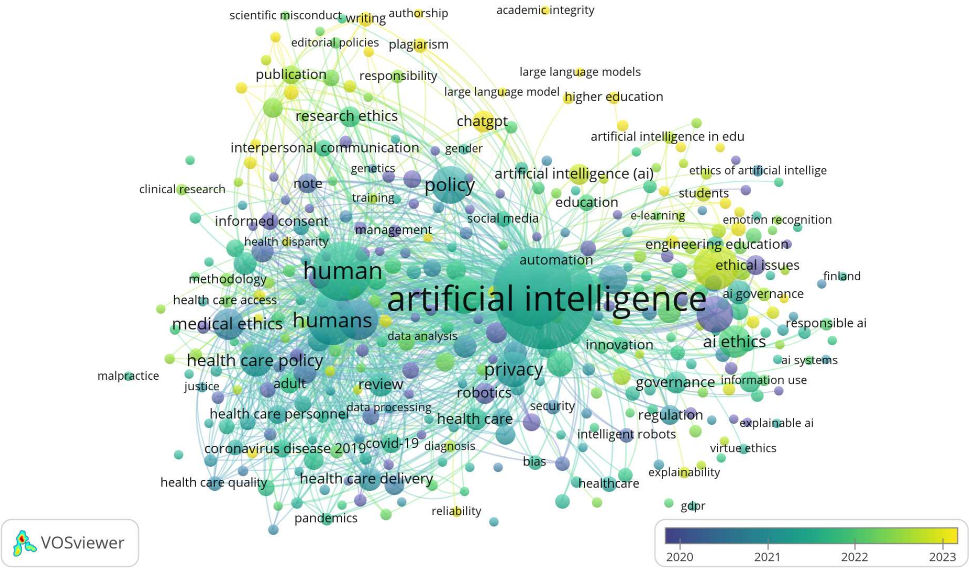


Figure 6. Overlay Visualization

where each node symbolizes a keyword. The node's size corresponds to the keyword's frequency in the dataset, and the lines connecting the nodes represent co-occurrence associations. The colors are most likely associated with distinct years or time intervals, indicating the progression of subjects throughout time. For example, prominent concepts such as "artificial intelligence," "ethics," "human/s" and "policy" indicate that these are major topics in the discipline. The relationships between these concepts and other keywords demonstrate interconnected fields of research.

The network's color overlay reflect the historical evolution of study focus from year to year. In Figure 6, the color change from purple to green and yellow shows us the change of keywords over the years. Accordingly, while words such as "philosophical aspect, research, law, robotics, morality" were used in early 2020, "ethics, artificial intelligent, human, data privacy" became some of the words that started to be seen more in 2021. Looking at 2023 and beyond, the words "ethical technology, chatgpt, human-centric, plagiarism, algorithmics, trust-worthy ai" started to be used more.

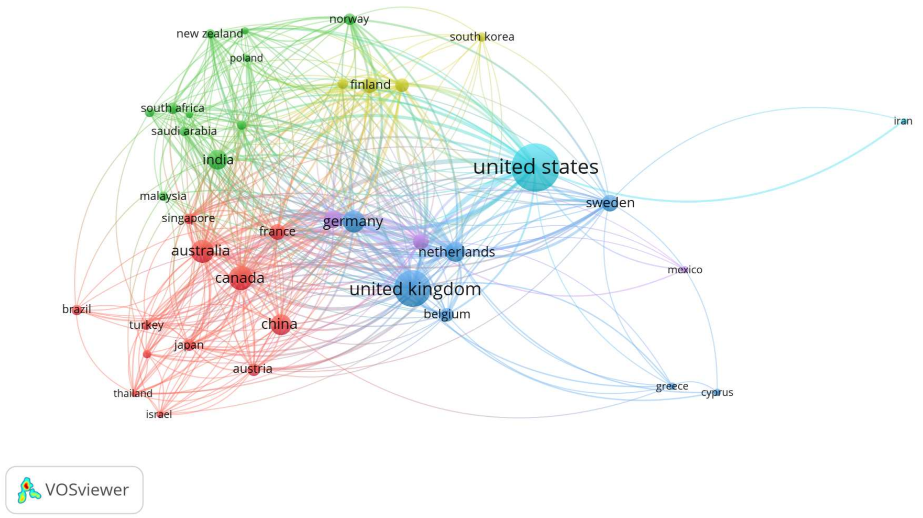


Figure 7. Distribution of Connections Between Countries - Network Visualization

Each node in Figure 7 represents a country, and the size of the node is likely to be proportional to the volume of publications or degree of activity. In determining the bibliographic coupling between countries, 41 countries with a minimum of 5 publications were used, out of 110 countries that met this requirement. The lines serve as indicators of the strength of collaboration or relationships; the thicker and more numerous the lines, the stronger the connections. The presence of central nodes, such as the United States, the United Kingdom, Canada, and India, suggests that these countries are important hubs in the network and are likely to have a greater volume of international collaboration or output. The United States has the most publications, with 238, and the United Kingdom the second most, with 142.

Figure 8 has been widely used to represent transnational collaboration based on academic papers or similar data over time. Colors on the lines and nodes can represent different years and show the progression of coupling over time. Cooler colors represent more recent years, warmer colors represent older years. Countries with a central geographic location and significant influence, such as the United States, the United Kingdom, and Canada, often play a major role in large and deep-rooted international bibliographic coupling. In addition, technologically advanced countries such as India, China, and Japan have also seen growth in coupling on these issues after 2022.

3.5 pyBibX

pyBibX is a Python package that use Artificial Intelligence to improve bibliometric analysis, which is an essential tool in scientific research. This tool offers valuable insights into current research trends, identifies influential researchers, and evaluates the effect of scientific publications. Conventional approaches to bibliometric analysis frequently require significant effort and consume a considerable amount of time. The integration of Artificial Intelligence can significantly enhance accuracy and efficiency. With the increasing

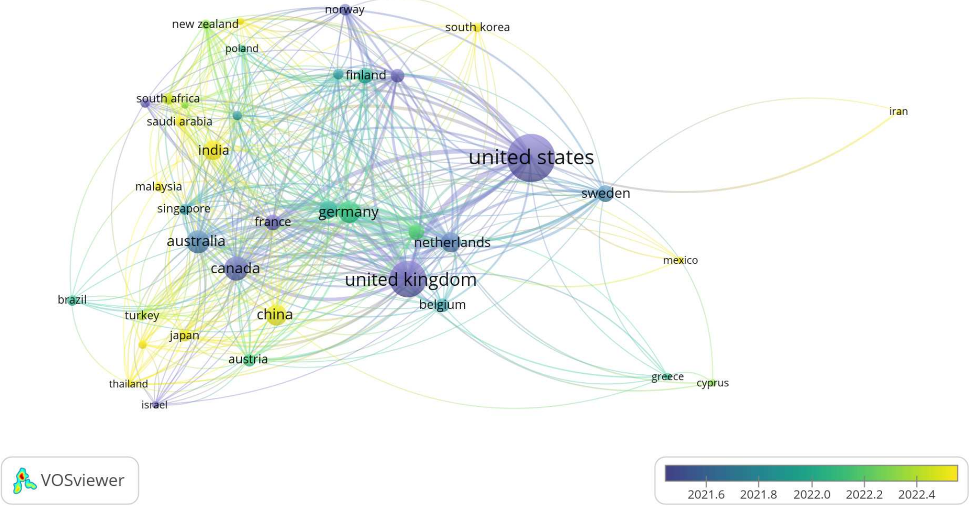


Figure 8. Distribution of Connections Between Countries - Overlay Visualization

number of scientific publications, there is an urgent requirement for sophisticated bibliometric systems that can effectively manage large amounts of data. Artificial intelligence (AI) techniques efficiently analyze vast datasets in an unbiased manner, providing researchers with unparalleled insights. These findings contribute to funding decisions, assist in strategic planning, and influence policies aimed at promoting scientific research. Therefore, the creation of an AI-driven bibliometric analysis tool holds great potential for providing significant advantages to the scientific community and society as a whole[24].

pyBibX is a freely available library for conducting bibliometric and scientometric analyses. It makes use of data from Scopus, WoS, and PubMed. The prominent characteristics of this system include its network capabilities, which encompass Citation, Collaboration, and Similarity Analysis. In addition, the library integrates AI functionalities such as Embedding vectors, Topic Modeling, Text Summarization, and other common Natural Language Processing activities. Additionally, it seamlessly connects with chatGPT. The library currently combines bibliographic sources by using the DOI and preprocessed Title columns for deduplication. This work introduces a tool that distinguishes itself from pyBibX in two distinct manners: firstly, it provides users with the ability to customize instructions for preprocessing databases, and secondly, it conducts deduplication for a user-defined number of columns, taking into account similarity as well as DOI and Title [25].

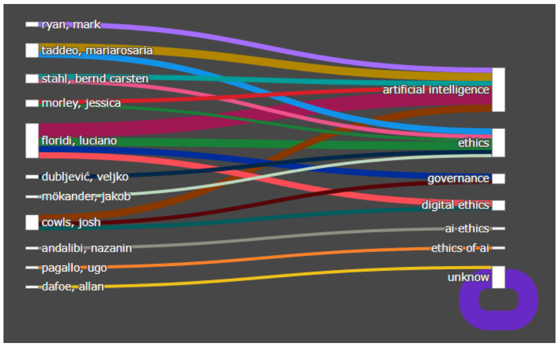


Figure 9. Sankey Diagram of authors and keywords

Sankey diagrams are a visual depiction that illustrates the movement or connection of something. Figure 9 displayed a diagram illustrating the top 20 authors and phrases that were either often used or had significant links among the papers analyzed.

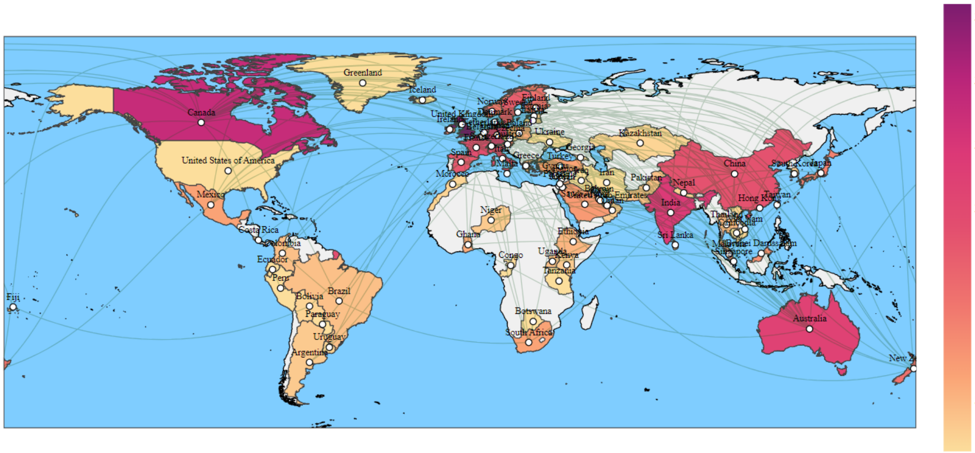


Figure 10. Collaboration Between Countries

This form of analysis, as in Figure 10, can be employed to evaluate the degree of collaboration across distinct nations in diverse scientific endeavors, such as scientific publications or research undertakings. For instance, it can furnish data regarding the countries that collaborate, the extent of their cooperation on specific subjects, and so on. These evaluations are crucial for formulating plans to enhance international research collaboration and for comprehending the global dissemination of scientific information. In the figure, the degree of cooperation between countries is represented by temperature colors.

4 Results and Discussion

During this research, we conducted a comprehensive analysis of the interrelated framework supporting research in artificial intelligence (AI) and ethics and policy. Choosing a bibliometric technique, we conducted a study by analyzing Scopus, a major database of scientific publications. We extracted relevant data by selecting specific keywords that aligned with our research focus. The use of Vosviewer software and the pyBibx library enabled the exploration of this data and revealed the complex networks of keywords and the global commonalities they contained.

When we examined the visualizations produced by Vosviewer, we noticed the prevalence of certain phrases such as “AI, policy” and “ethics”, which emphasized their importance in academic discussions. The analysis of the country collaboration network emphasized the strategic importance of the United States, the United Kingdom, and Canada, and emphasized their key roles as hubs for comprehensive scientific production and exchange. The shift from individual connections to density depiction provided a clearer understanding of the general patterns of international research collaboration.

Based on these observations, we can predict the path of future research and the patterns of international collaboration. This analysis highlights the need to use bibliometric approaches to measure the current state of academic progress. The findings of this study suggest that it would be useful to prioritize and emphasize growing areas where AI overlaps with other disciplines. In addition, it is suggested to encourage the formation of collaborative networks and support research projects that specifically address the gaps observed in less connected regions. The dynamic structure depicted in these graphical depictions is evidence of an ever-evolving scientific pursuit, where collaboration across disciplines and global cooperation are not only advantageous but also necessary for significant progress.

4.1 Limitations of the Study and Future Directions

Research limitations in bibliometric studies often arise from inherent limitations of the data sources. The databases used may not cover all relevant information, thus introducing potential biases. Another limitation is the reliance on keywords, which may miss subtleties in the study topics or may not capture emerging patterns that are not yet firmly established in the vocabulary. Future research could be improved by using other metrics, such as altmetrics, that take into account the impact of research on the Internet and social media platforms. This will provide a more comprehensive view of the impact of research. Furthermore, qualitative analyses have the potential to enhance the quantitative emphasis of bibliometric methods by capturing the essence and meaning behind statistical data.

5 Conclusion

Bibliometric analysis of AI ethical policy literature has shown that ethical issues have not kept up with the pace of AI development. It is essential to conduct more studies on ethics and ethical policies in the field of AI, which is advancing at a dizzying pace and where innovations are added every day. This is the way for the applications of technologies that we develop as humans to be truly beneficial for the benefit of society.

References

[1] Li R. Artificial intelligence revolution: How AI will change our society, economy, and culture. Simon and Schuster; 2020.

[2] Mitchell M. Artificial intelligence: A guide for thinking humans. Penguin UK; 2019.

[3] Davenport TH. The AI advantage: How to put the artificial intelligence revolution to work. MIT Press; 2018.

[4] Wamba-Taguimdje SL, Wamba SF, Kamdjoug JRK, Wanko CET. Influence of artificial intelligence (AI) on firm performance: the business value of AI-based transformation projects. Business process management journal. 2020;26(7):1893-924.

[5] Sjödin D, Parida V, Palmié M, Wincent J. How AI capabilities enable business model innovation: Scaling AI through co-evolutionary processes and feedback loops. Journal of Business Research. 2021;134:574-87.

[6] Sarioguz O, Miser E. Data-Driven Decision-Making: Revolutionizing Management in the Information Era. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023. 2024;4(1):179-94.

[7] Oke SA. A literature review on artificial intelligence. International journal of information and management sciences. 2008;19(4):535-70.

[8] Kuddus K. 1. In: Artificial Intelligence in Language Learning: Practices and Prospects. John Wiley & Sons, Ltd; 2022. p. 1-17. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119792437.ch1.

[9] Suta P, Lan X, Wu B, Mongkolnam P, Chan JH. An overview of machine learning in chatbots. International Journal of Mechanical Engineering and Robotics Research. 2020;9(4):502-10.

[10] Elgendy M. Deep learning for vision systems. Simon and Schuster; 2020.

[11] Janai J, Güney F, Behl A, Geiger A, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and Trends® in Computer Graphics and Vision. 2020;12(1–3):1-308.

[12] Parekh D, Poddar N, Rajpurkar A, Chahal M, Kumar N, Joshi GP, et al. A review on autonomous vehicles: Progress, methods and challenges. Electronics. 2022;11(14):2162.

[13] Aldoseri A, Al-Khalifa KN, Hamouda AM. AI-Powered Innovation in Digital Transformation: Key Pillars and Industry Impact. Sustainability. 2024;16(5):1790.

