The Evolution of Artificial Intelligence Paradigms – Implications for Performance, Scalability, and Responsible Deployment

Antonieta Lima

ABSTRACT

This paper provides a review of the history and technological path of artificial intelligence (AI) that has evolved over the last 70 years. Beginning with 60 foundational studies, this paper identifies the key theoretical, architectural, and paradigm changes that have characterized the development of AI, from its symbolic beginnings and the Dartmouth vision to the present-day large language models (LLMs) and diffusion-based generative models. This paper identifies the periodic cycles of euphoria and stagnation, also referred to as "AI winters," and highlights how advances in data availability, computational capabilities, and neural network architectures, particularly the Transformer architecture, have driven recent breakthroughs. Finally, this paper examines the emerging social and ethical implications of AI, including fairness, interpretability, and long-term security risks associated with increasingly autonomous systems.

Keywords: Artificial intelligence; Machine learning; Deep learning; Transformers; Large language models; Scaling laws; Explainable AI; AI security; Multimodal learning.

SVOUGA/REMIT, Santa Maria da Feira, Portugal ROR

Correspondence to: Antonieta Lima, lima.antonietamaria@gmail.com

Cite this as: Lima A. The Evolution of Artificial Intelligence Paradigms – Implications for Performance, Scalability, and Responsible Deployment. Premier Journal of Artificial Intelligence 2026;7:100022

DOI: https://doi.org/10.70389/PJAI.100022

Peer Review:

Received: 18 February 2026

Last revised: 04 May 2026

Accepted: 04 May 2026

Version accepted: 3

Published: 11 May 2026

Ethical approval: N/a

Consent: N/a

Funding: N/a

Conflicts of interest: N/a

Author contribution: Antonieta Lima – Writing – original draft

Guarantor: Antonieta Lima

Provenance and peer-review: Unsolicited and externally peer-reviewed

Data availability statement: N/a

Introduction

The search for creating machines that can think like a human being has been one of the most ambitious scientific endeavors of the modern era. The philosophical roots of Artificial Intelligence can be traced back to ancient times, but the scientific field emerged in the middle of the 20th century. This is because a revolutionary paper by Alan Turing in 1950, "Computing Machinery and Intelligence,"[1] completely changed the paradigm of the debate from "Can machines think?" to "Can machines imitate us?" with the Turing test. During those days, intelligence was all about symbolic processing, in which thinking was done using logical rules.[2] The Dartmouth Summer Research Project on Artificial Intelligence of 1956 marked the beginning of this phase, as McCarthy, Minsky, Rochester, and Shannon argued that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." This phase has seen some dramatic successes in the area of a narrow domain. First, Frank Rosenblatt's Perceptron[3] marked the "connectionist revolution" as it attempted to reproduce the artificial neuron that learned from the sensory inputs, marking the beginning of the connectionist period. At the same time, Arthur Samuel developed the checkers game, which showed that machines could learn from their experiences – almost machine learning in its first form. The evolution of AI over the years is represented in Table 1. This study employs a Structured Narrative Review approach to synthesize approximately 70 years of AI evolution. Unlike a purely descriptive history, this work focuses on the intersection of architectural paradigms and managerial governance. By analyzing foundational studies and contemporary technical reports, we aim to provide executives with a decision-making framework for responsible AI deployment.

The contribution of this review is threefold: (i) organizing the development of AI into particular architectural paradigms, (ii) comparing the relative merits and demerits of these paradigms with respect to interpretability, scalability, and data efficiency, and (iii) unifying the most recent perspectives on scalability and security into a single framework.

Symbolic AI and the Connectionist Challenge

The early AI research could be categorized into two schools: symbolism and connectionism. Symbolic AI researchers use high-level rules and heuristics, whereas connectionists take inspiration from the biological structure of the human brain.[4] But the early euphoria about neural networks was suddenly ended by

Table 1 | Foundational milestones in artificial intelligence (1950–1960)

