



Artificial intelligence: reflecting on the past and looking towards the next paradigm shift

Petar Radanliev

To cite this article: Petar Radanliev (28 Feb 2024): Artificial intelligence: reflecting on the past and looking towards the next paradigm shift, Journal of Experimental & Theoretical Artificial Intelligence, DOI: [10.1080/0952813X.2024.2323042](https://doi.org/10.1080/0952813X.2024.2323042)

To link to this article: <https://doi.org/10.1080/0952813X.2024.2323042>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 28 Feb 2024.



Submit your article to this journal [↗](#)



Article views: 655



View related articles [↗](#)



View Crossmark data [↗](#)



ARTICLE



OPEN ACCESS



Artificial intelligence: reflecting on the past and looking towards the next paradigm shift

Petar Radanliev^{a,b}

^aDepartment of Computer Sciences, University of Oxford, Oxford, UK; ^bSchool of Management, University of Bath, Bath, UK

ABSTRACT

Artificial intelligence (AI) has undergone major advances over the past decades, propelled by key innovations in machine learning and the availability of big data and computing power. This paper surveys the historical progress of AI from its origins in logic-based systems like the Logic Theorist to recent deep learning breakthroughs like Bidirectional Encoder Representations from Transformers (BERT), Generative Pretrained Transformer 3 (GPT-3) and Large Language Model Meta AI (LLaMA). The early rule-based systems using handcrafted expertise gave way to statistical learning techniques and neural networks trained on large datasets. Milestones like AlexNet and AlphaGo established deep learning as a dominant AI approach. Transfer learning enabled models pre-trained on diverse corpora to excel at specialised downstream tasks. The scope of AI expanded from niche applications like playing chess to multifaceted capabilities in computer vision, natural language processing and dialogue agents. However, current AI still needs to catch up to human intelligence in aspects like reasoning, creativity, and empathy. Addressing limitations around real-world knowledge, biases, and transparency remains vital for further progress and aligning AI with human values. This survey provides a comprehensive overview of the evolution of AI and documents innovations that shaped its advancement over the past six decades.

ARTICLE HISTORY

Received 21 October 2023

Accepted 3 February 2024

KEYWORDS

Artificial intelligence; machine learning; deep learning; natural language processing; artificial general intelligence

Introduction

Artificial intelligence (AI) is one of the most transformative technologies of our time. The quest to develop thinking machines that can simulate human intelligence has captivated computer scientists for over six decades. The field has undergone significant paradigm shifts driven by key algorithm innovations, data availability and computing power advancements. From humble beginnings in logic and rules, AI has rapidly evolved to statistically learn from large datasets using neural networks that contain millions of parameters. Applications that were once considered squarely in the realm of science fiction, like self-driving cars and conversational agents, are now part of our everyday lives.

This paper surveys the historical progress of AI from its origins in the 1950s to the current state-of-the-art. The survey comprehensively covers major innovations and breakthroughs that have shaped the advancement of AI over the past several decades. Beginning with the logic-based systems of the early pioneers, the paper traces the rise of machine learning and how techniques like deep learning have propelled AI capabilities in recent years. The scope and complexities of AI applications have

CONTACT Petar Radanliev ✉ petar.radanliev@cs.ox.ac.uk Department of Computer Sciences, University of Oxford, 7 Park Road, Oxford OX1 3QG, UK

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

expanded dramatically – from playing chess to captioning images, translating languages, and interacting naturally in dialogue.

However, despite the remarkable pace of progress, AI still needs more human intelligence and aptitude. Aspects like reasoning, creativity, and making ethical judgements remain challenging frontiers for AI. This paper discusses such limitations and sheds light on vital research issues and debates regarding the future development of AI technology. The survey provides a broad overview of historical milestones, current capabilities, and open challenges contextualising the promises and perils of rapidly advancing AI systems.

The logic theorist

In 1956, Allen Newell, Herbert A. Simon, and Cliff Shaw created the Logic Theorist computer program (Newell & Simon, 1956). This program was the first of its kind, designed to perform automated reasoning. Due to its revolutionary capabilities, it is often called the ‘first program with artificial intelligence’. Newell and Simon began developing Logic Theorist in 1955, before the field of artificial intelligence existed.

Herbert A. Simon, a political scientist who had already produced classic works in the study of bureaucracy and a theory of rationality that won the Nobel Prize, had valuable insights into critical thinking and decision-making abilities. In the 1950s, while observing a printer produce a map, he hypothesised that machines working with symbols could make decisions. He wondered if computers could simulate the thought process. The diagram in Figure 1, illustrates how a computer interprets various operators, such as ‘YES’ and ‘AND.’ The circles and Venn diagrams illustrate potential interpretations of data using the operators. ‘AND’ indicates that the information is present in both quantities, not just one.

Allen Newell, a scientist at the RAND Corporation who studied logistics and organisation theory, designed the printer that Simon observed. In 1954, Newell attended a presentation by artificial intelligence pioneer Oliver Selfridge on pattern matching. Newell had an epiphany when he realised

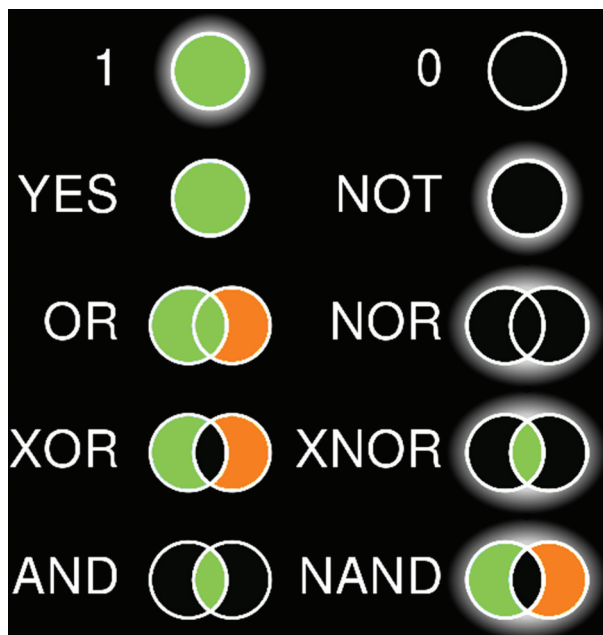


Figure 1. The logic theorist (Newell & Simon, 1956).

that the simple components of a computer could work together to achieve complex behaviour. He hoped that his theory could encompass human thought. 'Everything transpired in a single afternoon,' he would later say.

According to Pamela McCorduck's book 'Machines Who Think,' Newell had an overwhelming feeling of certainty that he was about to embark on a new path (McCorduck & Cfe, 2004). In the interviews with Pamela McCorduck, Newell argued that he is usually quite cynical and that he doesn't often get excited, but he did in this instance. He worked for 10 to 12 hours a day, completely absorbed in his work without the normal levels of consciousness that come with working, such as being aware of the consequences and ramifications.

Newell and Simon discussed the possibility of machines being able to 'think.' They then created the Logic Theorist (Newell & Simon, 1956), a program that could prove mathematical theorems from Principia Mathematica. Cliff Shaw, a computer programmer, also contributed to the program's creation. The first version of the program was manually simulated; the program was hand-written on 35 cards, as Simon recalls: 'In January 1956, we assembled my wife and three children together with some graduate students. To each member of the group, we gave one of the cards, so that each one became, in effect, a component of the computer program ... Here was nature imitating art imitating nature.'

They showed the program could prove theorems like a mathematician would. The Logic Theorist soon proved more theorems from Principia Mathematica. Newell and Simon formed a lasting partnership, founding one of the first AI laboratories at Carnegie Tech and developing a series of influential artificial intelligence programs and ideas, including GPS, Soar, and their unified theory of cognition.