[14] Huang Z, Shen Y, Li J, Fey M, Brecher C. A survey on AI-driven digital twins in industry 4.0: Smart manufacturing and advanced robotics. Sensors. 2021;21(19):6340.

[15] Osasona F, Amoo OO, Atadoga A, Abrahams TO, Farayola OA, Ayinla BS. Reviewing the ethical implications of AI in decision making processes. International Journal of Management & Entrepreneurship Research. 2024;6(2):322-35.

[16] Ayinla BS, Amoo OO, Atadoga A, Abrahams TO, Osasona F, Farayola OA, et al. Ethical AI in practice: Balancing technological advancements with human values. International Journal of Science and Research Archive. 2024;11(1):1311-26.

[17] Konidena BK, Malaiyappan JNA, Tadimarri A. Ethical Considerations in the Development and Deployment of AI Systems. European Journal of Technology. 2024;8(2):41-53.

[18] Whittlestone J, Nyrup R, Alexandrova A, Dihal K, Cave S. Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. London: Nuffield Foundation. 2019.

[19] Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and machines. 2018;28:689-707.

[20] Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: An overview and guidelines. Journal of Business Research. 2021;133:285-96. Available from: https://www.sciencedirect.com/science/article/pii/S0148296321003155.

[21] Klarin A. How to conduct a bibliometric content analysis: Guidelines and contributions of content co-occurrence or co-word literature reviews. International Journal of Consumer Studies. 2024;48(2):e13031.

[22] McAllister JT, Lennertz L, Atencio Mojica Z. Mapping a discipline: a guide to using VOSviewer for bibliometric and visual analysis. Science & Technology Libraries. 2022;41(3):319-48.

[23] van Eck NJ. Methodological advances in bibliometric mapping of science. EPS-2011-247-LIS; 2011.

[24] Pereira V, Basilio MP, Santos CHT. pyBibX–A Python Library for Bibliometric and Scientometric Analysis Powered with Artificial Intelligence Tools. arXiv preprint arXiv:230414516. 2023.

[25] Nikolić D, Ivanović D, Ivanović L. An open-source tool for merging data from multiple citation databases. Scientometrics. 2024:1-23.

AIPA's International Journal on Artificial Intelligence: Bridging Technology, Society and Policy
ISSN: 3062-097X
Published: 07 December 2024

Reliability And Outcome Bias Issues In AI-Driven Forecasting Practices

Mehmet Beceren[1*]

[1]**Queen's University Smith School of Business, Kingston, ON, K7L 3N6; FORA-Invest Research, Oakville, ON; ORCID: 0009-0000-0238-6304)**

ORIGINAL RESEARCH PAPER

Abstract

Amid all the hype around the economic potential of AI technologies, there is a growing risk of data analysis overkill in many applications. That risk is particularly high for the forecasting and decision-making models being proposed in social contexts such as economic policy, financial investment, and corporate decisions. Common research practices in those areas keep focusing on incidents of statistical discoveries. They omit the substantial reliability issues stemming from the nature of the data that offers very limited 'learning potential' for the machine learning (ML) algorithms. In this paper, I focus on the use of ML algorithms applied to such forecasting problems. I illustrate the reliability issues with a detailed example that builds a stock investment strategy by using the XGBoost algorithm on a large data set. The example demonstrates how easy it is to discover seemingly interesting random patterns when we fit over-parameterized models on historical data. The results also offer practical methods to investigate the statistical flukes and the reliability issues that are concealed by complex algorithms of artificial intelligence being blended with natural human ignorance, as seen in popular practice.

Keywords: forecasting, reliability, machine learning, asset pricing, factor investing

OPEN ACCESS
AIPAJ Vol:1, Issue:1

***Corresponding author**
mehmet.beceren@queensu.ca

***Submitted** 22 June 2024*

***Accepted** 15 November June 2024*

Citation
Beceren, M. (2024). Reliability And Outcome Bias Issues In AI-Driven Forecasting Practices, In AIPA's International Journal on Artificial Intelligence: Bridging Technology, Society and Policy, (Vol. 1, Number 1, pp. 25–40. DOI: 10.5281/zenodo.11200944

1 Introduction

> "*It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of light, it was the season of darkness, it was the spring of hope, it was the winter of despair*". Tale of Two Cities by Charles Dickens

This famous opening line of the Charles Dickens classic, *Tale of Two Cities*, works perfectly to encapsulate the main theme of this article in a nutshell. The simple and timeless language of the novel fits quite well to our data-obsessed times.

> "*It is the best of times, it is the worst of times, it is the age of artificial intelligence, it is the age of human ignorance, it is the epoch of data analytics, it is the epoch of statistical deception.*"

In the current digital age, there is a euphoric race both among businesses and academics to showcase the latest machine learning (ML) applications in their own practice areas. We see an exponential growth in ML-driven research output and commercial applications that utilise increasingly complex predictive models with ever-larger data sets. Amid all the buzz around the economic potential of artificial intelligence (AI) technologies, however, there is also a growing risk of data analysis overkill in many cases. The rush to catch up with the self-fulfilling 'AI revolution' wave is inevitably generating misused, misguided implementations alongside many fascinating products. That risk is particularly high for the forecasting and decision-making models being proposed in social contexts such as economic policy, financial investment, corporate strategy and such.

In this paper, I focus on the use of ML algorithms applied to forecasting problems. I discuss the unique nature and the limitations of historical data sets that have stochastic state-and-time dependent variables. I illustrate the specific issues with detailed examples from the financial investment strategy applications.

The main sources of concern about the excessive use of ML techniques to build decision models are as follows:

1. A unique sequence of historical events caused by incidental patterns of stochastic factors, and complex confounding effects, do not provide useful data sets that are sufficient to make reliable inferences about the future. In other words, unlike many successful applications such as image recognition, complex pattern discoveries within historical data sets, may not amount to ’learning’ or ’intelligence’ of any sort.

2. Although the Train-Test-Validation cycle of the ML algorithms may generate incidents of attractive back-test results (i.e. performance validations) on historical data, the relation between the performance metrics and future reliability may be highly uncertain.

3. In cases where (1) and (2) are true, there are significant reliability and outcome bias issues in ML-driven models. A forecasting solution that looks encouraging with historical data, may easily be an *over-fitted* fluke driven by a lucky draw from a large random set.

Surprisingly, neither academics nor professionals in social sciences tend to sufficiently address these serious issues. The hype to assign a flashy "AI" label on new products seems to trump the obvious reliability challenges. Probably fascinated by the speed and efficiency of ML algorithms, the data analysts seem to ignore the significant likelihood of making incidental, lucky discoveries with big data. Also, they tend to forget that a longer history, occurred and evolved with unique circumstances in time, does not necessarily mean a bigger data set with relevant and useful information.

The following sections will discuss the reliability risks in further detail along with some examples from the finance literature. At this point, however, it is probably a good idea to offer a bit more clarification about the concepts mentioned above for the non-expert reader.

1.1 Data mined flukes versus reliable insights

To understand the outcome bias and statistical flukes found in historical data analysis, let’s consider an extreme case where the target variable (i.e predicted or forecast variable) is completely random. Assume that you are the manager of a company named Lucky Bets Co. You believe in luck and in lucky people. You are in the gambling business, but you do not place bets on games. Instead, you place bets on lucky people. You provide funding for the gamblers that you think are lucky to win at the roulette table in return for a large share of the prizes they win.

The skillful data analysts of Lucky Bets Co. collect a large historical data set on many attributes of the addicted roulette players. The data set includes the players’ winning percentage over the past 5 years, amount of money they lost, age, height, profession, post code, shoe size, hair color, first letter of their names, star sign, and many others. The analysts divide the data set into Train, Test and Validation samples, and then let the ML algorithms run over-parameterized deep learning models, as they always do. After millions of iterations, the analysts provide a combination of attributes that predict a higher probability of winning at the roulette table. The results are confirmed in the Validation (hold-out) sample as well. All standard statistical measures check within the Test and Validation sub-samples.

What would you do? Would your expectation of winning probability change for the people with the right attributes? Assuming everyone plays the same game with the same odds, would you bet on the people with "statistically proven" success? Are there lucky characteristics, or lucky data analysts here?

Your betting decision actually does not matter. It will not change the odds of winning one way or another. The data analysts did not do much more than wasting electricity. They were lucky. Also, it was almost inevitable that they would find a fluke that works after so many iterations over countless combinations of gambler attributes. The historical results, no matter how statistically significant they may look, provide no guidance for the future outcomes that are completely random. The analysts just documented an observation bias - a lucky historical outcome with no implication for the future. That is because each roulette run is an independent random event by construct.

On the other hand, it may actually be a good strategy to go along with the model and promote it as the new, cutting-edge AI-Powered innovation by Lucky Bets Co. If, somehow, it catches another lucky episode, it may bring extra fame and fortune. (Actually, there are online betting companies, especially in sports betting, that offer AI models for their customers. See examples such as *DeepBetting*, *BetIdeas* or *Infinity Sports AI* among others.)

Typically, when there is a proposed forecast model, or a decision method, we are likely to see some instances of out-of-sample performance metrics as the key results. An instance of out-of-sample test is considered sufficient to prove the suggested model’s worth. The reliability risk and potential ’observation bias’ originating from the iterative data mining embedded in over-parameterized ML models are mostly downplayed. As a result, the real important question is mostly left unanswered.

Given that we are able to find some model that performed well in the past, how confident are we that the model will provide significant performance in the future as well? What is the correlation of the actual implemented results with the (out-of-sample) past performance that we could dig out by sifting through the data?

In the case of Lucky Bets Co., we know the answer. The correlation is zero. If we keep repeating the exercise of finding new hidden patterns with strong past performance, by utilising more and more data, and then we implement each model as a separate AI-powered betting strategy, we surely will find out that the documented past results have no relevance for future outcomes. Such an analysis would serve as the proper back-testing experiment to provide some guidance about the reliability of the methods.

Those experiments almost never show up in the results of ML-driven forecasting research, especially in the social science fields such as economics and finance. Both academic researchers and professionals keep showing instances of statistical discoveries, instead. Their common audience usually cannot distinguish the lucky coincidences hidden behind the complex and automated algorithms.

The computational power of the ML algorithms help the empirical researchers with the fast discovery of interesting patterns, but the findings might be just an '*observation bias*' - a fluke of the unique set of circumstances that might not repeat ever again. Therefore, when we try to import the predictive AI technologies to forecasting practices, one of the first questions to ask has to be: "*How similar is my case to Luck Bets Co.?*"

Many examples of empirical research output that are being promoted with sparkling AI labels might not be far from just another Lucky Bets exercise. It is common to find similar examples, especially in fields that rely on non-repeatable, state-and-time dependent data. Just to mention a few, Berman et al. (2021) [1], presents a model that integrates big data analytics with strategic planning to optimize business decisions; Lee and Chen (2020) [2] presents a machine learning model that predicts both employee success and retention; Chen and Guestrin (2016) [3] predicts political instability with ML models fitted onto social media data, and many others. In each study, we see some contemporaneous covariance among variables being documented with no in-depth discussion about cross-validation and reliability issues originating from particular methods and data samples used.

Another example, Erel et al. (2021) [4], presents results of decision tree models to select directors for corporate executive boards. The target variable used is "director success" which is some complex proxy measure constructed with authors' subjective discretion. It includes ad hoc indicators of shareholder popularity and company profitability. The ML algorithms run an over-parameterized decision tree model on a predetermined training sample and a fixed test sample. The model iterates over tens of different personal attributes, from gender and age, to the name of the university that the director graduated from. There is no cross-validation across different periods, industries, etc. There is no proper validation experiment over time either. The incident of the statistical results are particular to a very narrowly defined data construction process.

To find some interesting-looking pattern in large data sets does not require much skill since we have the technology to automatically iterate over pretty much countless parameter combinations. Those empirical research articles, and many other similar work, are arguably not that far away from the Lucky Bets case. Although the publications succeed in uncovering intriguing incidents of empirical results, future reliability of the findings, as a useful forecasting model, is a wide open question.

Historical data sets used for forecasting models in social contexts usually do not offer the breadth for proper cross-validation tests. After all, we have only one trail of the actual history. Therefore, AI methods that are employed successfully in other areas, may be unsuitable, or misleading, due to the irreducible over-fitting risk originating from the nature of the data sets. Quick and lazy ML applications with historical data require scrutiny within their own context since the standard data validation methods are mainly not feasible.

1.2 A special case: Financial asset pricing and investment strategy applications

Finance has been at the forefront of digital automation and the commercial use of AI technologies. Financial industry operates on an extremely digitized platform that produces immense amount of data, and the data universe is mostly accessible for analysis. Data collection is relatively easy and straightforward. Financial industry employees, especially on the trading and investment side, tend to be highly skilled in data analysis and coding practices. At the same time, the potential reward of successful forecasting models can be very high and fast especially in the trading and investment world.

In addition to the general economic backdrop that motivates the use of ML models in finance, the academic literature also provides some extra justification for the use of sophisticated predictive models in this field. For example, the investment management industry makes use of models inspired by the academic asset pricing literature. Contemporary empirical research in this field has developed around the Arbitrage Pricing Theory (APT), introduced by Ross (1976) [5], and the Stochastic Discount Factor concept, introduced by Merton (1973) [6], that lay out the framework for the empirical inquiries into the driving factors of financial asset returns. The seminal work by Fama and French (1992, 1993) [7],[8] and a large body of empirical work that followed the same path into the inquiry of asset returns, built a cultural tradition that is baked into the contemporary curriculum of finance education. The highly-regarded Chartered Financial Analyst (CFA) program also teaches the APT and related concepts that underpin the empirical inquiries into historical data to search for the drivers (factors) of asset returns.

The basic idea is that the financial asset returns are determined by their sensitivity to (potentially many) risk factors that the agents trade in the market place. It sounds like an axiomatic statement that opens up a wide gate for the inquiry of those elusive factors.

The complex and efficient predictive machinery offered by the recent developments in AI technology are welcomed as a powerful tool to work on the eternal questions of the investment industry and the asset pricing academics: *What drives the differences in asset returns? What should be the decision criteria to choose the assets to invest for the short or the long term?*

To answer those questions, quantitative finance professionals and academics dedicate a great portion of their work to building predictive models for the asset return dynamics. Common empirical research practice starts with an investigation of the so-called factors that show some covariance with the cross-sectional variance of asset returns in hand. Once the candidates for useful factors and trading signals are found, they are put into a back-testing process to validate their historical success. The instances of out-of-sample back-test results achieved over a selected period is usually considered as a sufficient experiment result. Reliability is mostly left out of the discussion.

With the advances in data access and computer power, the statistical discoveries became rather easy and fast. Sequentially, the number of academic publications showcasing the discovery of new factors started to grow rapidly during the early years of this century. From economic and financial indicators, to eccentric sentiment and risk measures, numerous variables are thrown into predictive models with the hope of finding some covariance patterns. The finance professionals started to implement such models for portfolio construction and proprietary trading practices at an accelerating pace, as well. By the time we reached 2010s, the asset pricing literature became a ’factor zoo’ as famously coined by Cochrane (2011) [9]. The criticism and warnings about the scientific quality of the empirical findings began to accumulate.

The critics highlighted two key observations. One, the published articles were presenting obviously over-fitted models that did not pass the statistical hurdle tests and the test of time. Two, the investment strategies based on the suggested factors mostly failed to deliver returns documented in their back-tests. In other words, the real out-of-sample tests proved that neither the predictive models nor the underlying theory was able to deliver a decent reliability over time.

The published statistical results were not necessarily wrong or careless, however. The issue was that the suggested models were not far from our Lucky Bets Co. example, again. People put too much faith in the instances of pattern discoveries driven from over-simplified models. Even the factors suggested by Nobel Prize winning Eugene Fama and Ken French’s work failed to repeat the documented patterns consistently, once they were implemented as real investment strategy products. See Carhart (1997), Fama, French and Carhart (2000), Fama and French (2015) [10, 11, 12] for more detail on that point.