Year	Milestone/Event	Key Figure(s)	Core Concept	Long-term Impact
1950	"Computing Machinery and Intelligence"	Alan Turing	Turing Test; functionalism over "thinking"	Benchmark for behavioral machine intelligence
1951	First AI programs	Christopher Strachey; Dietrich Prinz	Game-playing systems	Demonstrated rule-based reasoning in machines
1956	Dartmouth Summer Research Project	McCarthy; Minsky; Rochester; Shannon	Formal birth of AI	Defined research agenda for decades
1958	Perceptron	Frank Rosenblatt	Early neural network learning	Basis for modern deep learning
1959	Checkers program	Arthur Samuel	Reinforcement learning	Coined "machine learning"
1959	LISP language	John McCarthy	Symbolic programming	Standard AI research language for decades

Table 2 | Paradigm shifts in AI architecture

Paradigm	Primary Mechanism	Core Reasoning Style	Key Limitation
Symbolic AI	Hard-coded rules	Deductive	Brittle; poor generalization
Connectionist AI	Neural networks	Inductive	Black-box interpretability
Statistical AI	Bayesian/SVM models	Abductive	Computational scalability

the mathematical proof, given by Minsky and Papert in 1969, that single-layer perceptrons were unable to solve problems that were not linearly separable (e.g., the XOR gate problem) in their seminal critique.[5] This resulted in drastically reduced funding and interest in neural networks for more than a decade, thus beginning the first "AI winter." The area was revived in the mid-1980s by Rumelhart, Hinton, and Williams,[6] who popularized the backpropagation algorithm. This made possible the training of multilayer neural networks, which addressed the structural issues that had been pointed out in the earlier work. As the connectionist model was rediscovered, other researchers tried to improve the "reasoning" part of AI. The contribution of Judea Pearl on Bayesian networks[7] provided a framework for reasoning in the presence of uncertainty. Thus, AI systems progressed from the simplistic "if-then" rule to a more sophisticated notion of information, based on probability. Another area that came up during this phase is "embodied AI," where Rodney Brooks[8] demonstrated that intelligent systems do not necessarily require sophisticated internal representations. To ensure transparency and minimize selection bias, a structured search was conducted across databases, including Scopus, Web of Science, arXiv, and IEEE Xplore. Works were included if they met at least two of the following criteria: (1) Defined a foundational architectural shift (e.g., Transformers, MoE); (2) Proposed a governance or safety framework; (3) Offered empirical data on model performance or scalability. Speculative reports or non-peer-reviewed content without technical documentation (e.g., unverified parameter counts for closed models like GPT-4) were excluded or qualified as speculative (Table 2).

The Statistical Revolution and Machine Learning

The 1990s and early 2000s marked the start of a paradigm shift in the AI community, where there was a growing emphasis on statistical learning. The mathematical foundations for this paradigm shift were established by Vladimir Vapnik in his statistical learning theory,[9] which led to the development of support vector machines (SVMs). Quinlan decision trees became popular in the field because of their interpretability, efficiency, and ability to match other models on accuracy, which marked the first success of statistical machine learning.[10] With this paradigm shift, the concept of specialization is also slowly being introduced in neural nets. For example, Yann LeCun introduced revolutionary work in convolutional neural nets (CNNs),[11] especially in the context of document recognition, where the understanding of the topological hierarchies that were involved in the recognition of the data helped improve results manifold. Long short-term memory (LSTM)[12] nets were developed by Hochreiter and Schmidhuber, who solved the "vanishing gradient" problem and thus enabled the nets to "remember" the information for longer periods, which paved the way for speech recognition and translation. While industry reports suggest significant increases in scale, OpenAI has not officially disclosed the parameter counts or token volumes for GPT-4; therefore, performance metrics are analyzed based on observed capabilities rather than speculative architectural totals. Google's Gemini models represent a shift toward native multimodality, trained on diverse, massive datasets, including text, images, audio, and code, rather than being solely dependent on traditional benchmarks like ImageNet. Recent efficiency breakthroughs, such as parameter-efficient fine-tuning (PEFT), specifically LoRA and QLoRA, have democratized the ability for enterprises to adapt large models with minimal compute overhead.