The Dartmouth Workshop – the birth of artificial intelligence

The term 'artificial intelligence' was coined by John McCarthy in 1956 at the Dartmouth Workshop (James, 2006), and this event is considered the official birth of artificial intelligence as a new science (Crevier, 0000).

The 'thinking machines' field in the early 1950s had many names, including cybernetics, automata theory, and complex information processing (McCorduck & Cfe, 2004). The different names reflected the diversity of conceptual approaches. It was in 1955 when John McCarthy, assistant professor of mathematics at Dartmouth College at the time, decided to form a group to advance and clarify ideas regarding thinking machines. He wanted to avoid a narrow focus on automata theory and cybernetics, which was heavily focused on analogue feedback, and the possibility of having to accept the assertive Norbert Wiener as his mentor or argue with him (Nilsson, 2009). For that reason, he chose the name 'Artificial Intelligence' for the new field, which was neutral.

At the beginning of 1955, McCarthy approached the Rockefeller Foundation to request funding for a summer seminar at Dartmouth for about 10 participants (John et al., 1955). In June, he and Claude Shannon, a founder of information theory at Bell Labs at the time, met with Robert Morison, Director of Biological and Medical Research, to discuss the idea and potential funding. However, Morison was still determining whether money would be available for such a visionary endeavour (Kline, 2011). Finally, on 2 September 1955, McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon proposed the project and coined the term 'artificial intelligence.' (John et al., 1955).

One comment on the photo in Figure 2. Some have suggested that Trenchard More, another AI expert who attended the workshop, was the unknown person between Solomonoff and Shannon. However, according to a recent News Article by Grace Solomonoff in the IEEE Spectrum (Solomonoff, 2023), the person must be Peter Milner, who was working on neuropsychology at McGill University, in Montreal.



Figure 2. The Dartmouth Workshop. In the back row from left to right are Oliver Selfridge, Nathaniel Rochester, Marvin Minsky, and John McCarthy. In front on the left is Ray Solomonoff; on the right, Claude Shannon. The identity of the person between Solomonoff and Shannon remained a mystery for some time. (photo: Margaret Minsky found in (Solomonoff, 2023)).

The Perceptrons

Ten years after the discovery of the perceptron, which demonstrated that machines could be taught to perform specific tasks using examples, in 1969 Marvin Minsky and Seymour Papert published 'Perceptrons: An Introduction to Computational Geometry' (Minsky & Papert, 2017), an analysis of the computational capabilities of perceptrons for tasks.

Perceptrons played a crucial role in the emergence of neural networks and machine learning. The book highlighted the capabilities and limitations of single-layer perceptrons, which led to further research into multi-layer neural networks and a renewed interest in deep learning. Although the book has been criticised for its emphasis on the limitations of perceptrons in solving complex problems, this criticism prompted additional research that eventually led to the development of neural network architectures with greater capacity. Perceptrons is an essential historical document in artificial intelligence and machine learning. It paved the way for creating contemporary deep-learning techniques and is still referenced in AI literature.

In Chapter 12 of Perceptrons (see Figure 3), Minsky discusses multilayer networks briefly and states that he and Papert had yet to investigate multilayer perceptron networks. However, he did not state that they would not work. Given the hardware available then, he saw no reason why they wouldn't work as function approximation machines, but they appeared computationally expensive. A megabyte of memory cost over \$1 million in 1970 (approximately \$6 million today), and the fastest research machine, a KA10, had approximately 0.1 MIPS of integer arithmetic, forgetting floating point entirely. Even if someone had attempted to construct a usable multilayer perceptron network at the time, its performance would have been quite limited compared to the enormous deep networks used today for a single class of image recognition, even if they had used every available computer's RAM and CPU.

It took a long time for people to realise that perceptrons could be multilayered to overcome the limitations posed by Minsky and Pappert. In the 1960s, Warren McCulloch and Walter Pitts

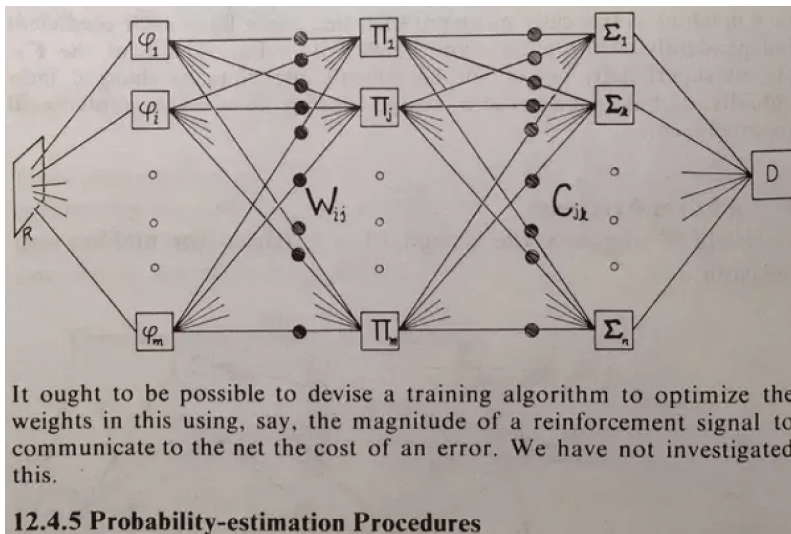


Figure 3. Chapter 12 of perceptrons (Minsky & Papert, 2017).

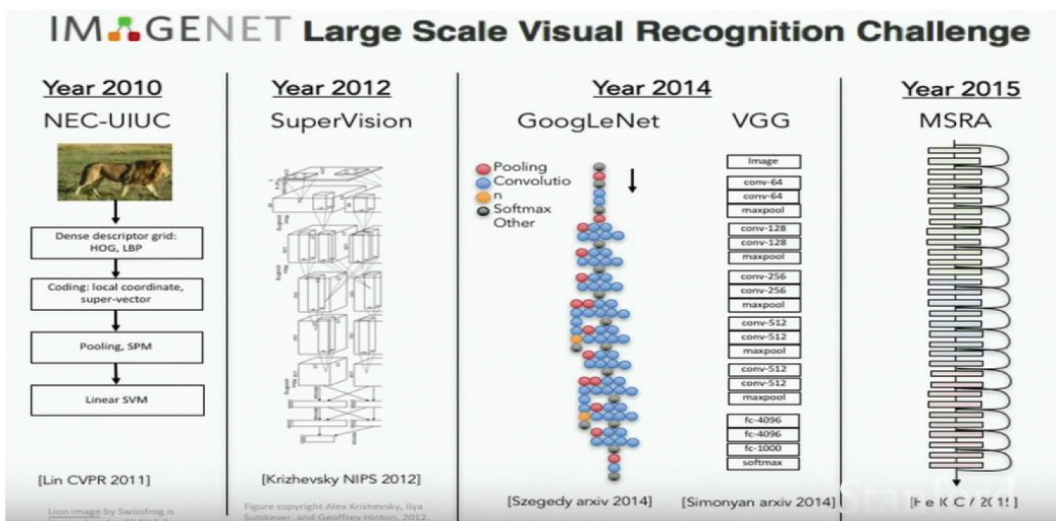


Figure 4. ImageNet classification with deep convolutional neural networks (Krizhevsky et al., 2012, 2017).

introduced the concept of using multiple perceptron layers, now known as artificial neural networks. However, the technology and understanding needed to realise the full potential of multilayered perceptrons were only available later. It was only in the 1980s and 1990s, when technological advancements and increased research in the field led to a greater understanding of the capabilities of artificial neural networks, that the limitations presented by Minsky and Pappert were overcome.

The Perceptron's Demise was somewhat different. The 'Perceptrons' limitations are connected to its inability to differentiate between input categories. In other words, unless input categories were 'linearly separable', the 'Perceptrons' could not differentiate between categories, and this proved problematic, as many important categories did not fall into specific input category. For example, the inputs to an XOR gate that generate a 1 output (such as 10 and 01) cannot be linearly separated from those that do not produce a 1 output (00 and 11).