As a result of humbling real-world validation experiences in the financial markets, the discussions on the potential uses of ML-driven or other type of predictive models started to shift from euphoria to skepticism, especially over the past 10 years. At this point, we can probably say that finance is more advanced in the discussions about reliability compared to other social science fields.

The discussions are evolving in three main paths. The first path can be called the ’scientific quality’ argument. Studies such as Bailey and Prado (2013, 2014), Prado (2020), Harvey et al. (2016), [13, 14, 15], [16, 17] present strong arguments about the ’data mining’ and ’over-fitting’ issues. They discuss the rampant use of statistical overkill and careless back-test practices spoiled by the ease of access to computational tools and large data sets. The criticism raised by Prado and Harvey is mainly about the errors, tricks and and biases in statistical inference. They are valid and crucial points that highlight the risk of false discoveries and wrong inferences made in common research practices.

However, the failures of the forecasting models in this field are not necessarily driven by the lack of diligence in statistical analysis. It is driven by the fact that there is an irreducible reliability issue caused by the natural instability of the system dynamics. To argue for scientific quality of the predictive models applied to naturally unpredictable dynamics is somewhat redundant. After all, it is impossible to determine the causes of model failure with confidence when the model is too simple relative to the stochastic complexity of the system in hand. As the past research experience showed time and time again, no matter how robust your statistical results may be, the estimated model may fail to perform, or become irrelevant, simply because of the evolving complexity of the system not being captured by the available historical data.

Second path can probably be categorized as the ’benign over-fitting’ effort. Studies such as Kelly et al. (2022) [18, 19] do not find the risk of over-fitting as an impediment to ML-based iterative search for hidden patterns. Instead, they try to develop the machinery that let over-parameterized, over-fitted models to automatically iterate towards a rather distilled form. They also let the ML algorithms to adjust over time, and over different states, and also let the algorithms discover those adjustment rules independently from the data. This line of research focuses on methods to distill signals without being limited by theory, or any other priors. It is probably a step in the right direction with a powerful inspiration, but reliability is still mostly missing in the discussion. Instances of good-looking back-tests are presented without a demonstration of how reliably the complex models might perform relative to simpler decision rules over time.

The third path suggests an alternative use case for the ML algorithms. The work by Chean and Zimmerman (2020) , and Chen and Valikov (2021) [20, 21] embraces ML-powered intentional data-mining to investigate the reliability of the models proposed by the ’factor zoo’ literature mentioned above. The approach is a leap from simply documenting another discovery of factors towards an analysis of real out-of-sample performance. With a multitude of different data-mined correlations that can easily present some historical performance, this line of work aims to establish a benchmark for the value and usefulness of the models that claim to have some prediction power.

I think Chen’s work is an example of how the AI technologies can bring a significant disruption to social sciences and forecasting practices. By allowing the fast and automated search algorithms, ML models can help us to devise tools to help distinguish a humble analysis that provides insights to highly complex and fluid stochastic systems from a statistical fluke published with a dose of confirmation bias and academic hubris.

Meanwhile, although similar discussions happen in parts of the investment industry, the commercial pressure to roll out generic commercial products with a flashy AI-name continues. Take the "AI-labeled" exchange traded fund (ETF), QRFT - QRAFT AI Enhanced US Large Cap ETF, for example. This ETF relies on "AI-powered models" which are based on some back-tested historical correlations - not some "intelligence" gained by learning from very large big data sets as we see in other fields. As seen in the Figure 1, there is no convincing performance of any sort. The performance over the benchmark index converges to zero as you would expect from any Lucky Bets exercise. In academia, as well, we can observe an intellectual inertia to keep producing those incidental back-test results. The publication rate of such research will inevitably fade away as their value-added is tested over time.

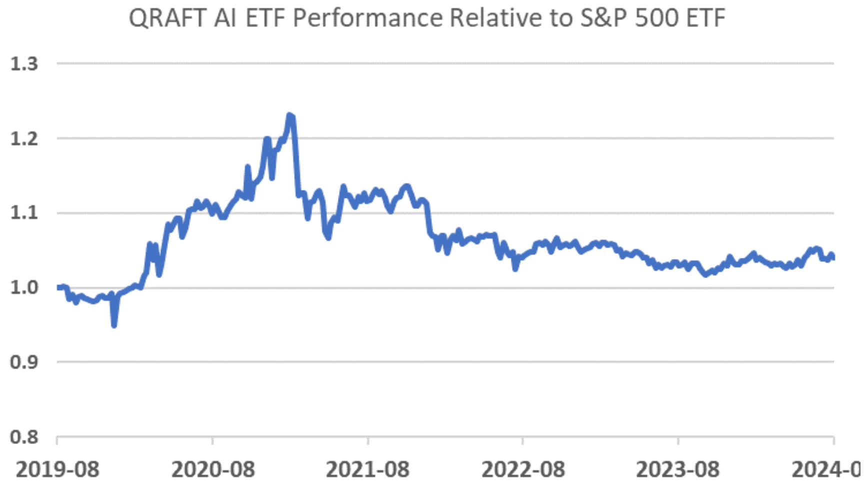


Figure 1. AI-Powered Large Cap US ETF Performance; *Source: www.qraftaietf.com*

Motivated by the contributions of all three paths followed by recent finance literature discussed above, the following section presents an example of an investigation into an ML-driven method applied to asset returns.

First, by using historical data on stocks and company characteristics, I run a decision-tree model (XGBoost) to intentionally data-mine the factors that distinguish the Winner (high future return) and Loser (low future return) stocks - similar to our Lucky Bets case. I demonstrate how easy it is to document some seemingly successful back-test when you are not much concerned about cross-validation. Then, I run a series of investigations to discuss how similar the case could be to the Lucky Bets scenario.

I do not use the ML techniques to show how we can predict Winners and Losers in the stock market. Instead, I utilise the power of ML to show how reliable the employed data and methods might be for the specific case in hand.

2 Material and Methods

Let's assume we have a problem of building an ML-based stock selection method that can possibly be turned into an ETF product similar to the one mentioned above. However, the financial literature does not offer much help about the predictors of stock returns. Although there are some obvious common sense approaches to portfolio construction and investment, there is no formula to predict which stocks will outperform the others over a certain period, say, the next 3 months or 2 years. Actually, there cannot be a formula because, if there was one, it would be instantly exploited and vanish, anyway.

The markets facilitate exchange of expected risks and returns that fluctuate according to perceived opportunities and costs that vary across numerous agents over time and economic conditions. Incidental clusters of those expectations cause demand-supply imbalances to move the asset prices. Additionally, when the underlying assets deliver unexpected positive or negative economic performance, share prices adjust so as to remain consistent with changing conditions.

Although financial theory does not offer a magic formula, at least it provides the framework that allows empirical investigations for the elusive, incidental or persistent risk factors that drive returns.

The equation for the expected asset return R$_{i,t}$ for the asset i at time t is given by:

$$E[R_{i,t+1}] = \Gamma_t(\beta_{i,t} \cdot X_t)\ ) \tag{1}$$

where:

- R$_{i,t+1}$ is the return to be realized at time $t + 1$, E is the expectation operator,
- $\beta_{i,t}$ is the $N \times 1$ vector of exposure of asset i to the observed N factors X_t,
- X$_t$ is the $(1 \times N)$ vector of factors that are assumed to affect expectations,
- Γ_t is the time-specific function that translates observed factors to retun expectations

Here, one can think of X$_t$ as the set of themes and criteria that influences asset return expectations and portfolio preferences at a point in time. For example, they may be a popular theme such as AI to drive growth expectations for the share price of Nvidia lately. The exposure $\beta_{i,t}$ of Nvidia to the AI theme may be high while for a company such as Alcoa which is in the business of metal mining globally, $\beta_{i,t}$ may be zero. One can think of $\beta_{i,t}$ as traffic lights switching on and off over time differently for each stock as themes, risks and investors' preferences evolve.

The issue is that we do not know any of those parameters in that simple abstraction (1). We have some idea about what the investors generally consider, maybe factors such as profitability, volatility etc., but we have no idea how those considerations might translate into return performance at a point in time. We would like to believe that we have some intuitive list of what X$_t$ could consist of, but we do not have a clean method of measurement either. Therefore, equation (1) does not tell us anything other than 'whatever works!' offers no insights. (That pretty much sums up the field of asset pricing in finance.)

All we have is the historical realizations of R$_{i,t}$ and a data set of factors X$_t$ that we imagine, and hope, will show some covariance with future returns to help us distinguish the Winners and Losers. So, as one can easily see, the problem in hand is not much different from the Lucky Bets scenario discussed earlier.

Our case is probably a very good example of potential use cases of AI to solve complex problems without a specific formula. We observe some phenomena that is driven by complex interactions of unknown set of factors. We hope that the computational technology will be able to sift through huge data sets to generate useful predictions although we are not able to identify what exactly drives those predictions. Image recognition with deep neural networks is such a process. We cannot tell how exactly the image recognition works, but we see that computer algorithms trained on big-enough data sets can accumulate the cognitive experience to generate impressively accurate predictions. A deep learning model trained on millions of X-ray images, for example, comes close to obtaining a life-time experience of a doctor. That is made possible by being exposed to a very large number of instances of a well-defined problem.

To have access to *'the instances of a well-defined problem'* is the key issue that distinguishes forecasting from other problems. As we increase the size of our data set, by extending the history for example, we do not necessarily accumulate the instances to learn from. The phenomena that we register in our data sets are mostly the outcome of instances of unique or temporarily relevant circumstances. That is why, with historical finance data, we do not see the 'double descent' phenomenon that is remarkably demonstrated by Belkin (2021) [22]

Alonso and Sonam (2023) [23] applies Belkin's (2021) [22] methods to financial return data set and shows that the learning accuracy rate does not improve with larger data sets with more parameters. Alonso and Sonam (2023) [23] formally experiments with the financial data sets and documents how the historical data sets fail to show any potential for 'double descent'.

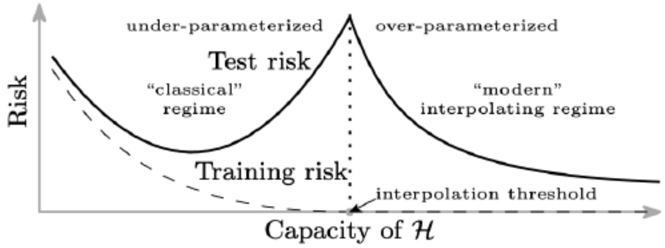


Figure 2. Double-descent of over-parameterized ML models shown in Belkin (2021)

In our case, we have a similar data set with (very) limited learning capacity. In our modeling exercise, we need to humbly accept that fact, and try to analyze what we can distill from the data set. As discussed in the earlier section, the main argument and motivation of this paper is the lack of such approaches in common ML-driven forecasting practices. There is too much focus on the instances of statistical findings, and too few discussion about how much luck is involved in those findings.

Our data set is the same as the one used in Guida (2020) [24]. The data is available through the book's Github. We have monthly data on the Total Return of 1212 global stocks over a 20-year period from 1998 to 2019. All the stock characteristics (features) to be used as predictors are scaled and normalised and they are ready to be used in ML algorithms. Not all stocks are alive throughout the 20 years. Some vanish, others emerge, as they always do, in the data. Therefore, we have an unbalanced panel of cross-sectional, time-series data with over 208K rows (roughly [20 year x 12 months x 1000 stocks]).

Along with the stock returns, there are also 93 different company characteristics such as valuation ratios, past returns, past volatility, accounting measures of profitability, growth, debt, capital expenditures, and many other similar variable with seemingly relevant economic measures. Of course, we do not know whether any of these variables make any reliable predictor of Winner or Loser stocks at any time. Although the variables seem to have financially meaningful labels, they are not necessarily different from any random number in relation to their predictive value for future stock returns.

Our prior is that we have some function given in (1) that will partly reveal itself in the large data set in hand. The data is aligned such that a model can be fitted as:

$$R_{i,t+1} = \Gamma_t'(\beta_{i,t}' \cdot Z_t) \ + \epsilon_{i,t+1} \tag{2}$$

where:

- $\mathrm{R}_{i,t+1}$ is the return to be realized at time $t+1$,
- $\beta_{i,t}'$ is the estimate of exposures of asset i to the observed N factors Z_t,
- Z_t is the $(1 \times N)$ vector of factors that we have in hand with no causal relation with the returns, necessarily

- Γ_t' is the time-specific estimated function that translates observed factors to future returns observed

We transform the problem to the following form:

$$rank[R_{i,t+1}] = rank[\Gamma_t'(\beta_{i,t}' \cdot Z_t)] + \phi_{i,t+1} \tag{3}$$

because we are interested in the Rank of the future returns across stocks at a point in time. We set R$_{i,t+1}$ as the Next 3-Month Return. For example, in 2009-December, we would like to predict the Rank of returns over the 3 months from 2010-Jan to 2010-Mar. At each point in time (i.e. each Month in the data set), we define the top 80% as Winners = 1, and the bottom 20% as Losers (Winners = 0).

To fit a tree-based model, we can use the XGBoost (Extreme Gradient Boosting) algorithm. XGBoost builds an ensemble of trees sequentially, where each tree corrects the errors of the previous ones by focusing on the hardest-to-predict cases. The algorithm incorporates regularization to prevent over-fitting. It is popular in categorization (1 vs. 0) problems. The model output includes decision trees similar to the Figure 3 below.

As a start, let's pick a small portion of the large data set. Let's take the first 3 years as the *Train*, and pick the 3 months immediately after the *Train*, as the *Test* sample. Our hope is that the model will train on the past 36 months as the 'most relevant' period to forecast the Winner and Loser stocks in the next 3-month period.

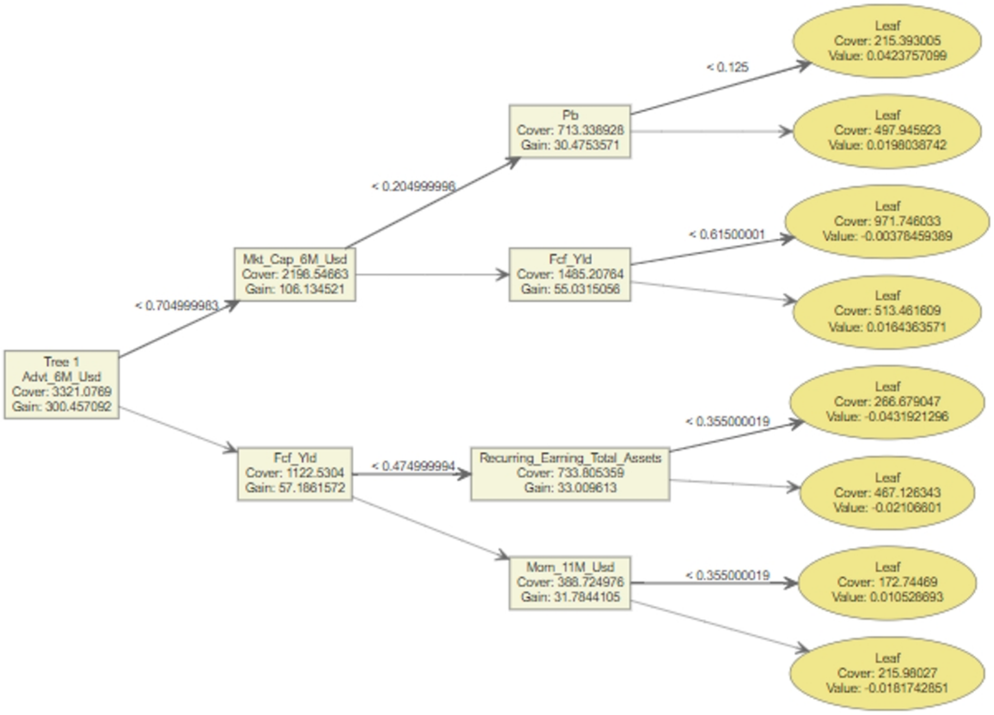


Figure 3. A partial picture of an example XGBoost tree

To find the best-performing model, we enable hyper-parameter tuning and let the gradient descent algorithm iterate over various parameter combinations and pick a model based on the AUC (Area Under the Curve) measure based on the ROC (Receiver Operating Characteristic) curve. For our given sub-sample, the AUC numbers as seen in Figure 4.