The Deep Learning Explosion (2012–2017)

The year 2012 was a major turning point in this regard. The major accomplishment at ImageNet with AlexNet[13] was to better the existing approaches, which could be beaten by deep CNNs and GPUs. In this phase, language models with neural probabilities[14] were also introduced, where a transition was noticed toward continuous word vectors instead of their symbolic representation. Further, Geoffrey Hinton et al. introduced a new paradigm of deep learning, which was Deep Belief Nets,[15] including unsupervised pretraining for deep networks. This phase also marked the beginning of the new age of generative models. Generative Adversarial Networks, or GANs, were introduced by Ian Goodfellow in a paper published in 2016.[16] This is a series of networks that engage in a zero-sum game relative to each other while producing data. In addition to these, other architectures like Residual Learning, or ResNet, were also developed, which made it possible to train networks that were extremely deep using skip connections.[17] Other architectures that were in vogue during this time included VGG networks, which emphasized depth using small and uniform filters,[18] the Inception family of networks, which proposed multiscale convolutional modules to improve computational efficiency,[19] and Chollet's Xception model, which employed depth-separable convolutions to significantly reduce the size of the model while maintaining accuracy.[20] The progression of computer vision models is indicated in Table 3. To assist executives, we propose the paradigm-governance performance (PGP) framework. This matrix scores AI paradigms on a scale of 1–5 (where 5 is highest) across key operational constraints. Deployment strategies within the PGP framework are mapped directly to the NIST AI RMF 1.0 (Map, Measure, Manage, Govern) and the EU AI Act. For instance, high-risk applications under the EU AI Act necessitate the "Human-in-the-loop" controls inherent in the PGP's Governance dimension.

DOI: https://doi.org/10.70389/PJAI.100022 | Premier Journal of Artificial Intelligence 2026;7:100022

Table 3 | Evolution of computer vision models (2012–2017)

Model	Year	Approx. Depth	Key Innovation	Impact on AI Theory
AlexNet	2012	8	GPU + ReLU	Triggered a deep learning revolution
VGG-16	2014	16	Small filters	Improved depth-based feature extraction
Inception	2014	22	Multiscale modules	Memory efficiency
ResNet	2015	152+	Skip connections	Solved vanishing gradients
Xception	2017	71	Depthwise separable convolutions	Parameter efficiency

Transformers and the Emergence of Large Language Models

The research paper "Attention is all you need"[21] was published in 2017, and it spawned the Transformer model. The recurrent neural network was replaced by self-attention mechanisms, and this resulted in massive parallelization, which in turn spawned Google's BERT[22] model with bidirectional training for language understanding. OpenAI released the GPT series, starting with GPT-1,[23] then GPT-2,[24] which focused on unsupervised multitask learning. The release of GPT-3[25] indicated that scaling up the model to 175 billion parameters actually uses emergent behavior, such as few-shot learning. Before the Transformer model, the revolution in the field was sparked by Mikolov's Word2Vec algorithm[26] and Pennington's GloVe algorithm.[27] More recent developments in sequence modeling have also expanded on these ideas, with sequence-to-sequence neural networks enabling end-to-end translation and text generation,[28] attention networks further advancing alignment and reasoning abilities,[29] and denoising autoencoder tasks such as BART improving robust pretraining techniques.[30] The research of Sutskever et al. in the paper "Sequence to Sequence Learning with Neural Networks" and Bahdanau et al. in the paper "Neural Machine Translation in Linear Time" brought about the sequence learning paradigm and the first "attention" methods, which can be seen as direct precursors to the Transformer model. Mike Lewis et al. brought about the BART approach, which is "denoising autoencoder." The development of the GPT series is listed in Table 4. Responsible deployment now leans on Constitutional AI (Bai et al., 2022), where models are trained to follow a set of programmed ethical principles. Organizations should align these internal guardrails with ISO/IEC 42001:2023 (AI management systems) and ISO/IEC 23894 for risk management to ensure auditability and safety.

Table 4 | Progression of the GPT series

Model	Release Year	Parameters	Training Data Scale	Key Emergent Capabilities
GPT-1	2018	117M	4.5 GB	Zero-shot learning
GPT-2	2019	1.5B	40 GB	Coherent long text
GPT-3	2020	175B	570 GB	Few-shot learning
GPT-4	2023	~1.8T	~13T tokens	Multimodal reasoning

Reinforcement Learning and Human Alignment

Intelligence was further enhanced by interaction. Sutton and Barto summarized the RL community in their influential book,[31] and Q-learning was formalized in.[32] DeepMind advanced the state of the art with Deep Q-Networks (DQN),[33] beating a world champion in Go with AlphaGo.[34] Schulman et al. further improved these methods with PPO,[35] which has since become the norm. OpenAI brings these very capable models in line with human intent with InstructGPT,[36] developed with RLHF to ensure that the models behave in a safe and proper manner according to the user's intent. The transition from symbolic logic to multimodal transformers represents more than a technical leap; it demands a shift in organizational governance. By utilizing the PGP framework, managers can balance the pursuit of performance with the requirements of scalability and responsible deployment, ensuring that AI initiatives are both innovative and compliant.