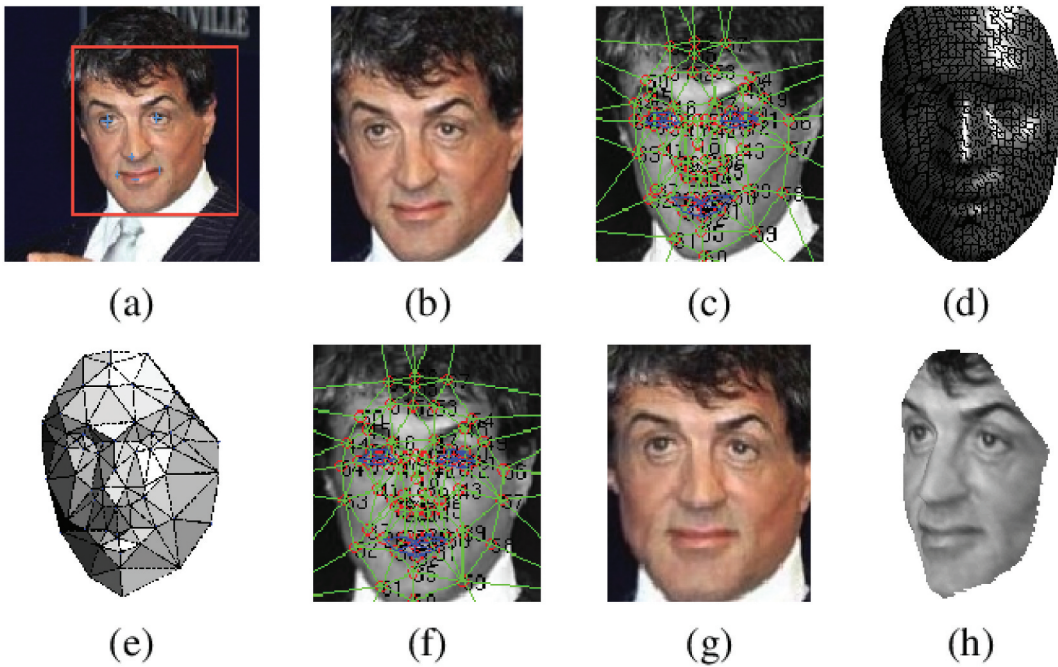


Figure 5. DeepFace: Closing the gap to human-level performance in face verification’ (Taigman et al., 2014).

Eliza

ELIZA is a computer program developed by Joseph Weizenbaum at MIT between 1964 and 1967 (Weizenbaum, 1976) - (Weizenbaum, 1966). It was created to explore communication between humans and machines. The program used a pattern-matching and substitution technique that made users feel like the program understood what was being said, even though it did not comprehend the conversation (Marianna & Höltingen, 2018)– (Norvig, 1992). Although ELIZA was originally written in MAD-SLIP, the language capability was provided in separate scripts using a Lisp-like notation. The most famous script, DOCTOR, simulated a psychotherapist of the Rogerian school and used script-dictated rules to respond to user inputs with non-directional questions (Bassett, 2019; Dillon, 2020). ELIZA was one of the first chatterbots, also known as a chatbot, and it was one of the first programs to attempt the Turing test.

Even today, we can still find online representations of ELIZA, such as the New Jersey Institute of Technology basic Rogerian psychotherapist chatbot (Ronkowitz, 0000), which can be used to test the functionality of the DOCTOR. Weizenbaum was concerned by the lay responses to ELIZA, which led him to write ‘Computer Power and Human Reason: From Judgment to Calculation’ (Benjamin et al., 1976). In this book, he explains the limitations of computers and emphasises that anthropomorphic views of computers reduce humans or any other life form. Weizenbaum clarified in the 2010 independent documentary film ‘Plug & Pray’ that only those who misunderstood ELIZA referred to it as a sensation.

David Avidan, an Israeli poet, was fascinated by the relationship between technology and art. He wanted to explore using computers in literary composition, so he had several conversations with an APL implementation of ELIZA. He published these conversations in English and Hebrew under the title ‘My Electronic Psychiatrist – Eight Authentic Talks with a Computer’ (Avidan, 2010). In the introduction, he described it as a constrained writing experiment.

Many ELIZA-based programs have been written in different programming languages. Some Sound Blaster cards for MS-DOS computers included a program called Dr. Sbeitso, which functioned

similarly to the DOCTOR script. Other adaptations of ELIZA had a religious theme, such as those featuring Jesus (both in a profound and humorous context) and an Apple II variant titled 'I Am Buddha.' In 1980, 'The Prisoner' incorporated ELIZA-style interaction into its gameplay. In 1988, the British artist and Weizenbaum friend Brian Reffin Smith created two art-oriented ELIZA-style programs in BASIC, titled 'Critic' and 'Artist,' which ran on two separate Amiga 1000 computers. They were displayed at the 'Salamandre' exhibition at the Musée de la Berry in Bourges, France. Visitors were expected to facilitate their conversation by typing into 'Artist' what 'Critic' had said and vice versa. The fact that the two programs were identical was kept secret. The psychoanalyse-pinhead command in GNU Emacs once simulated a session between ELIZA and Zippy the Pinhead. Due to copyright concerns, the Zippyisms have been removed, but the DOCTOR program remains.

ELIZA has been mentioned in popular culture and continues to inspire artificial intelligence programmers and developers. It was also featured in a Harvard University exhibit titled 'Go Ask A. L.I.C.E.' in 2012, which commemorated the 100th birthday of mathematician Alan Turing. The exhibit examined Turing's lifelong fascination with human-computer interaction, highlighting ELIZA as one of the earliest implementations of Turing's ideas. In 2021, ELIZA was awarded a Legacy Peabody Award.

CYC: using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks

The article titled 'CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks' by D. Lenat (Lenat et al., 1985), explores the challenges of AI knowledge acquisition and the fragility of expert systems. The author proposes the CYC system as a solution to these problems.

The article examines the 'expert system fragility' and how current expert systems exhibit 'fragility,' which refers to their poor performance when confronted with scenarios that deviate from their explicitly programmed knowledge. Lenat discusses the difficulty of inputting knowledge manually into AI systems, which slows down the process of making these systems more intelligent and versatile.

The CYC system aims to incorporate common sense reasoning to make artificial intelligence (AI) systems more adaptable to different situations and overcome their fragility. The paper discusses the use of analogical reasoning by CYC to expand its knowledge base and adapt to different situations. It also introduces the 'Braid' logical reasoner that employs probabilistic rules and custom unification functions to enhance CYC's reasoning abilities. This paper is a seminal work in artificial intelligence, especially in the field of expert systems.

Watson and jeopardy by IBM

IBM's Watson marked a significant milestone in artificial intelligence and natural language processing (Chandrasekar, 2014b). In 2011, Watson participated in the popular game show 'Jeopardy!' and showcased a sophisticated mix of data analytics, machine learning, and language comprehension to compete against human contestants (Yang et al., 2020).

Watson is engineered as an analytical engine capable of gathering data from multiple sources in real-time, conducting swift analysis, and evaluating the confidence of its responses (Baker, 2011). The machine's game-playing strategies were meticulously crafted using an accurate simulation model (Zadrozny et al., 2015). It could determine the risk and reward scenarios in a high-stakes game show environment and adapt its wagering and response strategies accordingly (Ferrucci, 2012).

During its appearance on 'Jeopardy!', Watson demonstrated its capabilities by dominating the game (Lewis, 2012; Zadrozny et al., 2015). The machine's adaptability was evident in the improvement of its performance. Watson's architecture was also designed to provide an intuitive user experience, which subsequently found applications in fields beyond game shows (Beltzung, 2013).