We see that the *Test* AUC tapers off quickly while *Train* fit is improved with iterations. This is not surprising since the useful information content of the data is limited in a similar fashion to the experiments conducted by [23].

The selected final model shows an ROC curve in Figure 5. The predictive ability looks poor but in the financial markets context, marginal improvements in the probability of picking Winners versus Losers may

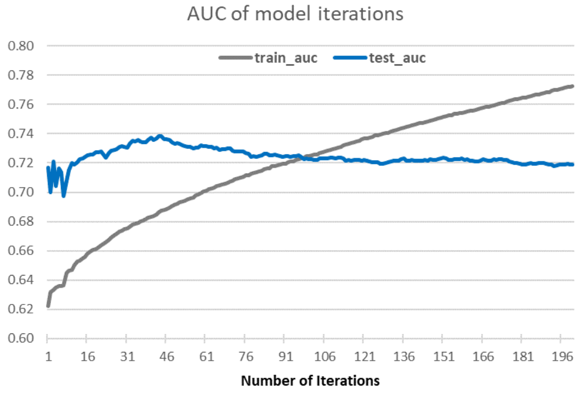


Figure 4. AUC of Test and Train over 200 Iterations

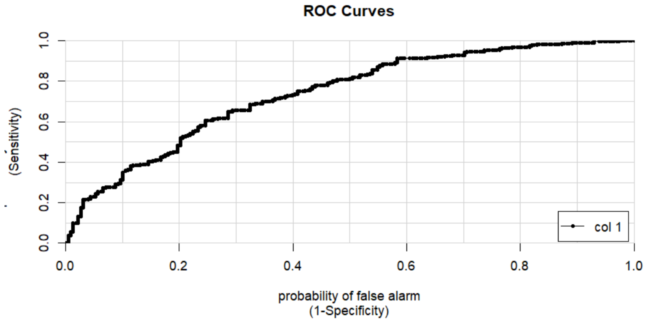


Figure 5. ROC Curve suggests very limited predictive potential but it is better than expected for stock returns

have significant economic meaning. The idea is not to reach high accuracy, such as in the X-ray image recognition problems, but to raise the odds somewhat, even if it is small.

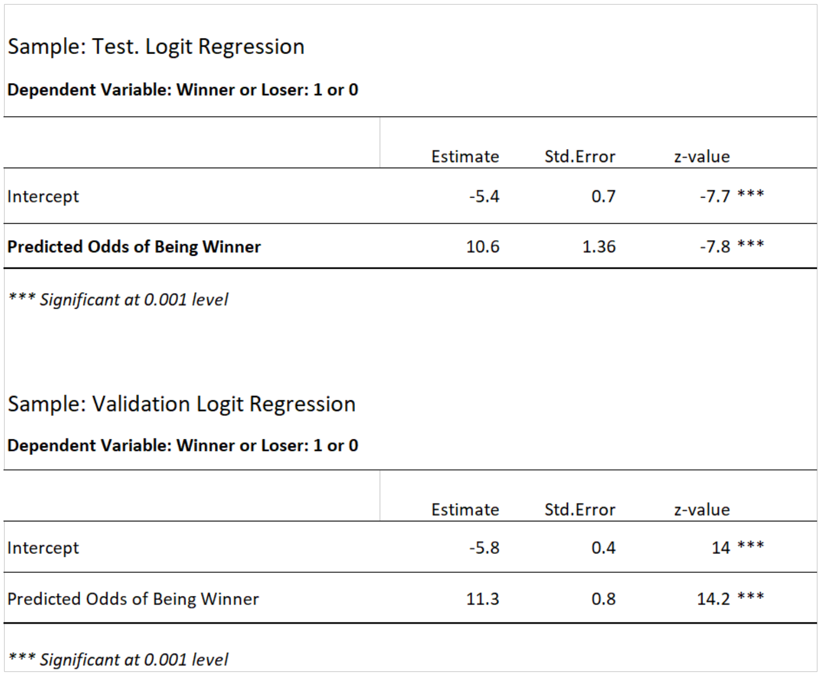
Imagine running a hedge fund managing $10 billion, a 1% increase in the odds may amount to non-negligible gains. Therefore, in the context of stock returns, the results look interesting, and even remarkable. When we carry the model to a *Validation* sample that is later than the *Test* sample, we see that a similar outcome occurs.

Table 1 below presents the results of the Logit regressions of Predicted Probability on the Realised Probability of selecting Winners. Both the *Test* sample and Validation sample results confirm that the model-estimated probabilities have a statistically significant correlation with the actual outcomes. That is quite encouraging.

The *Test* sample used to produce the results is in 2002. If we were in 2003 now, and we had run the same method to get these results in 2003, would we recommend the ML-driven stock selection strategy as a useful model? Maybe, if we believed that the results are repeatable in the future. However, we did not produce any evidence on how repeatable the results could be.

At this point, it is important to remember the discussion about the 'instances of statistical results' being published, and sometimes commercially implemented. In the examples discussed earlier, an in many other similar work, the researchers report the incidents of interesting results appearing in their data sample, but do not proceed with further discussions on reliability or future usefulness. They conclude their work with the reporting of the statistical instances without analysing how easy it might be to find a fluke with the data and the ML machinery in hand. Those results do not reflect any 'learning' or 'AI', just like the results shown here so far do not.

Now, let's develop our example further by utilising more of the data sample. How would the results look if we were to re-run the modeling exercise over other periods, and then look at the performance of the portfolios that might have been constructed with the help of the ML-driven models?



Sample: Test. Logit Regression

Dependent Variable: Winner or Loser: 1 or 0

	Estimate	Std.Error	z-value
Intercept	-5.4	0.7	-7.7 ***
Predicted Odds of Being Winner	10.6	1.36	-7.8 ***

**** Significant at 0.001 level*

Sample: Validation Logit Regression

Dependent Variable: Winner or Loser: 1 or 0

	Estimate	Std.Error	z-value
Intercept	-5.8	0.4	14 ***
Predicted Odds of Being Winner	11.3	0.8	14.2 ***

**** Significant at 0.001 level*

Table 1: Do the predicted odds actually help predict the Winners?

When we repeat the exercise over different, consecutive samples and show that we are able to establish a relation between the odds predicted by the ML-driven model and the real odds of catching the Winners, we might have an 'AI-powered' strategy for stock investing.

It is common practice to apply a moving-window sampling to partition the time series data into *Train* and *Test* sub-samples so that the chronological consistency is maintained in the process. Randomized sampling over time does not work with time-series data due to the risk of look-ahead bias. Especially in the investment strategy development practices, researchers run the model-driven portfolio decisions over time with moving samples to demonstrate how the portfolios could have performed if the same decision rules or modeling methods were applied. It is called back-testing. Many academic publications also use the same procedure to validate their predictive modeling. (See Kelly et al. (2022) and Harvey et al. (2019) [19, 25] for a couple of examples.)

To see whether our ML-driven portfolio decision rule could work over time, let's repeat the XGBoost model fitting exercise over consecutive moving samples and construct portfolios according to the predicted odds of catching Winner stocks. As mentioned earlier, the objective is not to make highly accurate predictions of stock returns but to improve the odds for our bets in the gamble. At a point in time, we bet on roughly 150-200 stocks to buy (to go Long in finance lingo) and about the same number of stocks to sell (to go Short in finance lingo) out of about 1000 stocks. Among all those bets, if we can catch a few good ones, and avoid the bad outcomes each time, we can accumulate profits as we repeat the same process over and over.

We let our XGBoost model train over 36-month periods, as shown in Figure 9, then predict the Winner stocks in the consecutive 3-month Test period which is separated from the *Train* sample by +3-month gap to avoid any information leakage. We construct equal-weighted portfolios of stocks that are predicted to be likely to deliver Winner performance (i.e. top 80-percentile in that particular 3-month period) and we build another portfolio with the stocks that are predicted to be the least likely Winners. We calculate the return difference between the predicted Winner and Loser stock baskets for the period up to 2008. The accumulated return trajectory looks like the one shown in Figure 10.

The performance chart looks encouraging again. The AI machinery seems to be able to find a way to improve the odds of our 3-monthly bets on stocks. The evidence on the usefulness of the ML algorithms to guide the future stock return forecasts is accumulating, or it seems so.

Such cumulative return charts of back-tested portfolios are used widely as a historical validation tool in finance. Although it is helpful to run such experiments on historical data sets, the resultant performance charts may not reveal much about model reliability. In our case, for example, where we choose roughly 200

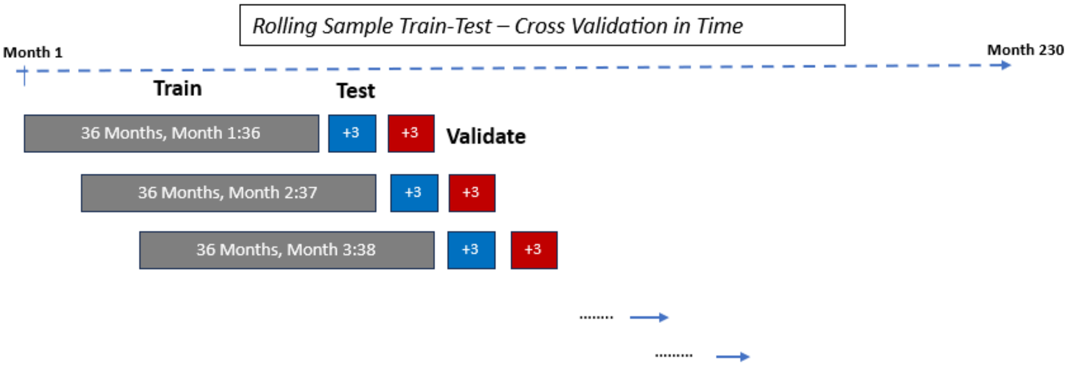


Figure 6. Model fitting with moving samples in time

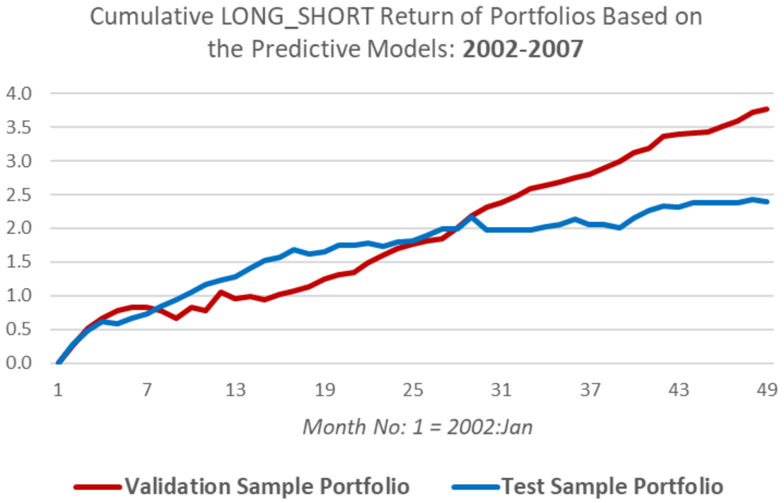


Figure 7. ML-driven model seems to deliver remarkable portfolio performance!

stocks among 1000, we have to acknowledge that there are countless (practically, pretty much infinitely many) portfolio combinations that may be shown to outperform another. It is highly unlikely not to randomly find a lucky portfolio among co many possible combinations.

On the other hand, if we divide our sample into much smaller sub-samples, 100 stocks among the available 1000 to fit our model, for example, the advantage of exploiting large data sets with ML algorithms fade away. Therefore, when we see back-testing exercises that are driven from ML models trained and tested on large data sets, we need to look into the drivers of results carefully to answer the following question: *Is the cumulative performance driven by a coincidental sequence of luck or by the accurate predictions of the model?*

In commercial applications, such as the AI-powered ETF products mention earlier, the questions about the probable sequential luck in their back-tests are completely omitted. Such an inquiry is against the commercial incentives to ride the AI wave of our time. Additionally, academics also tend to rely heavily on back-test results to show some evidence of validation for their models. Those practices are criticized in a growing number of papers such as [13], [14] and [25].

Now, let's make an attempt to shed some light onto the likelihood of 'sequential luck' in our case. We see that the ML-based model is able to help us accumulate positive returns with the historical sample prior to 2008. Are those positive returns driven by the models' successful predictions or are we picking up some lucky draws generated by the complex decision tree models?

In order to answer that question, we can run Logit regressions just like the ones presented earlier. If the 'predicted odds of being a Winner stock' correlates with 'actually being one of the Winner stocks' consistently over sequential samples, then we can build more confidence on the reliability of the data and the methods employed.

In Figure 8, ideally, we would like to observe the z-values pile up in the second quadrant, in and around the blue shaded area. We see that the dots are slightly tilted towards that area, but it is hard to argue for a significant cluster. Actually, if we remove 2-3 outliers from the picture, the chart becomes an evenly spread out scatter centered around zero. That suggests that some luck is involved in upward-trending back-tests.

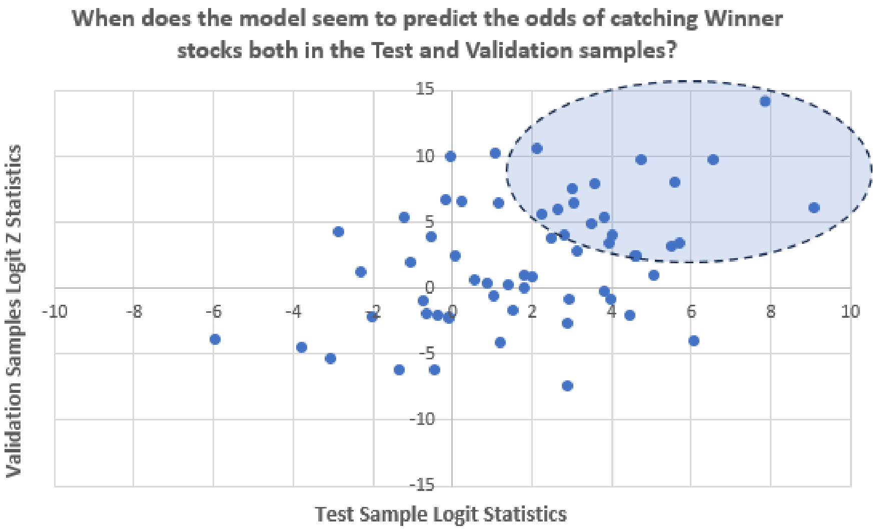


Figure 8. Z-value of Logit regressions of model prediction on real outcomes - Test vs. Validation samples

Random luck should converge to an average of zero success in the long run. You might have some lucky streak from time to time, but it tends to correct over time. When we extend our sample further into the following 10 years, we see that outcome.

In Figure 9, the *Test* sample continues to accumulate some positive return since, during the hyper-parameter tuning and model-selection process, the iterative algorithm uses the Test sample to optimize accuracy. However, when we try to implement that 'optimal model' in the following validation period, we see that the model does not bring any value.

If we were in 2009, for example, and got excited with the back-test results of our smart, AI-powered setup and implemented it as an investment strategy, we would end-up losing great sums of money- just like many other similar strategies do all the time.

