European Review, Vol. 33, No. 4, 416–425 © The Author(s), 2025. Published by Cambridge University Press on behalf of Academia Europaea. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

doi:10.1017/S1062798725100379

The Deep Learning Revolution in AI

SHALOM LAPPIN

School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, UK. Email: s.lappin@qmul.ac.uk

The early development of Artificial Intelligence (AI) in the latter half of the twentieth century was marked by limited, hand-crafted systems and fluctuating perceptions of the field's potential. Early research explored a range of paradigms – including symbolic, neural and probabilistic approaches – constrained by severe hardware and data limitations. Key technological advances, such as the invention of microchips, GPUs and later TPUs, significantly enhanced computational capacity, enabling more complex AI experimentation. Concurrently, the proliferation of digital data through the internet addressed longstanding bottlenecks in data availability. The most transformative shift, however, came from architectural innovations in neural networks, culminating in the deep learning revolution. This unfolded in two phases: the emergence of Recurrent and Convolutional Neural Networks, followed by the development of transformer-based models, which underpin today's Large Language Models (LLMs).

Introduction

During the early years of Artificial Intelligence, in the second half of the twentieth century, AI systems consisted of small-scale prototypes with very limited coverage. Many of their rules and architectural elements had to be hand-crafted. The field was generally regarded as speculative, with limited prospects for marketable engineering applications. It experienced alternating periods of hyperbolic optimism (AI summers) and acute scepticism (AI winters). Work in AI was often precariously perched on the margins of tech company research labs, and in the more remote parts of academic computer science programs (for a brief history of AI see Lappin 2025).

In this period, AI researchers experimented with a diverse set of approaches. These included symbolic systems (grammars, logics, constraint satisfaction provers, frames, and scripts), neural networks, (perceptrons, and multilayer feedforward

https://doi.org/10.1017/S1062798725100379 Published online by Cambridge University Press

networks), and probabilistic methods (Bayesian networks, Ngrams, Hidden Markov Models, and a variety of stochastic modelling procedures). Hardware limitations were a major factor in limiting progress in AI. Computers in the 1950s and 1960s ran on vacuum tubes, with data stored on magnetic tape, and programming done by punch cards. The processing and memory capacities of these devices could not support the rapid development and testing of large-scale systems.

The creation of more efficient computing hardware had a dramatic impact on AI. Intel introduced the first microchip central processing unit (CPU), its 4004, in 1971. In 1999, Nvidia released the first GPU processor, the GeForce 256. Unlike CPUs, GPUs consist of multiple processing units that perform a large number of operations in parallel. Google deployed its first tensor processing unit (TPU) in 2015, dedicated to linear algebraic operations required for the machine learning of neural networks.

The absence of significant quantities of digital data was a second major obstacle to progress in AI research. It was necessary to produce most of this data by hand, often separately for each task. The emergence of the internet in the 1990s generated a vast trove of digital data in a variety of modalities. Large quantities of text, visual images, videos, and audio files became easily accessible online.

However, the most important breakthroughs that generated the dramatic rise of current AI systems consisted of a series of radical architectural innovations in neural networks, which produced the deep learning revolution. This revolution came in two phases. The first yielded powerful Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs). The second produced transformers, which drive Large Language Models (LLMs).

Deep Learning Models Phase 1

Early feedforward neural networks lacked a memory to retain information from previous processing states for a sequence of input data. As a result, they were not able to identify and keep track of long-distance relations among elements of these sequences. The main subject-verb dependency in 1, and the left-right bracket matching in 2, illustrate relations of this kind.

1. *The candidates* interviewed for the position at the university where my friend teaches *give* a talk on their research.
2. $(_1\ (_2\ (_3\ (_4)_4\ (_5)_5)_3)_2\ (_6\ (_7\ (_8)_8\ (_9)_9)_7)_6)_1$

Elman (1990) introduced Recurrent Neural Networks, which are sequential processing systems with a memory that retains information from a previous state and passes it forward to its successor. Figure 1 shows the architecture of a simple RNN.

While simple RNNs are able to track certain long-distance relations, they do not filter or control the information that is passed forward to successive states. They are limited in their ability to recognize complex patterns in extended contexts.