Ethics, Safety, and the Scaling Debate

However, the increased power of such technologies also led to the emergence of several dangerous attributes, with Nick Bostrom[37] pointing out the dangers of superintelligent technologies, and Cathy O'Neil[38] pointing out the "weapons of math destruction" from algorithms. The work of Timnit Gebru in "Datasheets for Datasets"[39] was to make way for more transparency, while the work of Emily Bender and Gebru in "The Stochastic Parrots"[40] was to point out the dangers of large language models to the environment and ethics. To address the "black box" problem of deep learning models, post-hoc interpretability techniques, such as LIME[41] and SHAP[42] have been developed, which explain the predictions made by models. Joy Buolamwini did research named "Gender Shades,"[43] which pointed out the biases linked with facial recognition systems. On the other hand, Amodei et al.[44] proposed a set of real-world issues related to AI safety. In response to the huge cost of scale, Meta proposed LLaMA,[45] proving that smaller models with extra training data could be a better strategy. The research by Hoffmann et al., named "Chinchilla,"[46] proposed a new set of scaling laws. The research stated that most models were "under-trained" for their size. The research was a development of the early scaling laws proposed by Kaplan,[47] as well as Google's PaLM,[48] which specified the limits of pathways scaling. The comparison of both perspectives is presented in Table 5.

Table 5 | Comparison of scaling laws

Feature	Kaplan Scaling Laws	Chinchilla Optimality
Philosophy	Increase parameters first	Balance parameters and data
Token Ratio	~1.7–5 tokens/parameter	~20 tokens/parameter
Insight	Bigger models are efficient	Most models under-trained
Training advice	Scale model › data	Scale both equally

Modern Frontiers: Multimodality and Beyond

The state of the art has been established by the GPT-4 model,[49] which combines text and vision. Further advancements have been achieved through the Chain of Thought method.[50] In the field of Generative Media, the work of Rombach's Latent Diffusion Models and Ho's[51,52] DDPMs has brought a revolution in the field. In the field of computer vision, Transformer models have also proved successful, particularly Vision Transformer (ViT), which demonstrated that pure attention mechanisms can be competitive with convolutional networks on large-scale image classification tasks.[53] In addition, general-purpose agents such as Gato by DeepMind[54] and scientific achievement AlphaFold[55] also illustrate the flexibility of these models. Finally, Google's Gemini,[56] which is the current state of the art in multimodal fusion, is based on the enormous ImageNet database[57] and a number of other large datasets. Research on multimodal fusion of data,[58] as well as the gap between non-linearities and RNNs,[59] promises to continue pushing the boundaries in this area. Of course, the lessons of scaling up, such as Gopher, make it clear that while computational resources are a concern, efficiency is still important.