The simple case discussed above clearly demonstrates the importance of collecting as many instances of statistical results as possible to gauge the reliability of the models fitted to historical samples. Unfortunately, neither the financial industry nor the academic researchers seem to have the necessary focus on reliability due to the ongoing rush to produce the next interesting statistical machinery that seems to show an instance of predictive success. Many end up reporting their lucky draw with an 'outcome bias'.

Forecasting is not only about predictive accuracy but also about estimation of the model risk. Machine learning models that are over-fitted onto the single sequence of observed history carry substantial reliability risks. However, the trendy labels with suggestive words such as 'learning' and 'intelligence' seem to create some illusion about the models' limited capabilities especially with historical data in social contexts. The luck factor and observation bias hiding behind the complex algorithms is a much bigger problem than it is usually discussed.

3 Conclusion

'Learning' is essentially about figuring things out with experience. AI technologies allow the computers to gain and simulate experience by using large amounts of data. As long as we can define the objective and formulate the related optimization problem, iterations over patterns in large data sets help us distill the

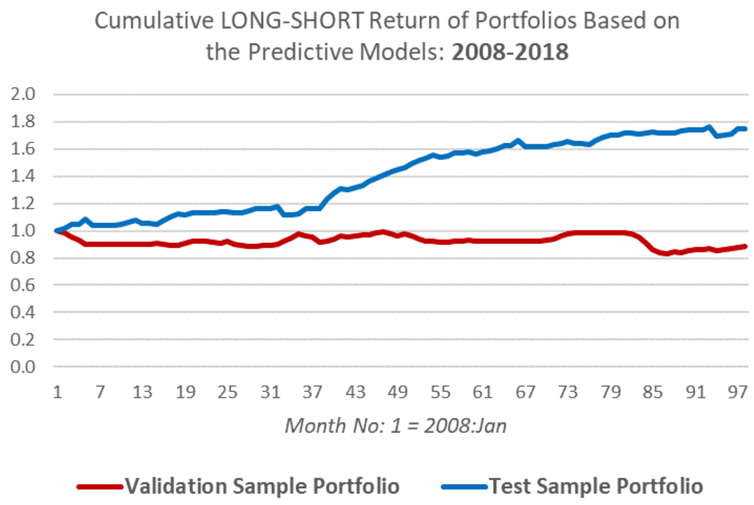


Figure 9. Performance in the post-2008 sample converge to zero

mechanisms that bring practical solutions. That is what we observe in many fields from automation to GO playing. The Boston Dynamics robot dog, Spot, for example, learns to walk over obstacles after processing the data of all previous falls to make improvements on its target.

A historical data set generally does not provide much information on repeated failures. It shows the outcomes captured in a certain set of circumstances that may radically change over time. To adopt AI techniques with such data sets might help us uncover some patterns that occurred in the past, but it would not necessarily yield any reliable predictions for the future.

The researchers in social fields, such as economics and finance, generally assume that an out-of-sample back-test can be used as an evidence of reliability. They present their results without discussing how likely it is to find such a back-test result simply by luck. In some cases, like the one developed in this paper, it may be very easy to find just by construct. The way we design the prediction target, the data we use, and the computational tools we implement might become a powerful combination to produce many statistical flukes. Therefore, while designing forecast models, the researchers need to extend their results into a detailed discussion on reliability.

Given that we are able to discover statistical patterns and validate them with the historical data, how useful should we expect those findings to be in the future?

To answer that question, we need to run empirical experiments to show what the results could have been if we had actually implemented similar models discovered in the past. The analysis and discussions in this paper offer some practical approaches to design such experiments. The results show that the incidental statistical discoveries may crumble easily, no matter how they may look convincing in the past. The finance literature is full of such examples.

As AI-powered applications proliferate many fields in business and academia, it is important to acknowledge the rising risk of statistical deception as a byproduct of careless and lazy model implementations. Extra care and regulation may be needed in the areas where artificial intelligence blends with too much natural human ignorance.

4 Discussion

The complexities related to the implementation of predictive machinery in financial investment and economic policy are actually go far beyond the reliability issues. For example, if large asset managers start implementing investment decision rules based on similar models, and if those models start to trigger correlated decision signals, they might generate self-fulfilling fluctuations in the market. Not only the actions become more predictable but also the models might induce cascades of decisions that are chasing each other.

Cascading actions are a common phenomenon in the financial markets. The AI models carry the risk of amplifying the cascades with automated herding behavior. Then the models drive their own validation success, and demise.

The nature of the problem is quite complex. In social settings, if enough people believe in something, that belief actually becomes the truth. If all believe that AI has huge economic potential and NVIDIA will be the company to benefit from that economic potential, and then invest in the company, NVIDIA stock price surges, pushes down cost of capital and triggers more investment decisions by the company, in a circular manner. AI-driven decision rules can work to build that self-fulfilling circuit. Therefore, the reliability and usefulness the models become rather fluid and stochastic.

An analogy from image recognition models would be as follows: In social settings and markets, if there are enough number of people that believes in a model that says the image is a cat, the image actually becomes a cat, no matter what it was to start with.

Such complexities will probably be the subject of other papers.

Acknowledgements

I thank Tony Guida for his valuable discussion in the AIFI - AI in Finance workshop, and for making the data sets available.

References

[1] Berman R, Sweeney P, Katari S. Machine learning for corporate strategy: An application to strategic decision-making in the pharmaceutical industry. Strategic Management Journal. 2021;42(7):1215-37.

[2] Lee J, Chen Z. Predictive analytics for employee success and retention: A machine learning approach. Journal of Business Research. 2020;116:372-80.

[3] Chen T, Guestrin C. Predicting economic and political stability using big data and machine learning. Political Analysis. 2016;24(3):293-311.

[4] Erel I, Stern LH, Tan C, Weisbach MS. Selecting Directors Using Machine Learning. The Review of Financial Studies. 2021;34(7):3226-64. Available from: https://doi.org/10.1093/rfs/hhaa133.

[5] Ross SA. The Arbitrage Theory of Capital Asset Pricing. Journal of Economic Theory. 1976;13(3):341-60.

[6] Merton RC. An Intertemporal Capital Asset Pricing Model. Econometrica. 1973;41(5):867-87.

[7] Fama EF, French KR. The Cross-Section of Expected Stock Returns. The Journal of Finance. 1992;47(2):427-65.

[8] Fama EF, French KR. Common Risk Factors in the Returns on Stocks and Bonds. Journal of Financial Economics. 1993;33(1):3-56.

[9] Cochrane JH. Presidential Address: Discount Rates. The Journal of Finance. 2011;66(4):1047-108.

[10] Carhart MM. On Persistence in Mutual Fund Performance. The Journal of Finance. 1997;52(1):57-82.

[11] Fama EF, French KR, Carhart MM. Characteristics, Covariances, and Average Returns: 1929 to 1997. The Journal of Finance. 2000;55(1):389-406.

[12] Fama EF, French KR. A Five-Factor Asset Pricing Model. Journal of Financial Economics. 2015;116(1):1-22.

[13] López de Prado M, Bailey DH. Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance. Notices of the American Mathematical Society. 2014;61(5):458-71.

[14] Bailey DH, López de Prado M. Backtest Overfitting. Journal of Computational Finance. 2014;18(2):21-36.

[15] López de Prado M. Why has Factor Investing Failed?: The Role of Specification Errors. SSRN Electronic Journal. 2020.

[16] Harvey CR, co authors. ... and the Cross Section of Returns. Review of Financial Studies. 2016;29(3):580-637.

[17] Harvey CR, Liu Y, Zhu H. Lucky Factors. SSRN Electronic Journal. 2016.

[18] Kelly B, co authors. The Virtue of Complexity in Return Prediction. SSRN Electronic Journal. 2022.

[19] Kelly B, co authors. Factor Models, Machine Learning, and Asset Pricing. NBER Working Paper Series. 2022;(w30599).

[20] Chen AY, Zimmermann T. Open Source Cross-Sectional Asset Pricing. Journal of Finance. 2020;75(3):1539-86.

[21] Chen AY, Velikov M. Zeroing in on the Expected Returns of Anomalies. Journal of Financial Economics. 2021;142(2):679-703.

[22] Belkin M. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. arXiv preprint arXiv:210514368. 2021. Available at arXiv: https://arxiv.org/abs/2105.14368.

[23] Noguer i Alonso M, Srivastava S. The Shape of Performance Curve in Financial Time Series. SSRN Electronic Journal. 2023.

[24] Guida T. Machine Learning for Factor Investing: R Version. Hoboken, NJ: Wiley; 2020.

[25] Harvey CR, co authors. Machine Learning in Finance: The Case of Missing Factors and Alternative Anomalies. SSRN Electronic Journal. 2019.

Appendix A: Data set

The data set on stock returns and attributes is a courtesy of the work by Guida (2020) [24]. The variable descriptions and exploratory data analysis can be found at: https://www.mlfactor.com/data-description.html

All feature variables are scaled and normalised. The details are not included to keep this document in manageable length.

The R codes used in this paper are available by request from the author.

AIPA's International Journal on Artificial Intelligence: Bridging Technology, Society and Policy
ISSN: 3062-097X
Published: 07 December 2024

A Review For Use Of AI and ML Techniques On Nuclear Power Technologies For The Last Decade

Veda Duman Kantarcıoğlu[1]*

[1]**Nuclear Engineering Department, Hacettepe University, ORCID: 0000-0001-6193-8359**

REVIEW

Abstract

Research in the field of nuclear technology increasingly focus on artificial intelligence (AI) and machine learning (ML) techniques to make nuclear power plants easier to operate safer, and more reliable. This review investigates the integration of AI and ML in nuclear power technologies over the past decade. We collected 725 research articles from five leading journals related to nuclear technology, categorizing them into five distinct groups identified by the International Atomic Energy Agency (IAEA) based on their focus research areas. This study aims to investigate the evolution of research topics over the years. We also examined the keywords used within these studies to obtain insights into prevailing trends. Furthermore, we summarized the AI and ML techniques employed across these articles to understand their applications in the nuclear sector. This study demonstrates experience using artificial intelligence-supported methodologies to improve various aspects of nuclear technology and promote innovation in the nuclear industry. Over the last decade, the use of AI and ML in research on nuclear power reactors has significantly increased. In 2018, there was a rapid increase in research articles on AI and ML applications; this trend has increased linearly over the last five years. The groups with the largest share in the published articles are prediction and prognosis, analytics, and optimization, respectively. However, research articles about automation for nuclear power have been increasing significantly in the last five years.

Keywords: artificial intelligence, machine learning, nuclear power plants

1 Introduction

In recent years, nuclear energy has begun to attract attention again as a low-carbon and reliable energy source. On the other hand, the rapid development of artificial intelligence (AI) technologies in recent years and the development of tools that enable the fastest integration of these technologies into different industries have opened the door to new capabilities that researchers can greatly benefit from in their research on nuclear power technologies. By adding new AI capabilities to the point reached so far in the field of nuclear engineering, the opportunity has arisen to optimize reactor design, performance and safety, and to achieve high efficiency and lower maintenance costs. Machine learning (ML) techniques are used to automate routine procedures and tasks. This increases reliability and reduces human or system errors. Predictive applications of AI are increasingly helping to monitor operations and detect anomalies [1].

Nuclear power technology has played a significant role in providing carbon-free energy for decades and has made significant contributions to global efforts to reduce greenhouse gas emissions. Reducing the greenhouse gas emissions globally to reduce the effects of climate change is now a requirement that is accepted all over the world. With the Paris climate agreement, countries have declared their goals and determination in this regard to the whole world. One of the most effective ways to reduce carbon emissions is to reduce the use of fossil fuels in electricity generation. Plans for this goal have also strongly highlighted nuclear power plants in recent years. Nuclear energy offers a low-carbon solution that can be implemented on a large scale within the required time frame and provides clean, reliable, and affordable electricity globally. According to the International Atomic Energy Agency (IAEA), by 2023, nuclear energy will account for approximately 10 percent of global electricity production and a quarter of all low-carbon electricity [2]. Nuclear power reactors with a capacity of approximately 413 gigawatts (GW) operating in 32 countries prevent global emissions of 1.5 gigatonnes (Gt) per year and global gas demand of 180 billion cubic meters (bcm) [3].

Concerns about climate change have accelerated the development of small and modular nuclear reactor

OPEN ACCESS
AIPAJ Vol:1, Issue:1

***Corresponding author**
veda.duman@hacettepe.edu.tr

***Submitted** 21 August 2024*

***Accepted** 28 September 2024*

Citation
Duman Kantarcıoğlu, V. (2024), A Review For Use of AI and ML Techniques On Nuclear Power Technologies For The Last Decade. In AIPA's International Journal on AI: Bridging Technology, Society and Policy. (Vol. 1, Number 1, pp. 41–55). DOI: 10.5281/zenodo.14293646

(SMR) designs in recent years. SMRs are nuclear fission reactors with an electrical power output of less than 300 megawatts (MWe). SMRs provide modular production, factor production, portability, and scalable sustainability [4]. SMRs offer enhanced safety features, simplified installation, and flexibility in operation. They can be built in a factory environment, transported to the field, and installed on-site. This reduces construction times and costs. With the lower accident risks and minimal radiological consequences, SMRs are good choices for carbon free electricity production. AI and ML technologies are frequently used for the development of SMRs.

The nuclear power industry has begun to leverage AI in areas such as automation, design optimization, data analytics, prediction, and insight extraction. According to the IAEA, today AI in Nuclear efforts aim to expand AI technologies from pilot studies to broader applications today [5]. AI applications enable precise monitoring and control of operating nuclear reactors and ML models are widely used to simulate reactor behavior. AI and ML constitute important areas of development in predictive maintenance and monitoring. ML algorithms are used especially in predicting failures. The aim is to analyze sensor data from various components of the nuclear reactor and ensure that maintenance is carried out without experiencing any accidents. These applications reduce operating costs. In addition, AI-supported optimization algorithms are used to improve reactor core design to increase fuel efficiency. Moreover, AI applications are integrated into real-time monitoring and control systems to develop automatic responses to detect anomalies and maintain reactor stability. Techniques such as neural networks and fuzzy logic also excel in managing uncertainties, which are among the most important issues in nuclear engineering calculations. In addition, AI-supported digital twins and virtual simulations provide valuable information for design, testing, and training, while AI-supported models optimize the management and disposal of nuclear waste. As AI and ML continue to develop, their applications in nuclear reactor technologies will also increase.

As AI technologies develop and their use becomes widespread, energy consumption levels are also increasing. New generation nuclear power reactors are seen as an important opportunity to meet these rising energy demands. Today, large technology companies are trying to purchase electricity directly from nuclear power plants [6]. As industries and societies move towards AI-driven systems, the demand for clean, reliable and continuous power sources will increase. With its high energy density and ability to operate independently of weather conditions, nuclear power is in a unique position to meet these needs [7].

This study examined the research articles of the last ten years that applied AI and ML techniques in the field of nuclear energy. Five journals directly related to nuclear technology were selected and 758 research articles were obtained online. Articles related to public perception, public opinion, optimization of electric power generation, and power forecasting were excluded. As a result, 725 research articles on nuclear energy reactors were included in the study. The abstracts of these articles were examined, and the articles were classified into five main groups defined by the IAEA for AI applications in nuclear energy. This analysis identified changes in the focus areas under these categories, and keyword analysis highlighted the most emphasized techniques and the nuclear reactor technologies to which these techniques are related.

1.1 The Evolution of AI and ML Applications in Nuclear Technology

The use of AI and ML in nuclear technology began in the early years of the development of these technologies. The chronological development of the integration of technologies can be summarized as follows [5]:

- 1960-1980: Monitoring reactor operating conditions
- 1980-1990: Expert systems for diagnose anomalies and prediction of failures (Following the Three Mile Island nuclear power plant accident in 1979, many methods, especially safety analyses, began to be developed to systematically structure nuclear safety)
- 1990-2000: ML for predictive maintenance and neural networks to model complex processes.
- 2000-2010: Reactor design and risk management
- 2010-2020: Control systems, operational efficiency, and fuel cycle optimization.
- 2020-: Advanced reactors and fusion power, optimizing operations,improving grid management, and improving cybersecurity, revolutionizing maintenance and safety procedures

Future AI-enabled autonomous systems are likely to revolutionize maintenance, inspection, and overall safety in nuclear operations [8]. As AI technologies develop, their impact on nuclear technology also increases.