Watson's triumph on 'Jeopardy!' was more than just a spectacle. It showcased the remarkable potential of artificial intelligence in real-time risk assessment, decision-making, and human language comprehension (Gliozzo et al., 2017). This achievement created a standard for future AI applications in various fields, such as healthcare and business, and sparked considerable interest in the potential of AI (Chandrasekar, 2014a).

It also marked a significant advancement in natural language processing and artificial intelligence, demonstrating the progress made in machines' ability to comprehend human language (C. T.-T. M., 2011)

AlphaGo

AlphaGo, an innovative artificial intelligence system developed by DeepMind (Li & Du, 2018), a subsidiary of Alphabet Inc., is a major milestone in AI research, particularly in reinforcement learning and deep neural networks (Y. Chen, 2023). AlphaGo's three primary features are:

- (1) Deep Neural Networks: AlphaGo uses deep neural networks in its decision-making process. These neural networks are trained with human gameplay and self-generated data to improve accuracy (Lapan, 2018), (Fu, 2016).
- (2) Monte Carlo Tree Search (MCTS): AlphaGo employs the MCTS algorithm to analyse a range of potential moves and outcomes in Go, significantly improving its decision-making capabilities.
- (3) Reinforcement Learning: AlphaGo uses reinforcement learning methods to enhance its performance incrementally. It competes against itself, learning from its successes and mistakes (J. X. Chen, 2016).

AlphaGo can apply its learning to different yet related tasks, such as transitioning from Go to Chess, despite not being its primary focus during development (Holcomb et al., 2018; Wang et al., 2016). The evolution of AlphaGo to AlphaGo Zero involved removing the need to train on human gameplay data, which streamlined and improved the learning process (Chang et al., 2016; Granter et al., 2017). The system carefully balances computational resources and algorithmic innovation, making it efficient and scalable. The accomplishments of AlphaGo go beyond board games and have far-reaching implications for optimisation problems, predictive modelling, and other complex decision-making scenarios (Egri-Nagy & Törmänen, 2020).

ImageNet classification

In 2012, Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton published a ground-breaking paper that revolutionised computer vision and deep learning (A. Krizhevsky, I. Sutskever, & G. E. Hinton, 2023; Krizhevsky et al., 2012). They introduced a deep Convolutional Neural Network (CNN) model trained on the ImageNet Large Scale Visual Recognition Challenge dataset, comprising about 1.3 million high-resolution images divided into 1000 categories (Tasci & Kim, 2015). In Figure 4, we can visualise the example of ImageNet classification with deep convolutional neural networks.

The contributions seen in Figure 4 are crucial when it comes to addressing the complexity of high-resolution images. In this regard, the paper proposed a deep CNN model with multiple layers (Krizhevsky et al., 2017). The article also popularised using Rectified Linear Units (ReLU) as activation functions, which helped accelerate the training process. To prevent overfitting, the dropout technique was implemented. This technique randomly sets a portion of output features to zero during training. Furthermore, to augment the training dataset, techniques such as image translation and horizontal reflection were used (A. Krizhevsky, Sutskever, & Hinton, 2023).

The paper has set a new standard in the ImageNet competition by significantly reducing the top-five error rates. This breakthrough has played an instrumental role in advancing research in deep learning and computer vision. Moreover, the novel architecture has opened doors for industrial

applications like object detection and face recognition. The impressive results of the paper were achieved by training a large model made possible by the parallel computing power of GPUs (Alex et al., 2023).

The architecture used in this project demanded a significant amount of computational resources to operate smoothly. Furthermore, despite their efficiency in image processing, deep convolutional neural networks (CNNs) have been criticised for being ‘black boxes’ due to their complex structure, making it difficult to interpret the reasoning behind their decisions. As a result, researchers have been exploring various methods to increase the transparency and explainability of deep CNNs.

DeepFace

In 2014, Yaniv Taigman and his colleagues published a paper titled ‘DeepFace: Closing the Gap to Human-Level Performance in Face Verification’ (Taigman et al., 2014), which significantly advanced facial recognition technology. They implemented deep learning techniques to overcome face verification issues, which significantly improved performance metrics. The method achieved an accuracy rate of 97.35% on the Labelled Faces in the Wild (LFW) dataset. DeepFace uses a nine-layer deep neural network that automatically identifies intricate structures in high-dimensional data (See Figure 5). However, building such a deep network was a technological challenge due to the computational resources needed for training.

In Figure 5, we can visualise the DeepFace’s ability to create a 3D model of a face using piecewise affine transformation is a remarkable feature. This advanced technique allows for more accurate face alignment, which is critical for precise face verification, especially when facial features may be fragile or variable. With DeepFace’s 3D model, face recognition systems can overcome these challenges and achieve greater reliability.

According to a recent study, DeepFace’s face verification accuracy rate is equivalent to human-level performance, reducing the error rate of existing technologies by more than 27%. However, this development has raised ethical concerns about technology that can outperform humans in specific tasks.

The research conducted by Taigman and his team in 2014 has significantly impacted the field of facial recognition. Their use of deep learning techniques for face verification set a new standard and inspired further research in various application domains. Since then, their novel approach has led to a paradigm shift in the methodology of facial recognition research. The paper makes significant contributions to the field but has some limitations. As the paper was published some time ago, the architecture’s depth and computational efficiency may now be considered outdated. Moreover, the paper does not provide in-depth analysis of ethical considerations, particularly those related to false positives and negatives that are particularly crucial in high-stakes applications such as law enforcement and security.

Bidirectional encoder representations from transformers (BERT)

In 2018, Google researchers unveiled the Bidirectional Encoder Representations from Transformers (BERT) language models (Devlin et al., 2019). A literature review conducted in 2020 revealed that BERT had rapidly become a standard baseline for Natural Language Processing (NLP) experiments in less than a year. The model has been analysed and improved in over 150 research publications (Rogers et al., 2020). There are two BERT models available for the English language [49]: BERTBASE, which boasts 12 encoders, 12 bidirectional self-attention heads, and 110 million parameters; and BERTLARGE, which features 24 encoders, 16 bidirectional self-attention heads, and 340 million parameters. Both models were originally trained on the Toronto BookCorpus (800 million words) (Zhu et al., 2015), and English Wikipedia (2,500 million words).

Bert is an encoder architecture that employs only transformers (Alaparthi & Mishra, 2020a, 2020b). The architecture comprises three main components: embedding, encoders, and un-embedding. The

embedding component transforms a collection of tokens encoded one-hot into a collection of vectors representing the tokens. Encoders, which are the transformers, transform the array of vector representations by applying transformations to them. The un-embedding module is necessary for pre-training, but is usually not required for subsequent tasks. Instead, the vector representations of the text input, the final representation vectors output at the end of the encoder stack, are used as a vector representation of the text input. A more compact model can then be trained on top of this. To translate each English word into an integer code, BERT uses WordPiece, whose vocabulary size is 30,000 words. Any token not found in the vocabulary is replaced with [UNK] for 'unknown'.

The embedding layer comprises three components: word embeddings, position embeddings, and token type embeddings. The word embeddings component takes an input token as a one-hot vector of 30,000 dimensions, equivalent to BERT's vocabulary size. The position embeddings component is accountable for absolute position embedding, similar to word embeddings, but only applies to time stamps ranging from 0 to 511, as BERT has a context window 512. The token type embeddings component is comparable to word embeddings but only for a vocabulary of 0 and 1. Tokens appearing after the [SEP] are of type 1, while all remaining tokens are type 0. The three outputs of these components are added and then passed through a LayerNorm, which generates an array of 768-dimensional representation vectors. Finally, after passing through 12 Transformer encoders, the representation vectors are de-embedded by an affine-Add & LayerNorm-linear.