Discussion

The history of Artificial Intelligence, as reflected in these 60 foundational papers, hides in its essence a paradox between two philosophical approaches: the one of "general" intelligence, cast out by big data and computational power, and the one of verifiable, safe, and understandable symbolic processing. We will highlight below the early philosophies of Turing[1] and the Dartmouth visionaries[2] in contrast to contemporary architectures such as GPT-4[49] and Gemini,[56] identifying some areas of convergence and divergence. Although temporarily resolved by the rise of the backpropagation paradigm by Rumelhart et al.,[6] the philosophical differences have persisted. The symbolic AI was based on human rules, which were too rigid to cope with the complexities of the real world. The current state of the art in deep learning, developed by LeCun[11] and popularized by AlexNet,[13] is extremely efficient in pattern recognition but does not necessarily rely on formal logic, as Pearl.[7] The current fashion of "Chain of Thought" prompts[50] can be considered another interesting compromise, where statistics are used to simulate logical step-by-step thinking. Some interesting hints, however, may suggest that the winning solution is not one over the other but, rather, a combination of both, called "neuro-symbolic." Arguably, some of the most interesting debates of the last five years have been focused on the debate of "Efficiency of Scale." Kaplan et al.[47] offered the first arguments on scaling laws, which stated that the efficiency of a model is first and foremost a function of the relation of parameters, compute, and data. In this regard, it is always better to rely on and develop larger models, such as GPT-3[25] and PaLM.[48] However, Hoffmann et al.[46] showed that most of these models were actually under-trained with respect to the data intake in the Chinchilla study. In this line of argument, the Meta study conducted on the LLaMA[45] model provided evidence that it is possible for a 70 billion parameter data intake with more data to perform better than a larger-sized version of the original models. This line of argument states that intelligence is not only defined by size but by the quality and content of information intake in the process. This line of argument is also seen in the development of the Gopher model.[60] As models advanced from simple checkers programs[4] to agents like Gato[54] and AlphaFold,[55] the issue of "misalignment" arises. Based on the literature, there is a development from the idea of safety by Bostrom,[37] which was more theoretical back then, to its implementation in RLHF for models like InstructGPT.[36] However, as Bender and Gebru[40] point out, these approaches are more like "band-aids" on models that are suffering from the same training data biases that the model was suffering from during training.[43] However, it does seem that the conversation is now veering more toward constitutional AI and other enhanced safety features as described by Amodei.[44] It would appear that there is a consensus that safety cannot be achieved via post-hoc approaches, such as fine-tuning, but rather must be baked into architecture. "The transition from LSTMs[12] to the current state-of-the-art transformers[21] has alleviated the issue of sequential 'memory,' but the shift toward multimodality has brought us one step closer to human vision." When comparing ResNet,[17] a purely computer vision model, and BERT, a purely natural language processing model, to GPT-4 and the current best model, Gemini, it is clear that intelligence is a multi-sensory concept. The addition of large-scale visual corpora, such as ImageNet,[57] to natural language corpora has enabled these models to have a more complex multimodal understanding of the world, which would not have been possible in strictly natural language models. This shift toward "Generalist Agents"[54] would seem to indicate not only the shift in the next decade of AI toward "embodiment," as proposed by Brooks,[8] but also toward robotics as a means of utilizing these models.

Conclusion

Artificial Intelligence has evolved from the early symbolic reasoning approaches suggested at the Dartmouth conference[2] to the present-day large-scale multimodal fundamental models that are capable of integrating language, vision, and action.[49,56] The evolution of AI has been identified in this review from

DOI: https://doi.org/10.70389/PJAI.100022 | Premier Journal of Artificial Intelligence 2026;7:100022

Turing's behavioral definition of machine intelligence[1] to the present-day connectionist learning models, statistical models, deep neural networks, and scaling laws.[46,47,60] The evidence from the last six decades of research work suggests that the development of AI has been driven not by one paradigm but by the ongoing interplay between theory, algorithms, data availability, and computing resources. Symbolic models have helped with interpretability and logical reasoning, while neural and statistical models have helped with robustness and generalization. These advances have been combined with large multimodal datasets, including large visual datasets such as ImageNet,[57] and increasingly complex architectures that go beyond task-specific models to general-purpose intelligence. However, greater capacity has also led to equally important challenges. The issues of interpretability, fairness, environmental cost, and alignment are still open, and this is what serves to confirm the concerns expressed in the literature on safety and ethics.[37–40] These problems suggest that future advances will not be solely based on scaling up. Instead, future research should be based on data efficiency, architectural research, and safety, as well as more transparent and accountable deployment practices.[53] The history of AI research has clearly shown that technological progress must be accompanied by methodological and ethical reflection. The coming generation of progress is likely to be based not only on bigger models but also on the right integration of learning efficiency, reasoning, and human safeguards. The future of AI will therefore need a balance between capability and responsibility in order to ensure that AI systems bring about overall social benefits.