1.2 IAEA's Grouping of AI Applications in Nuclear Power

AI falls into logic- or knowledge-based AI and data-driven AI. AI refers to a collection of technologies that produce systems capable of tracking complex problems in ways similar to human logic and reasoning. ML technologies learn how to complete a particular task based on large amounts of data. According to the report of IAEA, the main opportunities for AI to achieve a positive impact on the nuclear power industry can be grouped into the 5 areas: Automation, Optimization, Analytics, Prediction and prognostics and Insights [5].

Studies on automation are basically investigating new system innovations to increase reliability and efficiency by minimizing human intervention in routine tasks. Automating repetitive processes provides easier traceability and faster detection of system errors, as well as leading to the reduction of human errors. Thus, it is possible to speed up the processes during the operation of nuclear power plants. All these developments can provide cost savings as well as increased productivity and more efficient and effective use of human resources. The summary of content for studies on automation as described by IAEA is given below:

- High-pressure tasks in nuclear plants increase human error; data science can automate these processes.
- ML improves complex data analysis and defect detection in inspections.
- ML detects anomalies in control rod drive mechanisms (CRDMs), reducing human evaluation needs.
- Anomaly detection and operator awareness are enhanced with ML and physics-based models.
- Drones and natural language processing (NLP) streamline inspections and decision-making.

Optimization focuses on improving the efficiency and effectiveness of complex operations in all processes from the design, installation, operation and dismantling of nuclear power plants. The main focuses of research in this area are to find the best possible solution within physical and economic constraints by maximizing output, increasing efficiency, minimizing costs, or improving overall performance. Optimization techniques are generally developed to ensure that complex processes in the operation of nuclear power plants continue as smoothly and efficiently as possible. The summary of content for studies on optimization as described by IAEA is given below:

- Data science optimizes inventory management and outage scheduling in nuclear power.
- ML enhances scheduling and radiation mitigation.
- AI improves design processes, safety, and cost efficiency.
- AI balances multiple goals in reactor control.
- AI addresses optimization in in-core fuel management.

Research in the field of analytics aims to improve the quality of existing models and deepen the understanding of the analyzed systems. The importance of both theoretical and experimental studies in the development of nuclear reactor technologies is very great. Since the early years of commercial use of nuclear power reactors, important physical phenomena have been simulated with simulations and experiments that are physically impossible or too costly to be carried out in a computer environment. In this way, data can be produced for neutronic and thermal hydraulic calculations, and again, with sector-specific calculation tools, the behaviors of reactors and all components of the power plant under different conditions are tried to be predicted. All developed analytical tools and techniques help to identify patterns, trends and insights that can be used to improve analytics, decision-making, improve models and optimize system performance by collecting, processing and interpreting data. Hence, predictions can become more accurate to develop better strategies. The summary of content for studies on analytics as described by IAEA is given below:

- AI techniques support long-term research benefits.
- Expedite characterization and validation of materials for new designs.
- Develop quality assurance practices for additively manufactured components.
- Create complex models for improved accuracy in decision-making.
- Enhance model validation and support digital twin applications.

Prediction and prognosis focus on predicting the behavior of systems and their components under different conditions or possible failures, thus improving the planning and execution of maintenance activities. Predicting failures in advance is important for preventing serious nuclear power plant accidents, managing accidents in the most effective way, and mitigating their consequences. By using data-driven models and algorithms, proactive maintenance systems can be established as potential problems can be predicted before they occur. The summary of content for studies on prediction and prognosis as described by IAEA is given below:

- Data science for predicting events and assessing asset conditions.

- Tools for planning maintenance and reducing unexpected downtimes.
- Monitoring operation data for abnormal conditions and timing of inspections.
- Advanced simulation tools not fully utilized by end-users.
- AI addresses prediction challenges with mathematically rigorous algorithms.

Data from operating experience and experiments are used to generate insights. For this purpose, experience from a single reactor can be used, as can data from various reactors around the world. This data is evaluated, and important lessons are learned, and these conclusions are shared with nuclear reactor operators around the world. In this way, it is possible to gain insight into the possible outcomes of some operating processes, examples of good practice, and actions that may have negative consequences. These insights are used to improve operating conditions and reduce errors. The summary of content for studies on insights as described by IAEA is given below:

- Thousands of reactor years of operational experience.
- Extensive libraries of validation experiments.
- Data science technologies for best practices and decision-making.
- AI applications for maintenance record assessments.
- Challenges with language and jargon specificity.

2 Material and Methods

This study investigates the areas of focus and the trend of change in research articles published by researchers in the field of nuclear power in the last ten years. The research articles published in the last ten years by five important journals in which research articles in the field of nuclear energy and nuclear technology are published were included in the scope of the study. The following two search terms were used in the searches: "nuclear power" + "AI", "nuclear power" + "machine learning". A total of 758 research articles were found. Articles related to public perception, public opinion, optimization of electrical power generation, and power forecasting were excluded.725 of research papers on nuclear power reactors were included in the analysis.

Table 1. Number of Reviewed Research Articles

Journal Name	Number of Papers Accessed	Number of Papers Reviewed
Nuclear Engineering and Technology	142	126
Annals of Nuclear Energy	241	232
Nuclear Engineering and Design	184	179
Progress in Nuclear Energy	139	138
Journal of Nuclear Science and Technology	52	50
TOTAL	758	725

The selected research articles were classified according to the IAEA grouping explained in previous section. In this way, it was tried to determine the trends in the subjects of the researchers' research in the last 10 years. In the second part of the study, the keywords used by the researchers in the articles were analyzed and a second data was tried to be obtained for the areas where the researchers focused. The second information obtained from the keywords is the AI and ML techniques used. In addition, the articles were examined and the tools used were determined. The techniques and tools determined to be used are given in the following section.

3 RESULTS

3.1 The Outstanding AI And ML Applications in Nuclear Power Research For The Last Decade

As a result of keyword analysis of the research articles included in this review, the word cloud given in Figure 1 was created. This cloud is an important visual in terms of seeing the most common purposes for which AI and ML applications are used in nuclear power plants. A significant portion of the reviewed articles focused on fault diagnosis, and topics such as accident analysis, error analysis to prevent accidents, fault detection, anomaly detection, and control systems also came to the fore. As a result, it can be said that analyses related to nuclear safety are clearly visible in the studies. Analyses such as uncertainty and sensitivity analyses also clearly show themselves in the studies.

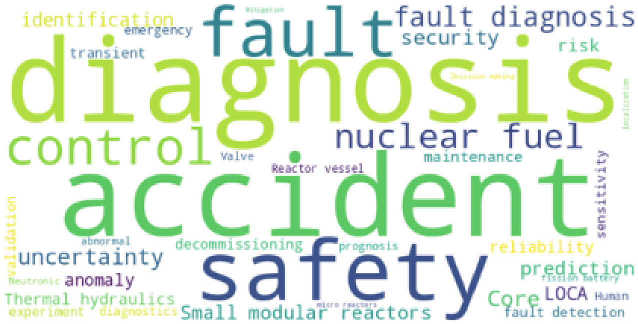


Figure 1. Word-cloud of nuclear research areas for which AI and ML techniques are used

Among the reactor types that stand out in the reviewed articles regarding the use of AI and ML techniques in the development of nuclear power technologies are Pressurized Water Reactor (PWR), Boiling Water Reactor (BWR), High-Temperature Gas-Cooled Reactor (HTGR), Sodium Fast Reactor, Fast Breeder Reactor, Small Modular Reactor (SMR), Advanced Gas-Cooled Reactor (AGR), Pressurized Heavy Water Reactor (PHWR) [9] [10] [11] [12] [13] [14] [15] [16]. Important components of nuclear power reactors, Steam Generators, Reactor Core, Fuel Assemblies, Containment Structures and Cooling Systems, take up an important place among the reviewed studies. [17] [18] [19] [20]

Studies on fuel types, core array optimization, fuel lattice design, fuel performance modeling, spent fuel, etc. are among the studies conducted on nuclear fuels. [21] [22] [23] [24]

Another important area where AI and ML techniques are used is nuclear safety and risk management. In the reviewed articles, it was observed that they focused on Probabilistic Risk Assessment (PRA), Safety Analysis, Accident Diagnosis, Fault Diagnosis and Identification, Safety Margin Analysis, Uncertainty Quantification, Risk-Based Approach, Safety Critical Systems. Fault diagnosis has been the most focused area in the last 5 years. [25] [26] [27] [28] [29] [30] [31] [32]

Many topics such as Loss of Coolant Accident (LOCA), Large-Break Loss-of-Coolant Accident (LBLOCA), Station Blackout, Severe Accident Analysis, Post-CHF Analysis are the topics that researchers focus on regarding accident situations in nuclear power plants [33] [34][35] [36]. In addition, many studies have been encountered regarding the safety systems of nuclear reactors. Some of these systems are Reactor Core Isolation Cooling (RCIC) System, Passive Systems, Emergency Response Systems [37] [38]. Many topics related to nuclear safety such as effective management of emergencies, radiation protection, Dosimetry, Environmental Radioactivity, Radiation Shielding, Radioactive Effluents, Atmospheric Dispersion attract the attention of researchers. [39] [40] [41]

Numerous studies have been encountered regarding the Thermal Hydraulics field, which is an important component in the efficient and safe operation of nuclear reactors. It has been observed that the areas of particular focus in this regard are Critical Heat Flux (CHF), Thermal Stratification, Flow Regime, Heat Transfer and Flow Correlations, Thermal Mixing and Stratification, Heat Exchange Coefficient. [42] [43] [44] [45]

It is understood that AI and ML applications have also begun to be widely used in Decommissioning and Waste Management. It is seen that the research in this field focuses on topics such as Decommissioning and Dismantling Processes, Waste Management, Radioactive Waste Disposal, Radioactive Waste Repackaged Drums and In-Situ Decommissioning. [46] [47] [48] [49]

One of the areas where AI and ML applications are most widely used in nuclear power technologies is Instrumentation and Control. In this study, topics such as Digital Instrumentation and Control (I and C), Control Room Radiological Habitability, Automatic Control Systems, Feedback Control, Voltage Regulator and Turbine Speed Control are among the research topics. [50] [51] [52]

Modeling and simulation are of vital importance in the development of nuclear power technologies. Important decisions in processes such as reactor designs, approval of designs, and licensing are made based on the results of analyses made with modeling and simulation tools. In the analyses conducted, it was seen that AI and ML applications were frequently used in this field in studies conducted in the last 10 years. Topics such as Computational Fluid Dynamics (CFD), System-Level Modeling and Simulation, Multi-Physics Simulations, Physics-Based Modeling and Data-Driven Modeling have come to the fore [53] [54] [55]. The increasing interest in new technologies such as Digital Twin Technology, High Fidelity Simulation, Exascale Computing and Augmented Reality for Educational Purposes is also evident in the reviewed papers. [56] [57] [58]

3.2 The Outstanding AI And ML Techniques In Nuclear Power Research For The Last Decade

As a result of keyword analysis for AI and ML techniques in nuclear power research, the word cloud given in Figure 2 was obtained. This cloud shows the most common AI and ML techniques used for research in nuclear power plants. The most prominent techniques in the cloud include neural networks and their varieties. Deep neural networks, convolutional neural networks (CNNs), and recurrent neural networks (RNNs) are some of them. Neural networks are widely used for fault diagnosis, predictive maintenance, and real-time monitoring in nuclear systems [59] [60] [61] [62]. Generative Models and Deep Belief Networks are applied to develop predictive maintenance systems and improve fault diagnosis [63]. In addition to experimental data, they can generate new data samples to work with larger data sets. They can learn complex representations from large data sets. Hybrid and Ensemble Methods such as CFD-ANN in the cloud are also prominent. These are developed by combining the strengths of different models to improve the prediction performance. [64]. Time Series Analysis Techniques such as LSTM (Long Short-Term Memory) networks and time series deep learning are used to analyze dynamic data that are of great importance in critical processes such as reactor transitions and safety analysis scenarios [65]. They help in predicting and diagnosing time-dependent events.

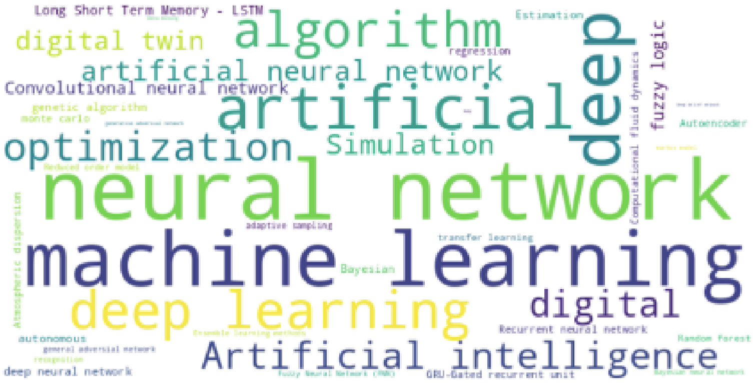


Figure 2. Word-cloud of nuclear research areas for which AI and ML techniques are used.

These techniques facilitate anomaly detection, system prognosis, and improved decision-making by leveraging large datasets and complex patterns. Techniques such as random forests, support vector machines (SVM), and genetic algorithms are applied for tasks like fault detection, optimization of reactor designs, and analysis of reactor behavior [66] [67] [68]. These algorithms are utilized for handling imbalanced datasets, feature selection, and predictive modeling. Genetic algorithms, ant colony optimization, and other optimization techniques are applied for reactor fuel design, control system optimization, and multi-objective problem-solving. These algorithms help in finding optimal solutions for complex reactor configurations and operations. Reinforcement Learning and Deep Reinforcement Learning techniques are used for operator support, diagnosis, and control automation [69] [70]. They help in optimizing reactor operations and improving safety functions by learning from interactions with the environment and adapting control strategies accordingly. Algorithms for pattern recognition are employed in digital neutron spectroscopy and signal processing [71] [72]. These methods help in identifying and analyzing complex patterns within nuclear data, enhancing safety and security measures and monitoring capabilities. Bayesian Networks and Bayesian Machine Learning techniques are used for probabilistic risk assessment, safety analysis, and updating models with new data [73] [74] [75] [76].

AI methodologies such as Principal Component Analysis (PCA), Autocoders, and Graph Convolutional Networks are used for data dimensionality reduction, feature extraction, and pattern recognition. [77] [78] [79] [80] [81] [82]. Deep learning models and reinforcement learning are used to optimize and improve system performance and fault detection. [83] [84] [85] [86][87]. Various optimization methods, including genetic algorithms, ant colony systems, and multi-objective optimization, are applied to improve reactor designs, fuel management, and operational efficiencies [88] [89] [90] [91]. These techniques help improve processes and system configurations for better performance and safety. Dimensionality Reduction and Data Visualization Methods like principal component analysis (PCA), autoencoders, and singular value decomposition (SVD) are used for reducing the complexity of data, enhancing feature extraction, and visualizing high-dimensional data for better analysis and interpretation [92].

Studies investigating the use of automated systems for automation, monitoring, control, and diagnosis in nuclear power plants are increasingly concentrated. Tools such as Hardware in the Loop (HIL) simulations and programmable logic controllers (PLCs) are increasingly important to automate processes and increase operational reliability [93] [94].Analytical tools and methods are used to extract actionable insights from large data sets. Techniques such as statistical analysis, data visualization, and pattern recognition help understand system behavior and predict potential problems. [95] [96] [97] [71].