Hochreiter and Schmidhuber (1997) constructed a more powerful RNN, the Long Short Term Memory network (LSTM). Each of the processing units of an LSTM

https://doi.org/10.1017/S1062798725100379 Published online by Cambridge University Press

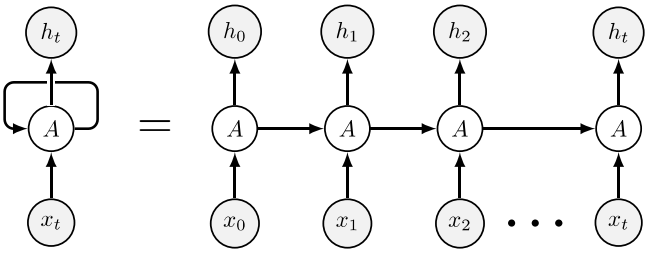


Figure 1. A simple Recurrent Neural Network (from Lappin 2021).

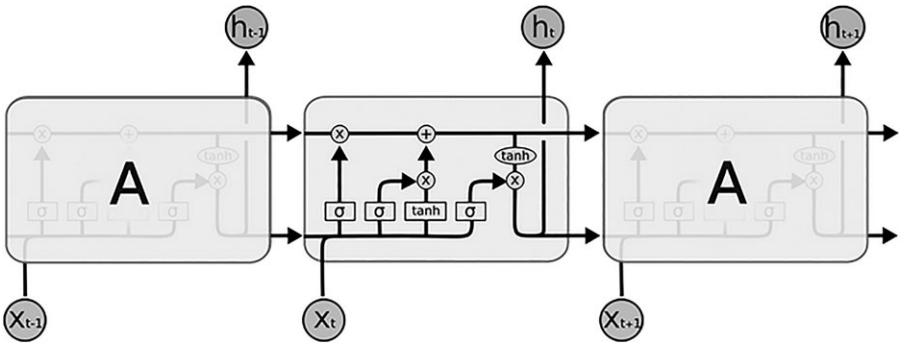


Figure 2. An LSTM (from Olah 2015, with permission).

applies three functions to its input, as filtering gates, to determine which information is passed on to the next processing phase. The input data string is represented as a vector, which is a sequence of real numbers. Figure 2 displays the structure of an LSTM.

Convolutional Neural Networks (LeCun *et al.* 2010) identify a feature map from input data by recognizing subcomponents of these data. A pooling layer reduces the dimensions of the features to produce a compressed map, and to render it stable under small variations. Successive convolutional and pooling layers produce increasingly higher-level representations of the objects in the data. A softmax function generates a probability distribution over different possible classifications of the vector. CNNs have achieved good results for applications in visual image identification and speech recognition. Figure 3 illustrates the architecture of a CNN.

Neural networks learn to determine weights for the elements of their input data, which cause their processing units to 'fire' (produce a specified value as output) when these weights are above a given threshold. Training a neural network involves assigning random values to their inputs, and then incrementally correcting these against the data to which it is exposed. The error rate of the network's output is calculated as the gradient (angle) of a loss function. The learning cycles seek to minimize the values of the loss function. They proceed down the slope of the function

https://doi.org/10.1017/S1062798725100379 Published online by Cambridge University Press

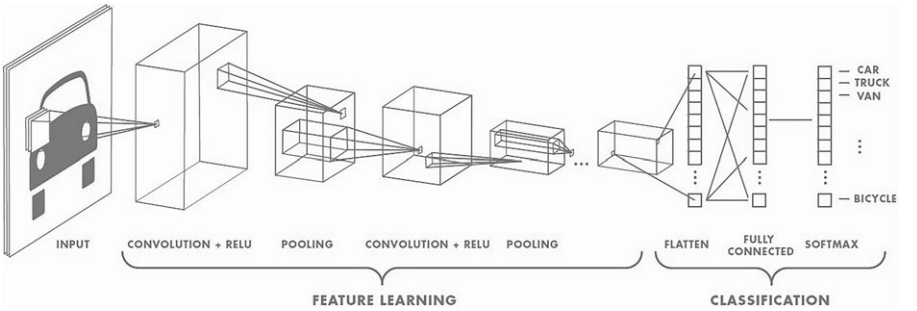


Figure 3. A CNN (from Saha 2015, with permission).

by incremental correction, until a statistically estimated optimal point is reached. This learning procedure is known as *stochastic gradient descent*.