The BERT model of contextual representation supports various natural language processing (NLP) tasks. Google created it and has revolutionised the NLP landscape. BERT's architecture is based on the Transformer design and differs from conventional models. It reads the text in both directions, considering all layers' left and right contexts. This contrasts with unidirectional models that read text from left to right or right to left. BERT is first pre-trained on a large corpus of text and then fine-tuned for specific tasks. The pre-training phase includes masked language modelling and prediction of the next sentence. BERT is good at understanding the context in which words appear, leading to more precise representations. It can recognise the significance of common words such as 'to' in a sentence.

BERT has many applications, including search engine optimisation, sentiment analysis, named entity recognition, and machine translation. With an astounding number of citations, BERT has fundamentally shifted the paradigm in NLP and serves as the foundation for many derivative models and applications.

BERT vs BART

There is a key difference between BERT and BART in terms of their pre-training tasks. In BERT, certain words in the input text are replaced with a special token, and the model is trained to predict the original words. On the other hand, in BART, the input text is corrupted with random tokens, and the model is trained to reconstruct the original text. These different pre-training tasks result in distinct advantages and disadvantages for each of the models.

BERT is highly recognised for its exceptional performance in context- and relationship-sensitive tasks like sentiment analysis and question-answering. The pre-training task of BERT encourages the model to acquire context-aware word representations, making it well-suited for tasks that require understanding words in different contexts.

On the other hand, BART is known for its superior performance in tasks requiring complex language processing like text summarisation and machine translation. The BART pre-training task encourages the model to acquire representations insensitive to noise and variations in the input text. This makes BART suitable for tasks requiring the processing of noisy, ambiguous, or multilingual text.

Another key difference between BERT and BART is the transformer's design. BERT utilises a transformer architecture with a multilayer encoder, while BART uses a transformer architecture with a multilayer encoder-decoder. This architectural difference leads to distinct computational and memory requirements for each model.

Table 1. Comparison between BERT and BART in a summary table.

Feature	BERT (Bidirectional Encoder Representations from Transformers)	BART (Bidirectional and Auto-Regressive Transformers)
Architecture	Transformer-based, predominantly Encoder	Transformer-based, includes both Encoder and Decoder
Training Direction	Bidirectional	Bidirectional, but also auto-regressive in the decoding phase
Training Tasks	Masked Language Model (MLM) and Next Sentence Prediction (NSP)	Multiple tasks like text generation, summarisation, etc.
Contextual Understanding	Yes, both left and right context	Yes, with enhanced sequence-to-sequence tasks
Fine-tuning Required?	Required for specific tasks	Required for specific tasks
Token Handling	Masks certain tokens for prediction	Entire sequences can be masked
Applications	Text Classification, Named Entity Recognition, Sentiment Analysis	Text Summarisation, Translation, Question-Answering
Developer	Google	Facebook AI
Year of Introduction	2018	2019
Notable Derivatives	RoBERTa, DistillBERT	mBART, BART-large
Impact on NLP	Pioneering, set new standards for NLP tasks	Extended capabilities, especially in sequence-to-sequence tasks

The architecture of BERT is relatively simple, which makes it easy to use and compatible with various devices. However, this simplicity also makes BERT less suitable for tasks that involve managing lengthy text sequences or generating text. On the contrary, BART's architecture is more complex, which makes it a better fit for tasks that require handling lengthy text sequences or text generation.

The table below compares two innovative models, BERT and BART, in the field of Natural Language Processing. The [Table 1](#) highlights their unique characteristics, methodologies, and applications. It aims to help users understand the models better by presenting complex data in an easily digestible form. Whether you are a researcher or an industry professional, this table offers a curated summary of the most pertinent information, which can be useful in decision-making and scholarly discussions.

The development of BERT's bidirectional architecture and contextual understanding revolutionised NLP. It paved the way for the creation of many applications and subsequent models. Based on BERT, BART further expanded its capabilities with a sequence-to-sequence architecture, making it more suitable for generative tasks such as text summarisation and translation. Both models have played a significant role in the progress of NLP and have been widely embraced in academic and business environments.

ChatGPT

ChatGPT possesses several advantages over other AI models. One of its major strengths is the ability to fine-tune responses and provide highly specific and relevant contextual replies, which can be customised to cater to the requirements of a particular domain or subset of users. As a generative model, ChatGPT has the flexibility to produce content versatily, making it more adaptable and often compared favourably to other generative AI models like Google Bard. It is a self-sufficient chatbot that allows seamless user interaction without additional productivity tools, making it a flexible solution for various applications.

The vast underlying language model of ChatGPT enables it to produce a wide range of responses to user queries, making it highly interactive and adaptable. Furthermore, since it is based on OpenAI's architecture, ChatGPT benefits from strong community support, including tuning guides, pre-trained models, and an extensive library of utilities that other AI models may lack. The use of Reinforcement Learning from Human Feedback (RLHF) makes ChatGPT unique

since human AI trainers provided the model with conversations where they played both parts according to OpenAI.

ChatGPT is considered a superior AI model for several reasons. It displays impressive versatility and has a firm grip on natural language comprehension. This makes it capable of engaging in meaningful conversations with sensitivity to context, making it a valuable tool for multiple applications such as customer service and content creation. Users frequently find ChatGPT's responses more engaging and useful, thanks to its ability to generate coherent, context-sensitive text. This significantly improves the user experience compared to older AI models that may produce less coherent responses.

ChatGPT is known for its ability to generate high-quality text on various topics. It is preferred for content creation and summarisation tasks due to its superior text generation capabilities. Furthermore, ChatGPT is designed to be easily integrated with several platforms, making it accessible to developers and businesses. This ease of use contributes to its widespread adoption and popularity. ChatGPT and other advanced AI models represent a point of no return in artificial intelligence. It can effectively communicate in plain English, which brings artificial intelligence closer to seamless human-computer interactions. OpenAI, the organisation behind ChatGPT, continually enhances and refines the model based on user feedback and scientific research. This dedication to ongoing development ensures that ChatGPT remains at the forefront of artificial intelligence technology.

LLaMA (large language Model meta AI)

LLaMA (Large Language Model Meta AI) is a set of large language models (LLMs) developed and released by Meta AI in February 2023. On 18 July 2023, in collaboration with Microsoft, Meta announced the launch of LLaMA-2, the next generation of the LLaMA family. LLaMA-2 has three model sizes: 7 billion, 13 billion, and 70 billion parameters. The model's architecture remains largely unchanged from that of LLaMA-1 models, but 40% more data was used to train the foundational models, making LLaMA-2 more advanced than its predecessor.

LLaMA-2 consists of foundational models and dialogue fine-tuned models, referred to as LLaMA-2 Chat. Unlike LLaMA-1, all models in LLaMA-2 are released with weights and can be used for many commercial purposes for free. However, due to certain restrictions, some have disputed the description of LLaMA as open source by the Open Source Initiative, which is responsible for maintaining the Open Source Definition.

LLaMA is built on the transformer architecture, which has been the standard architecture for language modelling since 2018. However, there are some minor differences in the architecture used by LLaMA compared to GPT-3. For instance, LLaMA uses SwiGLU (Google, 2020) activation function instead of ReLU, rotary positional embeddings (Su et al., 2021) instead of absolute positional embedding, and root-mean-squared layer-normalisation (Zhang & Sennrich, 2019) instead of standard layer-normalisation (Ba et al., 2016). Additionally, LLaMA has an increased context length of up to 4K tokens between, which is twice the length used in LLaMA 1.

Claude

Claude, developed by Anthropic, is designed to excel in certain specific areas. It is optimised for answering open-ended questions that require a certain depth of understanding and capability for nuanced responses. Claude has also been tailored for tasks such as providing helpful advice, searching, writing, and editing. It reportedly has a higher performance in maths, reasoning, and coding skills compared to previous models by Anthropic. Claude can be integrated with LangChain to create a Search-Powered Personal Assistant AI App and is highly capable of advanced research and application testing. Although not yet as capable as GPT-4, Claude 2 performs well on standardised tests and is improving rapidly. These features differentiate Claude from other AI models, making it a tailored solution for specific research and applications. Claude Pro can handle up to

100,000 tokens, providing a more extensive user experience than ChatGPT Plus, which only supports 32,000 tokens.