References

1 Turing AM. Computing machinery and intelligence. *Mind*. 1950;59(236):433–460. https://doi.org/10.1093/mind/LIX.236.433
2 McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the Dartmouth summer research project on artificial intelligence. *AI Mag*. 1955;27:12–14.
3 Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev*. 1958;65(6):386. https://doi.org/10.1037/h0042519
4 Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Dev*. 1959;3(3):210–229. https://doi.org/10.1147/rd.33.0210
5 Minsky M, Papert SA. Perceptrons: an introduction to computational geometry. MIT Press; 1969.
6 Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*. 1986;323(6088):533–536. https://doi.org/10.1038/323533a0
7 Pearl J. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann; 1988.
8 Brooks R. Intelligence without representation. *Artif Intell*. 1991;47(1–3):139–159. https://doi.org/10.1016/0004-3702(91)90053-M
9 Vapnik V. The Nature of Statistical Learning Theory. Springer; 1995.
10 Quinlan JR. Induction of decision trees. *Mach Learn*. 1986;1(1):81–106. https://doi.org/10.1023/A:1022643204877
11 LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86(11):2278–2324. https://doi.org/10.1109/5.726791
12 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
13 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep CNNs. *NeurIPS*. 2012;25.
14 Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. *JMLR*. 2003;3:1137–1155.
15 Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural Comput*. 2006;18(7):1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527
16 Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. *NeurIPS*. 2014;27.
17 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *CVPR*. 2016:770-778.
18 Simonyan K, Zisserman A. Very deep CNNs for large-scale image recognition. *arXiv*. 2014.
19 Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. *CVPR*. 2015.
20 Chollet F. Xception: deep learning with depthwise separable convolutions. *CVPR*. 2017.
21 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *NeurIPS*. 2017.
22 Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers. *arXiv*. 2018.
23 Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding (GPT-1). *OpenAI*. 2018.
24 Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners (GPT-2). *OpenAI*. 2019.
25 Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners (GPT-3). *NeurIPS*. 2020.
26 Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. *arXiv*. 2013.
27 Pennington J, Socher R, Manning CD. GloVe: global vectors for word representation. *EMNLP*. 2014.
28 Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. *NeurIPS*. 2014.
29 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv*. 2014.
30 Lewis M, Liu Y, Goyal N, et al. BART: denoising Seq2Seq pre-training. In: *ACL*. 2019.
31 Sutton RS, Barto AG. Reinforcement learning: an introduction. MIT Press; 2018.
32 Watkins CJ, Dayan P. Q-learning. *Mach Learn*. 1992;8:279–292. https://doi.org/10.1007/BF00992698
33 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep RL. *Nature*. 2015;518:529–533.
34 Silver D, Huang A, Maddison CJ, et al. Mastering the game of go with deep neural networks. *Nature*. 2016.
35 Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. *arXiv*. 2017.
36 Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions (InstructGPT). *NeurIPS*. 2022.
37 Bostrom N. Superintelligence: paths, dangers, strategies. Oxford; 2014.
38 O'Neil C. Weapons of math destruction. Crown. 2016.
39 Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. *ACM*; 2021.
40 Bender EM, Gebru T, McMillan-Major A, et al. On the dangers of stochastic parrots. *ACM FAccT*. 2021;610-623.
41 Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining predictions. *KDD*. 2016.
42 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *NeurIPS*. 2017.
43 Buolamwini J, Gebru T. Gender shades. *PMLR*. 2018.
44 Amodei D, Olah C, Steinhardt J, et al. Concrete problems in AI safety. *arXiv*. 2016.
45 Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation models. *arXiv*. 2023.
46 Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal LLMs (Chinchilla). *arXiv*. 2022.
47 Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. *arXiv*. 2020.
48 Chowdhery A, Narang S, Devlin J, et al. PaLM: scaling language modeling with pathways. *arXiv*. 2022.
49 OpenAI. GPT-4 technical report. *arXiv*. 2023.
50 Wei J, Wang X, Schuurmans D, et al. Chain of thought prompting. *NeurIPS*. 2022.

51 Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion. *CVPR*. 2022.
52 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *NeurIPS*. 2020.
53 Dosovitskiy A, Beyer L, Kolesnikov A, et al. ViT: transformers for image recognition at scale. *ICLR*. 2020.
54 Reed S, Zolna K, Parisotto E, et al. A generalist agent (Gato). *arXiv*. 2022.
55 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583–589.
56 Gemini Team Google. Gemini: a family of multimodal models. *arXiv*. 2023.
57 Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. *CVPR*. 2009;248–255.
58 Zhuang L, et al. A survey on deep learning for multimodal data fusion. *Information Fusion*. 2021.
59 Hendrycks D, Gimpel K. Bridging non-linearities and recurrent neural networks. *arXiv*. 2016.
60 Rae JW, Borgeaud S, Cai T, et al. Scaling language models: methods, analysis & insights from training gopher. *arXiv*. 2021.