This procedure is generally applied through *backpropagation* (introduced by Rumelhart *et al.* 1986), where one feeds the difference between the desired and the actual weights of the network backwards along the connections among the units, to calculate updates for the weights of the training inputs. Backpropagation can encounter one of two serious difficulties when used in training RNNs, CNNs, and feed-forward networks. If it is not constrained, the gradient of the loss function can become vanishingly small, or it can explode into unmanageably large values.

RNNs and CNNs were the primary vehicles of the first generation in the deep learning revolution. They greatly increased the accuracy and coverage of neural networks. They provided the first set of AI systems that offered the prospect of effective wide-coverage models that could be applied to a wide range of cognitively interesting AI tasks. These included, among others, machine translation, text understanding, text generation, information retrieval, image recognition, complex object classification, image description, medical diagnostics, and expert risk assessment.

Much of the power of Deep Neural Networks (DNNs) resides in the vector representations of the data on which they are trained. These representations, known as embeddings, express a variety of distributional relations among elements of these data. In the case of words, they provide a conceptual map of the co-occurrence patterns of the terms extracted from the training data. These patterns encode semantic and syntactic relations among the expressions of the corpora. Multimodal embeddings greatly expand the power and range of DNNs by permitting them to track relations between linguistic expressions and non-linguistic environments.

Adding *attention* (Bahdanau *et al.* 2015) to DNNs significantly improved their ability to identify subtle patterns in data, and to recognize connections among the elements of their input. An attention layer produces a dynamic context vector that is updated with each new token of the input sequence. The vector effectively preserves the information of each hidden state of the part of the DNN that encodes input data. It assigns different weights to these states on the basis of their relations to the output that the decoder produces. The attentional context vector learns to align the weights

https://doi.org/10.1017/S1062798725100379 Published online by Cambridge University Press

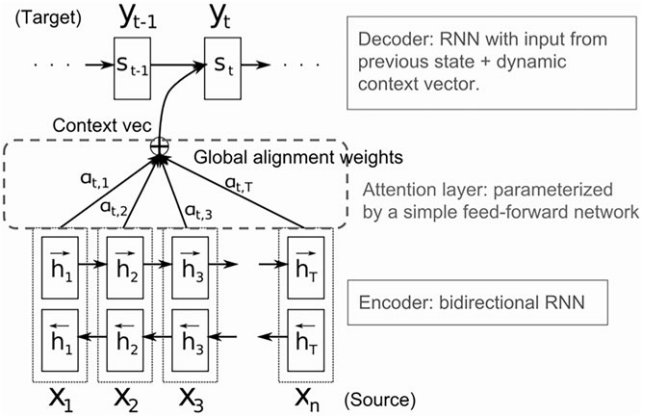


Figure 4. A bidirectional LSTM with an attention layer (from Bahdanau *et al*. 2015, with permission).

of the encoded input with the elements of the decoded output. Figure 4 indicates the structure of a bidirectional LSTM with an attention layer.

Embeddings and attention provide key elements of transformers, the DNNs that have given rise to the current phase of the deep learning revolution.

Deep Learning Models Phase 2

Vaswani *et al*. (2017) introduce transformers. These are encoder–decoder DNNs that consist entirely of blocks of multi-head attention units, connected by normalization and feedforward layers. An attention block computes self-attention weights for the token of an input sequence, where self-attention is the relative importance of each token in relation to the others. Each block of multi-head self-attention units in a transformer can be dedicated to a different dimension of information concerning its input. In this way, transformers can identify patterns and connections among the elements of the data on which it is pre-trained, across a wide range of features. As a result, transformers are generally more successful than RNNs and CNNs at recognizing long distance relations among elements in input. Figure 5 shows the architecture of a transformer.

Transformers are pretrained on large data sets from which they extract embeddings. They can then be fine-tuned for specific applications by adding new dedicated layers of attention heads, or by retraining some of its pre-trained layers. This is a procedure for customizing the model to a particular domain and a set of applications. Transformers can be efficiently trained because the blocks of attention heads can be trained independently of each other, in parallel. Due to their normalization layers and the fact that they retain information for input throughout all phases of processing, they avoid the vanishing and the exploding gradient problems that beset the earlier generation of DNNs.

https://doi.org/10.1017/S1062798725100379 Published online by Cambridge University Press