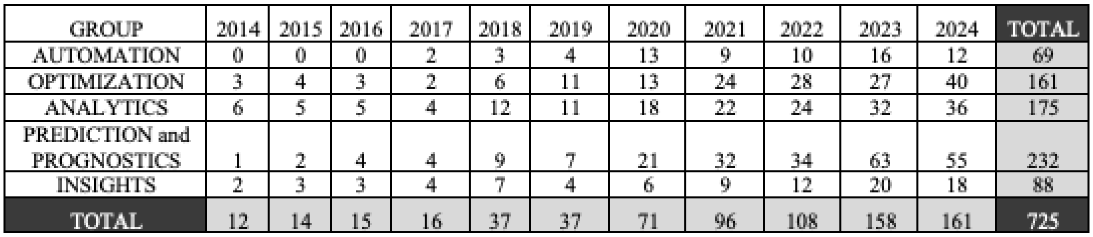
In recent years, predictive maintenance and prognostics have been key areas where analytics provide insights into equipment health and system reliability. Techniques such as Bayesian Neural Networks and time series analysis help predict failures and plan maintenance. Safety-related studies are also among the main areas of interest for applications in the field of AI and ML. These studies generally focus on probabilistic risk assessments, accident analysis, and fault diagnosis to improve nuclear power plant safety. Tools such as MELCOR and RELAP5/SCDAPSIM are widely used nuclear engineering software for safety analysis and uncertainty quantification [98] [99]. Methods such as risk-based analysis, sensitivity analysis, and hazard analysis are used to assess and reduce risks associated with the operation of nuclear power plants and radioactive waste management [100].

The research articles reviewed focus on innovative technologies such as digital twin technology and augmented reality to simulate and visualize complex systems, improve understanding, and increase operational control [101] [102].These technologies provide a virtual representation of physical systems for better monitoring and decision-making. Studies on AI and ML applications have also focused on radiation protection, radioactive waste management, and environmental impact assessments. It is understood that radiation dose, distribution, and waste processing assessment techniques are considered crucial to ensure safety and compliance [103] [104] [105].

3.3 Trends In Focus Of AI And ML Use In Nuclear Power Research For The Last Decade

In this study, a total of 725 articles were distributed into groups using the grouping approach made by the IAEA on the use of AI and ML in the nuclear power field. As a result of this study, the distribution given in Table 2 was obtained. Between 2014 and 2024, the annual number of published research papers in this field has increased by 13 times. Additionally, in parallel with the rapid advancements in AI technologies worldwide over the past five years, the use of AI and ML techniques in nuclear power research has also become more widespread.

Table 2. Grouping of the Reviewed Articles



GROUP	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	TOTAL
AUTOMATION	0	0	0	2	3	4	13	9	10	16	12	69
OPTIMIZATION	3	4	3	2	6	11	13	24	28	27	40	161
ANALYTICS	6	5	5	4	12	11	18	22	24	32	36	175
PREDICTION and PROGNOSTICS	1	2	4	4	9	7	21	32	34	63	55	232
INSIGHTS	2	3	3	4	7	4	6	9	12	20	18	88
TOTAL	12	14	15	16	37	37	71	96	108	158	161	725

According to Figure 3, 32 percent of the research in the last 10 years is in the field of prediction and prognostics, 24 percent in analytics, 22 percent in optimization, 12 percent in insights and 10 percent in automation.

In the graph given in Figure 4, it can be observed in which groups researchers use AI and ML techniques more in nuclear power technologies and the change in this focus point according to the years. A significant increase is observed in the number of articles using AI and ML techniques in 2018 and after. 668 of the 725 articles examined were published in 2018 and after. The total number of articles has increased significantly over the years and second sharp rise was in 2020 and onwards.

This trend indicates a growing interest and emphasis on AI and ML techniques in the field, likely driven by technological advancements and increased computational capabilities. The peak in 2023 and 2024 shows the highest level of research activity, suggesting that these technologies are becoming increasingly critical in recent applications.

Automation stands out with a notable increase, especially from 2018 onwards. This surge reflects the industry's focus on automating processes to enhance efficiency and safety, particularly in complex environments such as nuclear power plants. The consistent rise indicates that automation remains a priority area for research and development. Prediction and Prognostics is the dominant group as an emerging focus

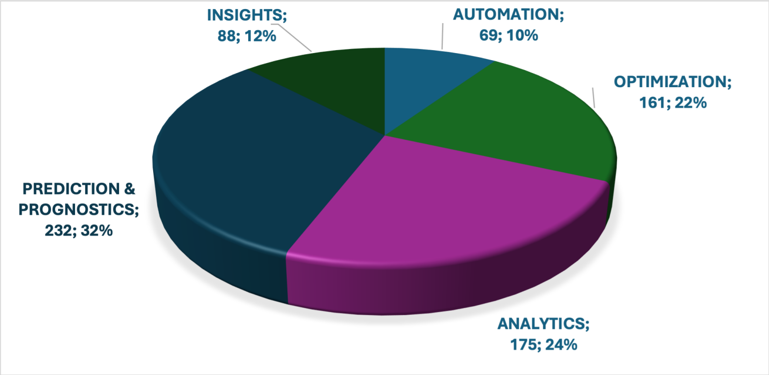


Figure 3. Distribution of Articles Due to the Groups.

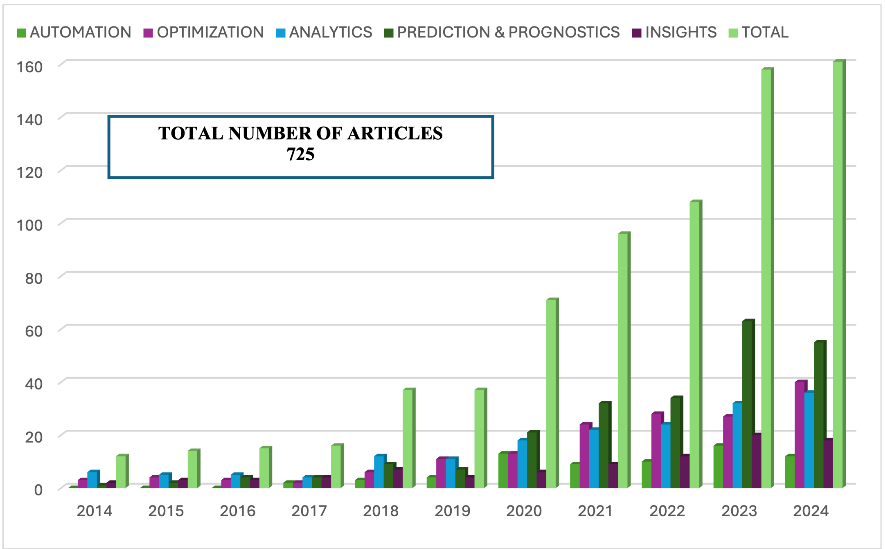


Figure 4. Evolving Research Trends in AI and ML Applications in Nuclear Technology.

area. Prediction and Prognostics and Optimization have seen marked growth since 2020, indicating a shift towards more sophisticated and predictive maintenance strategies. This trend suggests that researchers are increasingly exploring ways to predict equipment failures and optimize operations to prevent unplanned downtime and enhance safety protocols. The categories of Analytics and Insights have shown steady growth over the years, underscoring their importance in data interpretation and decision-making. The gradual increase in articles highlights a consistent effort to extract actionable insights from data, supporting informed decisions in complex systems. The sharp increases in recent years, particularly in advanced techniques like Prognostics and Optimization, reflect a maturation of AI and ML applications. This growth is likely driven by the successful implementation of these technologies in real-world scenarios, leading to increased confidence and further exploration in the field. Overall, the trends suggest a dynamic evolution in research priorities, with a clear shift towards automation, predictive analysis, and optimization, indicating that AI and ML are becoming integral to the future of the nuclear power industry.

4 CONCLUSION

This study provides a comprehensive review on the application of Artificial Intelligence (AI) and Machine Learning (ML) in nuclear power technologies over the past decade. By analyzing 725 research articles from five leading nuclear technology journals, key trends and developments were identified in automation, optimization, analytics, prediction and prognosis, and insights. The findings show that AI and ML are increasingly integrated into various aspects of nuclear power: A graphical analysis of the growth in publications over time highlights the increasing importance of AI and ML in nuclear power. The significant increase in research output since 2018 demonstrates the rapid adoption of these technologies, particularly in areas such as forecasting and prognosis, automation, analytics, and optimization; collectively, these constitute a significant portion of the reviewed literature. The summary of this review is as follows:

- Over the years, the use of AI and ML in research on nuclear power reactors has increased.
- In 2018, there was a significant increase in research articles with AI and ML applications. This trend has increased linearly over the last 5 years.
- The groups with the largest share in the published articles in the last 10 years are prediction and prognosis, analytics and optimization, respectively.
- On the other hand, a significant increase in automation research has been observed in the last 5 years.

As a result, AI and ML are becoming an integral part of the advancement of nuclear energy technologies and providing innovative solutions to complex challenges. However, the review also highlighted the lack of standardized benchmarking practices that limit the wider adoption and validation of AI and ML applications in nuclear engineering. Defining these benchmarks is of great importance to ensure the practical application of AI models in real-world nuclear scenarios. Continued research and development, along with the establishment of standardized benchmarks, will be essential to safely and effectively integrate these technologies and drive the future of the nuclear industry towards greater safety, efficiency and innovation.

References

[1] Vlasov A, Barbarino M. Seven Ways AI Will Change Nuclear Science and Technology. https://wwwiaeaorg/newscenter/news/seven-ways-ai-will-change-nuclear-science-and-technology. 2022.

[2] IAEA. Nuclear Power Reactors in the World. IAEA Reference Data Series No 2. 2024.

[3] IEA. A new dawn for nuclear energy? https://wwwieaorg/energy-system/electricity/nuclear-power. 2024.

[4] OECD-NEA. The NEA Small Modular Reactor (SMR) Strategy. https://wwwoecd-neaorg/jcms/pl-26297/the-nea-small-modular-reactor-smr-strategy. 2024.

[5] IAEA. Artificial Intelligence for Accelerating Nuclear Applications. Science and Technology, Non-serial Publications. 2022.

[6] Hiller J, Herrera S. Tech Industry Wants to Lock Up Nuclear Power for AI. https://wwwwsjcom/business/energy-oil/tech-industry-wants-to-lock-up-nuclear-power-for-ai-6cb75316. 2024.

[7] Aldrete B, Ward J, Pandise E. The AI industry is pushing a nuclear power revival — partly to fuel itself. https://wwwwsjcom/business/energy-oil/tech-industry-wants-to-lock-up-nuclear-power-for-ai-6cb75316. 2024.

[8] Banafa A. Nuclear AI: Pioneering the Future of Nuclear Technology. https://wwwbbvaopenmindcom/en/technology/artificial-intelligence/nuclear-ai-pioneering-the-future-of-nuclear-technology/. 2023.

[9] Qian H, Chen G, Li L, Zhang L, Yin X, Zhang H, et al. Development of supporting platform for the fine flow characteristics of reactor core. Nuclear Engineering and Technology. 2024;56(5):1687-97.

[10] Espinosa-Paredes G, Molina-Tenorio Y, Prieto-Guerrero A, Olvera-Guerrero OA. Linear or non-linear stability monitor in BWRs? Introducing a new non-linear monitor based on the fractal spectrum. Nuclear Engineering and Design. 2023;415.

[11] Bartnik R, Hnydiuk-Stefan A, Skomudeki W. Methodology for thermodynamic and economic analysis of hierarchical dual-cycle gas-gas nuclear power and CHP plants with high-temperature reactors and helium as the circulating medium. Progress in Nuclear Energy. 2023;158.

[12] Byun H, Gil Lee H, Kyu Kim B, Dong Song G, Lee B. Defect monitoring system of the internal structures of a sodium fast reactor using an artificial intelligence model. Nuclear Engineering and Technology. 2024;Article in press.

[13] Manimaran M, Shanmugam A, Parimalam P, Murali N, Satya Murty SAV. Software development methodology for computer based I and C systems of prototype fast breeder reactor. Nuclear Engineering and Design. 2015;292:46-56.

[14] Zhang B, Zhu H, Cheng S, Ma H. Sensor anomaly detection for small modular reactors utilizing improved autoencoder. Nuclear Engineering and Design. 2024;417.

[15] M West G, Wallace CJ, McArthur SDJ. Combining models of behaviour with operational data to provide enhanced condition monitoring of AGR cores. Nuclear Engineering and Design. 2014;272.

[16] Rani J, Roy AA, Kodamana H, Tamboli PK. Fault detection of pressurized heavy water nuclear reactors with steady state and dynamic characteristics using data-driven techniques. Progress in Nuclear Energy. 2023;156.

[17] Kazuyuki D, Hori T, Perrin S. Crack Depth Estimation of Non-Magnetic Material by Convolutional Neural Network Analysis of Eddy Current Testing Signal. Journal of Nuclear Science and Technology. 2019;57(4):401-7.

[18] Li X, Zheng Y, Du X, Xiao B. A new surrogate method for the neutron kinetics calculation of nuclear reactor core transients. Nuclear Engineering and Design. 2024;56(9):3571-84.

[19] Li W, Ding P, Xia W, Chen S, Yu F, Duan C, et al. Artificial neural network reconstructs core power distribution. Nuclear Engineering and Technology. 2022;54(2).

[20] Ruan F, Chen CH, Cheng Y, Wang JY, Chen LW. Study on evaluation method for nuclear emergency rescue measures at containment vessel. Annals of Nuclear Energy. 2021;151.

[21] Emily H Kwapis KCH Hongcheng Liu. Tracking of individual TRISO-fueled pebbles through the application of X-ray imaging with deep metric learning. Progress in Nuclear Energy. 2021;140.

[22] Abu Saleem R, Radaideh MI, Kozlowski T. Application of deep neural networks for high-dimensional large BWR core neutronics,. Nuclear Engineering and Technology. 2020;52(12).

[23] Montes-Tadeo JL, Perusquía-del Cueto R, Pelta DA, François JL, Ortiz-Servin JJ, Martín-del Campo C, et al. A hybrid system for optimizing enrichment and gadolinia distributions in BWR fuel lattices. Progress in Nuclear Energy. 2020;19.

[24] Ortiz-Servin JJ, Cadenas JM, Pelta DA, Castillo A, Montes-Tadeo JL. Nuclear fuel lattice performance analysis by data mining techniques. Annals of Nuclear Energy. 2015;80:236-47.

[25] QJones HR, Mu T, Kudawoo D, Brown G, Martinuzzi P, McLachlan N. A surrogate machine learning model for advanced gas-cooled reactor graphite core safety analysis, , Volume 395, 2022. Nuclear Engineering and Design. 2022;395.

[26] D'Onorio M, Glingler T, Molinari M, Maccari P, Mascari F, Mandelli D, et al. Nuclear safety Enhanced: A Deep dive into current and future RAVEN applications. Nuclear Engineering and Design. 2024;427.

[27] Li Z, Sun J, Tong J, Sui Z, Gang L. An accident diagnosis algorithm for HTR-PM based on deep learning methods. Progress in Nuclear Energy. 2019;115:140-50.

[28] Tan H, Guo Z, Feng Q, Zhao H, Wu Y Haoand Yu. The application of time series deep learning model to the fast prediction of parameters in the MSLB accident. Progress in Nuclear Energy. 2024;176.

[29] Kobayashi K, Kumar D, Alam SB. AI-driven non-intrusive uncertainty quantification of advanced nuclear fuels for digital twin-enabling technology. Progress in Nuclear Energy. 2024;272.

[30] Sahin E, Lattimer B, Allaf MA, Duarte JP. Uncertainty Quantification of Unconfined Spill Fire Data by Coupling Monte Carlo and Artificial Neural Networks. Journal of Nuclear Science and Technology. 2024;417.