The future evolution of generative AI

The survey has shown that generative AI models' quick development is exciting future opportunities. These models, such as GPT-3, DALL-E 2, and Stable Diffusion, are predicted to bring about the future of artificial intelligence (AI) systems producing original text, images, and multimedia content with little to no human involvement. This article explores some of the most promising research trajectories that may further advance generative AI.

The capabilities of foundation models are expected to increase significantly, thanks to larger language models like GPT-3 and code-generating models like Codex. These models, which have trillions of parameters and are trained on enormous multimodal datasets, could develop high-level skills such as logical reasoning, creativity, and extensive world knowledge. They can be customised for various downstream tasks using methods such as transfer learning and prompt-based learning, which can significantly expand the range of applications for generative AI.

Contemporary generative models still face challenges in controlling the style, content, and bias of the outputs generated. Future developments must focus on making generative AI more manageable, secure, and reliable. Using methodologies like steering, dynamic prompts, and constrained optimisation, generative models can be guided to produce results that precisely align with the users' intended properties.

Future generative AI models may be able to combine different modes of communication such as text, images, audio, video, and 3D environments. These models will differ from DALL-E, which focuses solely on images. By training these models end-to-end on a wide range of multimodal data at scale, they could develop a solid understanding of the world and produce multimedia content that seamlessly integrates all modes of communication. This can open up innovative possibilities that we cannot yet imagine.

It is becoming increasingly important to be able to identify machine-generated text, images, or audio as generative models produce content that is more and more similar to human-written content. To distinguish fake human creations from real ones, it is necessary to use adversarial training and multimedia forensic methods. Detecting and flagging generated content can help prevent the spread of false information and ensure the authenticity of human-created content.

Generative AI models can be specialised in fields such as drug discovery, materials science, and education, resulting in a significant impact. By pre-training custom models on pertinent textual data using the appropriate architecture, accessibility and human potential can be improved.

To sum up, the development of generative AI models can lead to ground-breaking applications in content creation, personalised recommendations, creative arts, and scientific discovery that are tailored to human needs. Continued research and responsible development of generative AI may support and enhance human creativity across all disciplines.

Discussion

The development of artificial intelligence has seen remarkable progress over the past few decades, as highlighted in the main text. From the fundamental logic-based systems like the Logic Theorist in 1956, AI has rapidly advanced to sophisticated deep learning models like ChatGPT to generate human-like conversations.

A key observation is a shift from rule-based expert systems to data-driven machine-learning approaches. Early AI systems like ELIZA used handcrafted rules and scripts to simulate intelligent behaviour. However, these rule-based systems needed to be revised, requiring extensive human effort to define domain-specific rules. The availability of big data and increased computing power

enabled the rise of statistical machine learning techniques like deep neural networks that can automatically learn representations from large datasets. As the main text shows, breakthroughs like AlexNet, BERT and ChatGPT are built on deep learning architectures trained on massive corpora.

An interesting aspect is the symbiotic relationship between advances in AI and the availability of data and compute resources. Larger datasets and models have directly translated to improved capabilities, as evidenced by the evolution from BERT to GPT-3 and LLaMA. However, scaling up model size requires proportional increases in computing power. The training of massive models like GPT-3 was only made possible through access to resources like cloud computing and specialised hardware like GPUs and TPUs. Democratising access to such resources will be key for future progress.

The survey in the main text also highlights the expanding scope of AI systems, from niche tasks like playing chess to multifaceted applications like visual recognition, language understanding and dialogue agents. Transfer learning techniques have accelerated this trend, allowing models trained on large datasets to be fine-tuned for specialised downstream tasks. The versatility of large language models like GPT-3 and BART demonstrates the generalisability of such foundation models.

However, despite the rapid progress, AI systems still need to catch up on human intelligence in many regards. Aspects like common sense, reasoning, creativity, and empathy remain challenging for AI. Hybrid approaches combining neural techniques with symbolic logic and knowledge representation can overcome some limitations. However, designing architectures that can acquire and leverage real-world knowledge in a scalable manner remains an open problem.

There are several limitations to the use of Artificial Intelligence (AI), one of which is the possibility of bias in both training data and models. This bias can perpetuate harmful stereotypes and lead to unfair outcomes. To combat this issue, there has been a recent focus on developing transparent, fair, and human-aligned AI systems. Techniques like adversarial training, causal modelling, and algorithmic recourse can help to mitigate biases and enable responsible deployment of AI technology.

In recent AI models, the focus has been on deep multilayer perceptron (MLP) models (Liu et al., 2021), which could be a new paradigm for computer vision. In a recent study, the authors compared the connections and differences between convolution, self-attention mechanisms, and token-mixing MLPs. They also analysed recent MLP-like variants and their applications. However, the authors concluded that MLP is not a new paradigm that can push computer vision to new heights. Instead, they suggest that the next paradigm will emerge with the development of a new computing hardware system that uses a non-weighted summation network, such as the Boltzmann machine. The authors also provided a historical perspective on MLP, convolution neural networks (CNNs), and Transformers and discussed their corresponding paradigms. The authors highlighted the importance of computing hardware in driving the development of paradigms in computer vision.

Overall, this survey has highlighted the significant impact of deep learning and data availability on advancing AI capabilities. However, there are still significant challenges to overcome to achieve artificial general intelligence that matches or exceeds human-level aptitude across all facets of intelligence. Addressing these open problems will require sustained research and responsible development of AI technology aimed at benefiting humanity.

Conclusion

This survey traces the remarkable progress of artificial intelligence from its origins in logic and symbolism to the data-driven deep learning techniques that drive modern AI. The key milestones and innovations that have advanced AI over the decades are analysed. However, as the last section elucidates, we are potentially on the cusp of another paradigm shift in AI capabilities.

Recent breakthroughs in generative models like DALL-E 2, GPT-3 and Stable Diffusion demonstrate astonishing abilities to produce original text, code, images, and other multimedia content. This signals a future where AI could acquire extensive world knowledge and high-level cognitive capabilities through continued training at an immense scale, unlocking creative potential we cannot yet fathom.

Realising this future would require sustained research into making such generative models more controllable, safe, and specialised for focused domains. Effective techniques for detecting machine-generated content will also grow in importance as AI creation becomes increasingly indistinguishable from human output. Responsible development and deployment remain vital, considering the disruptive implications of such sophisticated generative AI on economics, disinformation, and society.

The key findings and new knowledge that emerge are:

- There has been a shift from symbolic, rule-based AI to data-driven machine learning techniques. Modern AI breakthroughs are underpinned by neural networks trained on massive datasets rather than hand-coded rules.
- The availability of big data and increased computing power, especially GPUs, has catalysed the development of deep learning models. Access to resources like cloud computing has enabled the training of complex models with billions of parameters.
- Transfer learning and pre-training foundation models on large corpora have become dominant techniques. Models like BERT, GPT-3 and LLaMA-2 demonstrate how pre-training on diverse datasets leads to more generalised capabilities.
- The scope of AI has expanded from narrow applications like playing chess to multifaceted tasks like computer vision, natural language processing, speech recognition and dialogue agents. Systems can be fine-tuned for specialised downstream uses.
- However, despite rapid progress, AI still needs to improve in human intelligence in aspects like reasoning, creativity, common sense and empathy. Building architectures that can acquire and apply real-world knowledge remains an open challenge.
- There is increased focus on developing transparent, fair, and trustworthy AI systems that align with human values. Mitigating algorithmic biases and enabling the responsible deployment of AI is vital for its acceptance.