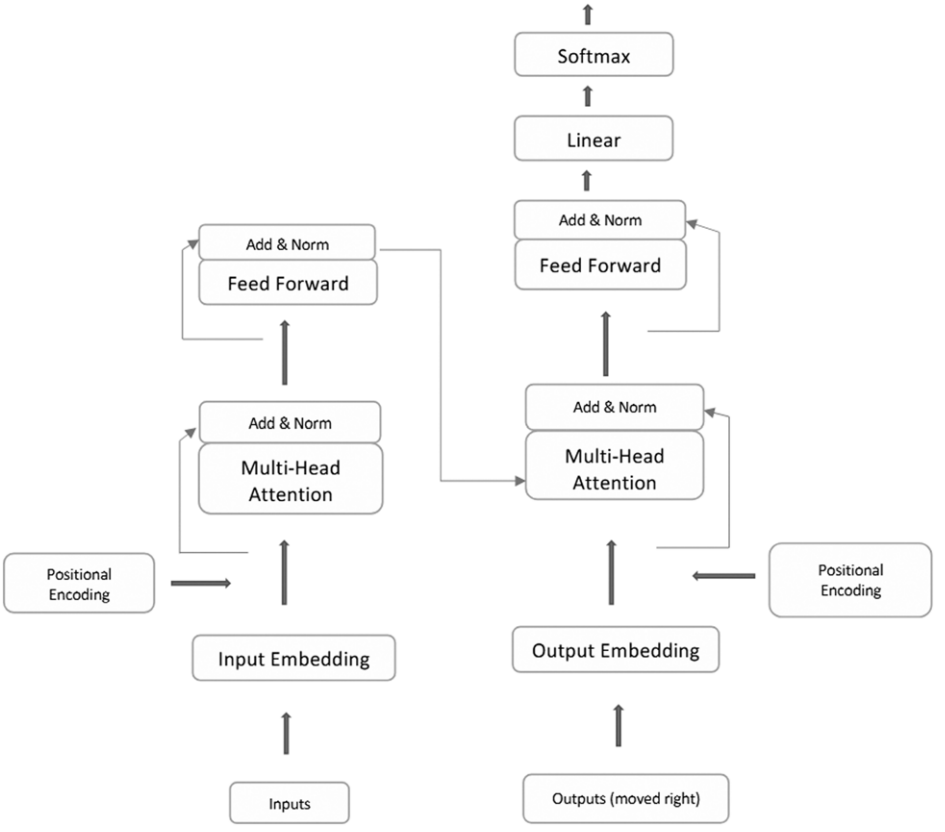


Figure 5. The architecture of a transformer (from Lappin 2021, based on Vaswani *et al.* 2017).

Transformers can be trained on multi-modal data without revising their architecture. Their attention heads can recognize connections among text, visual images and sound in the same way as they do among the terms of purely linguistic input. Transformers constitute a significant breakthrough in deep learning. They are driving the large data models that have yielded the impressive results currently on display in many areas of AI.

Bender *et al.* (2021) and Chomsky *et al.* (2023) claim that the LLMs that transformers sustain are little more than 'stochastic parrots' returning the data on which they are pretrained. In fact, this assertion is incorrect. They identify abnormalities in X-rays and predict new molecular structures for proteins, within the medical domain. They learn complex new verbal commands in virtual environments. They identify scenes in graphic images, and they produce accurate descriptions for these scenes. They generate new computer code to solve difficult programming

https://doi.org/10.1017/S1062798725100379 Published online by Cambridge University Press