[31] Wang Z, Xia H, Zhu S, Peng B, Zhang J, Jiang Y, et al. Combining models of behaviour with operational data to provide enhanced condition monitoring of AGR cores. Journal of Nuclear Science and Technology. 2021;59(1):67-77.

[32] Miki D, Demachi K. Fault detection of pressurized heavy water nuclear reactors with steady state and dynamic characteristics using data-driven techniques. Journal of Nuclear Science and Technology. 2020;57(9):1091-100.

[33] Park SH, Kim DS, Kim JH, Na MG. Prediction Of The Reactor Vessel Water Level Using Fuzzy Neural Networks In Severe Accident Circumstances Of NPPs. Nuclear Engineering and Technology. 2014;46(3):373-80.

[34] Choi GP, Yoo KH, Back JH, Na MG. Estimation of LOCA Break Size Using Cascaded Fuzzy Neural Networks. Nuclear Engineering and Technology. 2017;49(3):495-503.

[35] Lee JH, Yilmaz A, Denning R, Aldemir T. An online operator support tool for severe accident management in nuclear power plants using dynamic event trees and deep learning. Annals of Nuclear Energy. 2020;146.

[36] Song J, Kim S. A machine learning diagnosis of the severe accident progression. Nuclear Engineering and Design. 2024;416.

[37] Hawila MA, Kirkland KV. Turbopump scaling analysis and similarity level estimation for Texas A and M university RCIC system experimental test facility. Progress in Nuclear Energy. 2019;113.

[38] Huang Z, Miao H, Lind M, Zhang X, Wu J. Quantifying performance of passive systems in an integrated small modular reactor under uncertainties using multilevel flow modelling and stochastic collocation method. Progress in Nuclear Energy. 2022;149.

[39] Hvala N, Mlakar P, Grašič B, Božnar MZ, Kocijan J, Perne M. Surrogate tree ensemble model representing 2D population doses over complex terrain in the event of a radiological release into the air. Progress in Nuclear Energy. 2023;158.

[40] Sáez-Muñoz M, Cerezo A, Prieto E, Salvadó M, Hernandez IV, Duch MA, et al. Recent radiation protection activities related to nuclear facilities on the Iberian Peninsula,. Nuclear Engineering and Design. 2024;417.

[41] Alrammah IA. Analysis of nuclear accident scenarios and emergency planning zones for a proposed Advanced Power Reactor 1400 (APR1400). Nuclear Engineering and Design. 2023;407.

[42] Hedayat A. Developing a robust and flexible smart tool to predict a full range Critical Heat Flux (CHF) in different LWRs by using deep learning Artificial Neural Networks (ANN) via parallel multi-processing. Progress in Nuclear Energy. 2021;142.

[43] Sun X, Zhou K, Han X, Song K, Shi S, Yu W, et al. Prediction of time-varying inner wall temperature of surge lines by a dynamic neural network. Nuclear Engineering and Design. 2021;383.

[44] Breitenmoser D, Manera A, Prasser HM, Adams R, Petrov V. High-resolution high-speed void fraction measurements in helically coiled tubes using X-ray radiography. Nuclear Engineering and Technology. 2021;373.

[45] Guillen DP, Anderson N, Krome C, Boza R, Griffel LM, Zouabe J, et al. A RELAP5-3D/LSTM model for the analysis of drywell cooling fan failure. Progress in Nuclear Energy. 2020;130.

[46] Hume S, West G, Dobie G. A framework for capturing and representing the process to classify nuclear waste and informing where processes can be automated. Progress in Nuclear Energy. 2024;170.

[47] Invernizzi DC, Locatelli G, Brookes NJ. How benchmarking can support the selection, planning and delivery of nuclear decommissioning projects. Progress in Nuclear Energy. 2017;99:155-64.

[48] Kim SI, Lee HY, Song JS. A study on characteristics and internal exposure evaluation of radioactive aerosols during pipe cutting in decommissioning of nuclear power plant. Nuclear Engineering and Technology. 2018;50(7):1088-98.

[49] Slimák A, Nečas V. Melting of contaminated metallic materials in the process of the decommissioning of nuclear power plants. Progress in Nuclear Energy. 2016;92:29-39.

[50] Yockey P, Erickson A, Spirito C. Cyber threat assessment of machine learning driven autonomous control systems of nuclear power plants. Progress in Nuclear Energy. 2023;166.

[51] Tacke J, Borrelli RA, Roberson D. Advanced frequency-domain compensator design for subsystems within a nuclear generating station. Progress in Nuclear Energy. 2021;140.

[52] Yu J, Wilson JC, Dave AJ, Sun K, Forget B, Phillips B. Experimental demonstration of a data-driven control system for subcritical nuclear facility. Progress in Nuclear Energy. 2024;168.

[53] Oh C, Kim DH, Ik LJ. Application of data driven modeling and sensitivity analysis of constitutive equations for improving nuclear power plant safety analysis code. Nuclear Engineering and Technology. 2023;55(1).

[54] Yang J, Sui X, Huang Y, Zhao L, Liu M. Application of deep neural networks for high-dimensional large BWR core neutronics,. Annals of Nuclear Energy. 2022;179.

[55] Bao H, Dinh NT, Lane JW, Youngblood RW. A data-driven framework for error estimation and mesh-model optimization in system-level thermal-hydraulic simulation. Nuclear Engineering and Design. 2019;349.

[56] Park HS, Jang SC, Kang IS, Lee DJ, Kim JG, Lee JW. A detailed design for a radioactive waste safety management system using ICT technologies. Progress in Nuclear Energy. 2022;149.

[57] Yan M, Ma Z, Pan L, Liu W, He Q, Zhang R, et al. An evaluation of critical heat flux prediction methods for the upward flow in a vertical narrow rectangular channel. Progress in Nuclear Energy. 2021;140.

[58] Harazono Y, Ishii H, Shimoda H, Taruta Y, Kouda Y. Development of AR-based scanning support system for 3D model reconstruction of work sites. Journal of Nuclear Science and Technology. 2022;59(7):934–948.

[59] Jie M, Qiao P, Gang Z, Panhui C, Minghui L. Fault diagnosis method for Small modular reactor based on transfer learning and an improved DCNN model. Nuclear Engineering and Design. 2024;417.

[60] Liu J, Zhang Q, Macián-Juan R. Enhancing interpretability in neural networks for nuclear power plant fault diagnosis: A comprehensive analysis and improvement approach. Progress in Nuclear Energy. 2024;174.

[61] Koo YD, An YJ, Kim CH, Na MG. Nuclear reactor vessel water level prediction during severe accidents using deep neural networks. Nuclear Engineering and Technology. 2019;51(3):723-30.

[62] Choi J, Lee SJ. RNN-based integrated system for real-time sensor fault detection and fault-informed accident diagnosis in nuclear power plant accidents. Nuclear Engineering and Technology. 2023;55(3):814-26.

[63] Liu Yk, Zhou W, Ayodeji A, Zhou Xq, Peng Mj, Chao N. A multi-layer approach to DN 50 electric valve fault diagnosis using shallow-deep intelligent models. Nuclear Engineering and Technology. 2021;53(1):148-63.

[64] Ding J, Liu Q, Ke J, Deng M, Yu G, Liang Y. Development of a hybrid CFD-ANN method with multi-objective optimization for airfoil-finned PCHE used in Gen-IV nuclear systems. Progress in Nuclear Energy. 2024;175.

[65] Song J, Kim S. A machine learning informed prediction of severe accident progressions in nuclear power plants. Journal of Nuclear Science and Technology. 2024;56(6):2266-73.

[66] Gohel HA, Upadhyay H, Lagos L, Cooper K, Sanzetenea A. Predictive maintenance architecture development for nuclear infrastructure using machine learning. Nuclear Engineering and Technology. 2020;53(7):1436-42.

[67] Dongliang M, Yi L, Tao Z, Yanping H. Research on prediction and analysis of supercritical water heat transfer coefficient based on support vector machine,. Nuclear Engineering and Technology. 2023;55(11).

[68] Kim W, Lim C, Chai J. Study on evaluation method for nuclear emergency rescue measures at containment vessel. Nuclear Engineering and Technology. 2020;52(6).

[69] Radaideh MI, Shirvan K. PESA: Prioritized experience replay for parallel hybrid evolutionary and swarm algorithms - Application to nuclear fuel. Nuclear Engineering and Technology. 2022;54(10).

[70] Zhong X, Zhang L, Ban H. Deep reinforcement learning for class imbalance fault diagnosis of equipment in nuclear power plants. Annals of Nuclear Energy. 2023;184.

[71] El-Tokhy MS. Digital inspection approach of overlapped peaks due to high counting rates in neutron spectroscopy. Progress in Nuclear Energy. 2021;137.

[72] Lee H, Yu K, Kim S. Discrimination model using denoising autoencoder-based majority vote classification for reducing false alarm rate. Nuclear Engineering and Technology. 2023;55(10):3716-24.

[73] Katayama Y, Ohtori Y, Sakai T, Muta H. Development of supporting platform for the fine flow characteristics of reactor core. Journal of Nuclear Science and Technology. 2021;58(11):1220-34.

[74] Kadowaki M, Nagai H, Yoshida T, Terada H, Tsuduki K, Sawa H. Application of Bayesian Machine Learning for Estimation of Uncertainty in Forecasted Plume Directions by Atmospheric Dispersion Simulations. Journal of Nuclear Science and Technology. 2023;60(10):1194–1207.

[75] Nguyen A Tran Canh Hai ad Diab. Using machine learning to forecast and assess the uncertainty in the response of a typical PWR undergoing a steam generator tube rupture accident. Nuclear Engineering and Technology. 2023;55(9).

[76] Lee GG, Lee BS, Kim MC, Kim JM. Determining the adjusting bias in reactor pressure vessel embrittlement trend curve using Bayesian multilevel modelling. Nuclear Engineering and Technology. 2023;55(8).

[77] Mendoza M, Tsvetkov PV. An intelligent fault detection and diagnosis monitoring system for reactor operational resilience: Unknown fault detection. Progress in Nuclear Energy. 2024;171.

[78] Kim H, Kim J. Long-term prediction of safety parameters with uncertainty estimation in emergency situations at nuclear power plants. Nuclear Engineering and Technology. 2023;55(5):1630-43.

[79] Oh SW, Park JH, Jo HS, Na MG. Combining models of behaviour with operational data to provide enhanced condition monitoring of AGR cores. Nuclear Engineering and Design. 2014;272.

[80] Jo HK, Kim SH, Kim CL. Proposal of a new method for learning of diesel generator sounds and detecting abnormal sounds using an unsupervised deep learning algorithm. Nuclear Engineering and Technology. 2023;55(2):506-15.

[81] Wu M, Liu X, Gui N, Yang X, Tu J, Jiang S, et al. Prediction of the remaining time and time interval of pebbles in pebble bed HTGRs aided by CNN via DEM datasets. Nuclear Engineering and Technology. 2023;55(1):339-52.

[82] Liu J, Yang X, Macián-Juan R, Kosuch N. A novel transfer CNN with spatiotemporal input for accurate nuclear power fault diagnosis under different operating conditions. Annals of Nuclear Energy. 2023;194.

[83] Lee Y, Song SH, Bae JY, Song K, Seo MR, Kim SJ, et al. Surrogate model for predicting severe accident progression in nuclear power plant using deep learning methods and Rolling-Window forecast. Annals of Nuclear Energy. 2024;208.

[84] Zhou G, Peng Mj, Wang H. Enhancing prediction accuracy for LOCA break sizes in nuclear power plants: A hybrid deep learning method with data augmentation and hyperparameter optimization. Annals of Nuclear Energy. 2024;196.

[85] dos Santos MC, Pinheiro VHC, do Desterro FSM, de Avellar RK, Schirru R, Nicolau AdS, et al. Deep rectifier neural network applied to the accident identification problem in a PWR nuclear power plant. Annals of Nuclear Energy. 2019;133.

[86] Kim SG, Chae YH, Koo SR. Application of an open-set recognition method for detecting untrained accident scenarios in a nuclear power plant accident diagnosis model. Nuclear Engineering and Design. 2024;427.

[87] Popov E, Archibald R, Hiscox B, Sobes V. Artificial intelligence-driven thermal design for additively manufactured reactor cores. Nuclear Engineering and Design. 2022;395.

[88] Radaideh MI, Wolverton I, Joseph J, Tusar JJ, Otgonbaatar U, Roy N, et al. Physics-informed reinforcement learning optimization of nuclear assembly design. Nuclear Engineering and Design. 2021;372.

[89] Wang H, Gruenwald JT, Tusar J, Vilim R. Moisture-carryover performance optimization using physics-constrained machine learning,. Progress in Nuclear Energy. 2021;135.

[90] Rishehri HZ, Nejad MZ. Design and optimization of dual-cooled fuel assembly in a 12×12 configuration for NuScale SMR based on neutronic-thermohydraulic parameters using the combined ANN-GA approach. Progress in Nuclear Energy. 2023;163.

[91] Zhang Z, Guo Y, Tao Q. Dynamic multi-objective path-order planning research in nuclear power plant decommissioning based on NSGA-II. Annals of Nuclear Energy. 2024;199.

[92] Kim J, Lee D, Kim J, Kim G, Hwang J, Kim W, et al. Radioisotope identification using sparse representation with dictionary learning approach for an environmental radiation monitoring system. Nuclear Engineering and Technology. 2022;54(3):1037-48.

[93] Zhang F, Coble JB. Robust localized cyber-attack detection for key equipment in nuclear power plants. Progress in Nuclear Energy. 2020;128.

[94] Khatua S, Mukherjee V. Application of PLC based smart microgrid controller for sequential load restoration during station blackout of nuclear power plants. Annals of Nuclear Energy. 2021;417.

[95] Torisaki S, Miwa S. Robust bubble feature extraction in gas-liquid two-phase flow using object detection technique. Journal of Nuclear Science and Technology. 2020;57(11):1231–1244.

[96] Pinheiro VHC, Schirru R. Genetic programming applied to the identification of accidents of a PWR nuclear power plant. Annals of Nuclear Energy. 2019;124.

[97] Qian G, Liu J. Fault diagnosis based on gated recurrent unit network with attention mechanism and transfer learning under few samples in nuclear power plants. Progress in Nuclear Energy. 2023;155.

[98] Ayodeji A, Liu Yk. Support vector ensemble for incipient fault diagnosis in nuclear plant components. Nuclear Engineering and Technology. 2018;50(8).

[99] Lee TB, Jeong YH. Improvement of the subcooled boiling model using a new net vapor generation correlation inferred from artificial neural networks to predict the void fraction profiles in the vertical channel. Nuclear Engineering and Technology. 2022;54(12).

[100] Ebrahimzadeh A, Ghafari M, Moshkbar-Bakhshayesh K. Detection and estimation of faulty sensors in NPPs based on thermal-hydraulic simulation and feed forward neural network. Annals of Nuclear Energy. 2022;166.

[101] Liu Y, Mui T, Xie Z, Hu R. Benchmarking FFTF LOFWOS Test 13 using SAM code: Baseline model development and uncertainty quantification. Annals of Nuclear Energy. 2023;192.

[102] Herb J, Périn Y, Yum S, Mylonakis A, C, Demazière C, et al. ensitivity analysis in core diagnostics. Annals of Nuclear Energy. 2022;178.

[103] Liu J, Gong H, Wang Z, Li Q. Uncertainty analysis of dynamic mode decomposition for xenon dynamic forecasting. Annals of Nuclear Energy. 2023;194.

[104] Zhou W, Sun G, Yang Z, Wang H, Fang L, Wang J. BP neural network based reconstruction method for radiation field applications. Nuclear Engineering and Design. 2021;380.

[105] Heo Y, Lee C, Kim HR, Lee SJ. Framework for the development of guidelines for nuclear power plant decommissioning workers based on risk information. Nuclear Engineering and Design. 2022;387.