In conclusion, this survey provides a comprehensive overview of the evolution of AI, anchored in major innovations of the past. It also casts an anticipative eye towards opportunities and challenges posed by the new wave of increasingly capable generative AI systems. As AI progresses from narrow intelligence towards more general capabilities, aligning its development to benefit humanity remains imperative. This survey aims to serve as a reflective guidepost for navigating this important transition.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Engineering and Physical Sciences Research Council [EP/S035362/1] and the Economic and Social Research Council [ES/V003666/1].

References

- Alaparthi, S. and Mishra, M. (2020a, July). Bidirectional encoder representations from transformers (BERT): A sentiment analysis odyssey. Retrived October 20, 2023, from <http://arxiv.org/abs/2007.01127>
- Alaparthi, S., Mishra, M. (2020b). Bidirectional encoder representations from transformers (BERT): A sentiment analysis odyssey. *Arxiv.Org*. Retrived October 20, 2023, from <https://arxiv.org/abs/2007.01127>
- Alex, A., Krizhevsky, K., Ilya, I., Sutskever, S., and Hinton, G. E. ImageNet classification classification with deep convolutional convolutional neural networks. *pdfs.semanticscholar.org*. Retrived October 20, 2023, from <https://pdfs.semanticscholar.org/3cd5/a85dc9da55dc0b7aa7787ba49925f79b32e6.pdf>

- Avidan, D. (2010). Collected poems. In *Hakibbutz Hameuchad*. (Vol. 3, OCLC 804664009 ed). Mwsad Bya'liyy, Yrwšalayim: Hwsa'at haQiybws haM'whad. <https://search.worldcat.org/title/804664009>.
- Baker, S. (2011). Final jeopardy: The story of Watson, the computer that will transform our world. Retrived October 20, 2023, from [https://books.google.com/books?hl=en&lr=&id=CwLSrDUoVB0C&oi=fnd&pg=PP1&dq=Watson+and+Jeopardy!+by+IBM+\(2011\)&ots=TPZ0AN5zB2&sig=7nI37ZkwaS_vV6HohOQqSuzC5wM](https://books.google.com/books?hl=en&lr=&id=CwLSrDUoVB0C&oi=fnd&pg=PP1&dq=Watson+and+Jeopardy!+by+IBM+(2011)&ots=TPZ0AN5zB2&sig=7nI37ZkwaS_vV6HohOQqSuzC5wM).
- Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016, July). Layer normalization. Retrived October 20, 2023, from <http://arxiv.org/abs/1607.06450>
- Bassett, C. (2019, December). The computational therapeutic: Exploring Weizenbaum's ELIZA as a history of the present. *AI & SOCIETY*, 34(4), 803–812. <https://doi.org/10.1007/S00146-018-0825-9>
- Beltzung, L. (2013). Watson jeopardy! A thinking machine. Retrived October 20, 2023, from <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e123d71b927bdd5fd5873023804f06894b2831f>.
- Benjamin, K., John, M., & Joseph, W. (1976, June). Computer power and human reason. *ACM SIGART Bulletin*, (58), 4–13. <https://doi.org/10.1145/1045264.1045265>
- Chandrasekar, R. (2014a). Elementary? Question answering, IBM's Watson, and the jeopardy! challenge. Springer. Retrived October 20, 2023, from https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/article/10.1007/s12045-014-0029-7&casa_token=YvPwOZBql0AAAAA:NUQM01wcZCpFDAZk3yaBn15X3BPFgx_H5QrbDdfJOSn1z5PirgFsS6GcpDocyyeFKwBqTTNmNEsQJPB
- Chandrasekar, R. (2014b, March). Elementary? Question answering, IBM's Watson, and the jeopardy! challenge. *Resonance*, 19(3), 222–241. <https://doi.org/10.1007/S12045-014-0029-7>
- Chang, H. S., Fu, M. C., Hu, J., & Marcus, S. I. (2016). Google DeepMind's AlphaGo: Operations research's unheralded role in the path-breaking achievement. *Go.Gale.com*. Retrived October 20, 2023, from <https://go.gale.com/ps/i.do?id=GALE%7CA471000820&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=10851038&p=AONE&sw=w>
- Chen, J. X. (2016). The evolution of computing: AlphaGo. *leeeexplore.lee.org*. Retrived October 20, 2023, from <https://ieeexplore.ieee.org/abstract/document/7499782/>
- Chen, Y. Bayesian optimization in alphago. *arxiv.org*. Retrived October 20, 2023, from <https://arxiv.org/abs/1812.06855>
- Crevier, D. *AI : The tumultuous history of the search for artificial intelligence*. p.386.
- C. T.-T. M. (2011, June). <http://www.nytimes.com/2010/10/20/technology/20ibm-watson.html>, What is IBM's Watson. *whsfilmfestival.com*. Retrived October 20, 2023, from http://www.whsfilmfestival.com/Walpole_High_School_Film_Festival/Grammar_Articles_files/Smarter%20Than%20You%20Think%20-%20I.B.M.'s%20Supercomputer%20to%20Challenge%20'Jeopardy!%20Champions%20-%20NYTimes.com.pdf
- Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (Vol. 1, pp. 4171–4186). North American Chapter of the Association for Computational Linguistics. <https://doi.org/10.48550/arXiv.1810.04805>
- Dillon, S. (2020, January). The Eliza effect and its dangers: From demystification to gender critique. *Taylor & FrancisS DillonJournal for Cultural Research*, 2020-taylor & Francis, 24(1), 1–15. <https://doi.org/10.1080/14797585.2020.1754642>
- Egri-Nagy, A., & Törmänen, A. (2020). The game is not over yet—go in the post-alphago era. *Philosophies*, 5(4), 37. <https://doi.org/10.3390/philosophies5040037>
- Ferrucci, D. (2012). Introduction to 'this is watson, *leeeexplore.lee.org*. Retrived October 20, 2023, from <https://ieeexplore.ieee.org/abstract/document/6177724/>
- Fu, M. C. (2016). AlphaGo and Monte Carlo tree search: The simulation optimization perspective. *leeeexplore.lee.org*. Retrived October 20, 2023, from <https://ieeexplore.ieee.org/abstract/document/7822130/>
- Gliozzo, A., Ackerson, C., Bhattacharya, R., and Goering, A. (2017). *Building cognitive applications with IBM Watson services: Volume 1 getting started*. Retrived October 20, 2023, from [https://books.google.com/books?hl=en&lr=&id=7W0pDwAAQBAJ&oi=fnd&pg=PR5&dq=Watson+and+Jeopardy!+by+IBM+\(2011\)&ots=_BcdP51oRk&sig=KIFnRXFU2w0ShhHUnJiUhf-JeM](https://books.google.com/books?hl=en&lr=&id=7W0pDwAAQBAJ&oi=fnd&pg=PR5&dq=Watson+and+Jeopardy!+by+IBM+(2011)&ots=_BcdP51oRk&sig=KIFnRXFU2w0ShhHUnJiUhf-JeM)
- Google, N. S. (2020, February). GLU Variants Improve Transformer. Retrived October 20, 2023, from <https://arxiv.org/abs/2002.05202v1>
- Granter, S. R., Beck, A. H., & Papke, D. J., Jr. (2017). AlphaGo, deep learning, and the future of the human microscopist. *Meridian.Allenpress.com*, Retrived October 20, 2023, from <https://meridian.allenpress.com/aplm/article-abstract/141/5/619/194217>
- Holcomb, S. D., Porter, W. K., Ault, S. V., Mao, G., & Wang, J. (2018, March). Overview on deepmind and its alphago zero ai. *dl.acm.org*, 67–71. <https://doi.org/10.1145/3206157.3206174>
- James, M. (2006, December). The Dartmouth college artificial intelligence conference. *AI Magazine*. <https://doi.org/10.1609/AIMAG.V27I4.1911>
- John, M., Marvin, M., Nathaniel, R., & Claude, S. (1955). A proposal for the Dartmouth summer research project on artificial intelligence (1955). *AI magazine 27: Dartmouth Workshop*, 4. <https://doi.org/10.1609/aimag.v27i4.1904>
- Kline, R. R. (2011, October). Cybernetics, Automata Studies, and the Dartmouth Conference on Artificial Intelligence. *IEEE Annals of the History of Computing*, 33(4), 5–16. <https://doi.org/10.1109/MAHC.2010.44>