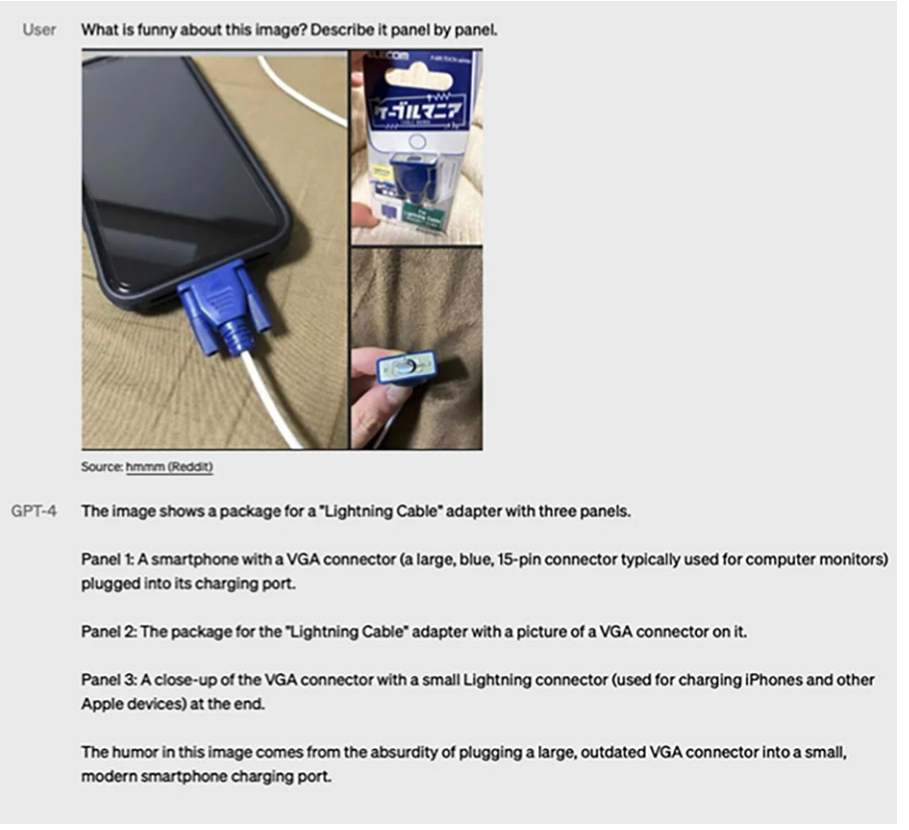


Figure 6. ChatGPT-4 interprets a sequence of images (from OpenAI 2023, arXiv.org non-exclusive license to distribute).

problems, and they create novel proofs of mathematical theorems. The inferences that sustain these achievements cannot be reduced to simple variations of previously encountered regularities in data.

Figure 6 exhibits ChatGPT-4's ability to interpret a sequence of photographic images and to identify the source of the humour in the main photograph in the left panel.

Problems with LLMs

It is important to recognize that transformers are designed to generate the next token in a sequence on the basis of the tokens encountered up to that point, or from the elements of the token's left and right context. They do not attend to the factual content of their input, and they have no way of doing so. LLMs frequently produce fluent, plausible text, which is entirely fictional. LLM 'hallucination' can create serious difficulties. In one instance, a lawyer in the US who relied on ChatGPT-3 to

https://doi.org/10.1017/S1062798725100379 Published online by Cambridge University Press

find legal precedents unknowingly went to court with a set of fabricated cases to support his argument (Weiser 2023). Devising effective procedures for verifying LLM output remains one of the major challenges posed by contemporary AI systems (see Lappin 2024, 2025, on LLM hallucination and the problem of verification).

LLMs have created, or intensified, a variety of significant problems, which pose important issues of public policy (Lappin 2025). As a partial list, consider the following. Pre-training LLMs requires very large amounts of energy and water. Chip production is also energy-intensive, and it produces toxic chemical waste. The servers used to run the online systems that use LLMs, and to store the data needed for these applications, occupy substantial tracts of land. The demands for energy and open land are causing serious environmental problems. These sorts of difficulties are not often associated with the AI revolution in public discourse.

It is primarily large tech companies that have the funds and computing resources to develop and train LLMs. They have achieved quasi-monopoly status in controlling the design and development of AI systems. Universities, public scientific research institutes and small start-ups are becoming clients of the tech companies, in fine-tuning and applying their deep learning models. Consequently, many problems in basic science and engineering that motivated earlier work in AI are being marginalized. In particular, some of the questions about the nature of learning and cognition are no longer in the focus of AI research. Most tech companies are not prioritizing the development of smaller, less computationally intensive models, which can identify significant patterns with much less training data. This sort of research demands long-term investment, and so it does not yield rapidly marketable results.

The capacity of LLMs to produce deep fakes in text, visual and audio modalities facilitates disinformation, identity theft, and a variety of other sorts of harm. AI-powered bots are pumping out false content on social media in the service of extremists, trolls and malign government actors. This stream of digital deception is undermining trust in foundational institutions, such as fair elections and public-health measures. It is also used to propagate racism, misogyny, and other forms of hate speech.

Assessment systems powered by AI models incorporate a variety of ethnic, religious, class, and gender biases in credit and hiring decisions, as well as recommendations for medical treatment. These models can cause serious discrimination against people from the populations against which the bias is encoded.