- Krizhevsky, A., Sutskever, I., and Hinton, G. ImageNet classification with deep convolutional neural networks (AlexNet). *ImageNet classification with deep convolutional neural networks (AlexNet)*. *Actorsfit.Com*. Retrived October 20, 2023, from <https://actorsfit.com/a/?ID=00450-9b1db208-6e87-450e-bc1d-78f9c16d5996>
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. Retrived September 3, 2023, from <http://code.google.com/p/cuda-convnet/>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Proceedings.Neurips.cc*. Retrived October 20, 2023, from <https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017, June). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. <https://doi.org/10.1145/3065386>
- Lapan, M. (2018). *Deep reinforcement learning hands-on: Apply modern RL methods, with deep Q-networks, value iteration, policy gradients, TRPO, AlphaGo zero and more*. Retrived October 20, 2023, from <https://books.google.com/books?hl=en&lr=&id=xKdhDwAAQBAJ&oi=fnd&pg=PP1&dq=AlphaGo&ots=wUgilj2h7C&sig=ZNDOD-qx8Hw3xK3sA6E4ZEsNB6A>
- Lenat, D., Prakash, M., Shepherd, M. (1985). CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *Ojs.Aaai.orgDB Lenat, M Prakash, M ShepherdAI magazine, 1985-ojs Aaai Org*. Retrived October 20, 2023, from <https://ojs.aaai.org/index.php/aimagazine/article/view/510/0>
- Lewis, B. L. (2012). In the game: The interface between Watson and jeopardy! *Ieeexplore.Ieee.org*. Retrived October 20, 2023, from <https://ieeexplore.ieee.org/abstract/document/6177728/>
- Li, F., & Du, Y. (2018). From AlphaGo to power system AI: What engineers can learn from solving the most complex board game. *IEEE Power and Energy Magazine*, 16(2), 76–84. <https://doi.org/10.1109/MPE.2017.2779554>
- Liu, R., Li, Y., Tao, L., Liang, D., & Zheng, H. T. (2021, November). Are we ready for a new paradigm shift? A survey on visual deep MLP. *Patterns*, 3(7), 100520. <https://doi.org/10.1016/j.patter.2022.100520>
- Marianna, B., & Höltingen, S. (2018). *Projektverlag* (1st ed., pp. 261). Projekt. <https://www.amazon.de/Hello-Eliza-Gespr%C3%A4che-Computern-Computerarch%C3%A4ologie/dp/3897334674>
- McCorduck, P. and Cfe, C. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. CRC Press. Retrived October 20, 2023, from <https://books.google.com/books?hl=en&lr=&id=r2C1DwAAQBAJ&oi=fnd&pg=PP1&dq=Pamela+McCorduck+%22Machines+Who+Think&ots=UnmXliuRtM&sig=JAh90Eu07MGvjS5OgFq1CMY6-gc>
- Minsky, M., & Papert, S. A. (2017, January). Perceptrons: An introduction to computational geometry. *Perceptrons*. <https://doi.org/10.7551/MITPRESS/11301.001.0001>
- Newell, A., & Simon, H. A. (1956). The logic theory machine a complex information processing system. *IRE Transactions on Information Theory*, 2(3), 61–79. <https://doi.org/10.1109/TIT.1956.1056797>
- Nilsson, N. J. (2009). *The quest for artificial intelligence*. Cambridge University Press. Retrived October 20, 2023, from <https://books.google.com/books?hl=en&lr=&id=nUJdAAAAQBAJ&oi=fnd&pg=PT5&dq=Nilsson,+N.,+The+Quest+for+Artificial+Intelligence,+Cambridge+University+Press,+2010&ots=2mO4INbkMD&sig=LBkxiZL6vD1wBDRbQWUhdznruA>
- Norvig, P. (1992). *Paradigms of artificial intelligence programming: Case studies in common LISP*. Retrived October 20, 2023, from [https://books.google.com/books?hl=en&lr=&id=QzGuHnDhVzIC&oi=fnd&pg=PR7&dq=Norvig,+Peter+\(1992\).+Paradigms+of+Artificial+Intelligence+Programming.+New+York:+Morgan+Kaufmann+Publishers.+pp.+151%E2%80%93154.+ISBN+1-55860-191-0&ots=pVvYzCW8tt&sig=5OaZec96OmCrmCvB3szp9c9xWrs](https://books.google.com/books?hl=en&lr=&id=QzGuHnDhVzIC&oi=fnd&pg=PR7&dq=Norvig,+Peter+(1992).+Paradigms+of+Artificial+Intelligence+Programming.+New+York:+Morgan+Kaufmann+Publishers.+pp.+151%E2%80%93154.+ISBN+1-55860-191-0&ots=pVvYzCW8tt&sig=5OaZec96OmCrmCvB3szp9c9xWrs)
- Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8, 842–866. https://doi.org/10.1162/tacl_a_00349
- Ronkowitz, K. ELIZA: a very basic Rogerian psychotherapist chatbot. New Jersey Institute of Technology,
- Solomonoff, G. (2023). The meeting of the minds that launched AI there's more to this group photo from a 1956 AI workshop than you'd think. *IEEE Spectrum, Guest Article(History Of Technology)*. <https://spectrum.ieee.org/dart-mouth-ai-workshop>.
- Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., and Liu, Y. (2021, April). RoFormer: Enhanced transformer with rotary position embedding. Retrived October 20, 2023, from <http://arxiv.org/abs/2104.09864>
- Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE conference on computer vision and, 2014-openaccess.thecvf.com*. October 20, 2023, from http://openaccess.thecvf.com/content_cvpr_2014/html/Taigman_DeepFace_Closing_the_2014_CVPR_paper.html
- Tasci, T. and Kim, K. (2015). Imagenet classification with deep convolutional neural networks. Retrived October 20, 2023, from <https://pdfs.semanticscholar.org/09b8/120cbc52e7df46122e8e608146289fddbdfa.pdf>
- Wang, F. Y., Zhang, J. J., Zheng, X., Wang, X., Yuan, Y., Dai, X., Zhang, J. and Yang, L. (2016). Where does AlphaGo go: From church-turing thesis to AlphaGo thesis and beyond. *Ieeexplore.Ieee.org*. Retrived October 20, 2023, from <https://ieeexplore.ieee.org/abstract/document/7471613/>
- Weizenbaum, J. (1966, January). ELIZA-A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45. <https://doi.org/10.1145/365153.365168>

- Weizenbaum, J. (1976). Computer power and human reason: From judgment to calculation. Retrived October 20, 2023, from <https://psycnet.apa.org/record/1976-11270-000>
- Yang, J., Chesbrough, H., and Hurmelinna-Laukkanen, P. (2020). The rise, fall, and resurrection of IBM Watson Health. Retrived October 20, 2023, from <https://oulurepo.oulu.fi/bitstream/handle/10024/27921/nbnfi-fe2020050424858.pdf?sequence=1>
- Zadrozny, W., Gallagher, S., Shalaby, W., & Avadhani, A. (2015, February). Simulating IBM Watson in the classroom. *dl.acm.org*, 72–77. <https://doi.org/10.1145/2676723.2677287>
- Zhang, B., Sennrich, R. (2019). Root mean square layer normalization. *Advances in Neural Information Processing Systems, 2019•proceedings.neurips.cc*. October 20, 2023 <https://proceedings.neurips.cc/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html>
- Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE international conference on computer vision, 2015-cv-foundation.org*. October 20, 2023 https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Zhu_Aligning_Books_and_ICCV_2015_paper.html