The prospect of large-scale automation through expert AI systems, and AI-supported robotics, across a wide spectrum of skilled and unskilled professions, raises the possibility of large-scale unemployment. In previous technological revolutions, job loss due to transformation in one industry was offset by the emergence of new areas of production and service. If AI-based models automate a large enough range of industries, a balancing demand for labour in emerging industries might not be available. This could generate major economic and social dislocation.

https://doi.org/10.1017/S1062798725100379 Published online by Cambridge University Press

Conclusions

The early period of AI saw the proliferation of diverse approaches to learning and representation tasks (feedforward neural networks, rule-based symbolic frameworks, and statistically driven models). Hardware and digital data limitations prevented most of these approaches from developing successful wide-coverage systems. Hardware innovation and radical changes in neural-network architecture gave rise to the deep learning revolution, which came in two phases.

The first phase saw the emergence of powerful RNNs and CNNs. In the second phase these systems increasingly gave way to transformers, which consist entirely of blocks of attention heads. The LLMs that they sustain have achieved striking improvements in performance and coverage across a wide spectrum of applications. A central element of this progress is the capacity of transformers to transfer high-accuracy learning to new tasks and domains, with minimal additional training. These models can apply their pretraining to new tasks through small-scale fine-tuning. They retain the same architecture over a large range of tasks. This avoids the need for task-specific redesign, or extensive training for each new application.

The deep learning revolution has produced powerful AI systems that are yielding substantial benefits in many different domains, from health care, through natural language processing, to music and the fine arts. However, it has also produced a series of pressing environmental, social and economic problems that require urgent responses. These issues should be the focus of informed public discussion. In many cases, effective regulation of these systems, and the large tech companies that produce and market them, is essential to derive social benefit from them.

Acknowledgements

This paper is a version of the talk which the author presented at the Academia Europaea annual conference 'Building Bridges', in Wroclaw, on 28 November 2024. I developed several of the ideas discussed in the paper in greater detail in a series of three lectures on the AI revolution, which I gave in the School of Electronic Engineering and Computer Science, Queen Mary University of London, in October/November 2024. I am grateful to the audiences of all four talks for helpful comments and feedback.

Some of the research summarized in this paper was supported by Grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg. I bear sole responsibility for the views expressed, and for any errors that remain.

https://doi.org/10.1017/S1062798725100379 Published online by Cambridge University Press

References

Bahdanau D, Cho K and Bengio Y (2015) Neural machine translation by jointly learning to align and translate. *Third International Conference on Learning Representations* (ICLR).

Bender E, Gebru T, McMillan-Major A and Shmitchell S (2021) On the dangers of stochastic parrots: can language models be too big? *FAccT '21*, Association for Computing Machinery, pp. 610–623.

Chomsky N, Roberts I and Watumull J (2023) The false promise of ChatGPT. *The New York Times*, 8 March 2023.

Elman J (1990) Finding structure in time. *Cognitive Science* **14**, 179–211.

Hochreiter S and Schmidhuber J (1997) Long short-term memory. *Neural Computation* **9**(8), 1735–1780.

Lappin S (2021) *Deep Learning and Linguistic Representation*. Oxford and Boca Raton: Taylor and Francis, CRC.

Lappin S (2024) Assessing the strengths and weaknesses of large language models. *Journal of Logic, Language and Information* (Special issue: Natural Logic Meets Machine Learning) **33**, 9–20.

Lappin S (2025) *Understanding the Artificial Intelligence Revolution: Between Catastrophe and Utopia*. Oxford and Boca Raton: Taylor and Francis, CRC.

LeCun Y, Kavukcuoglu K and Farabet C (2010) Convolutional networks and applications in vision. *Proceedings of 2010 IEEE International Symposium on Circuits and Systems* pp. 253–256.

Olah C (2015) *Understanding LSTM Networks*, blog, 7 August 2015.

OpenAI (2023) GPT-4 Technical Report. *arXiv*:2303.08774.

Rumelhart DE, Hinton GE and Williams RJ (1986) Learning representations by back-propagating errors. *Nature* **323**, 533–536.

Saha S (2015) A comprehensive guide to convolutional networks – the ELI5 way. *Towards Data Science*, 18 December 2015.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L and Polosukhin I (2017) Attention is all you need. *31st Conference on Neural Information Processing Systems* (*NIPS 2017*).

Weiser B (2023) Here's what happens when your lawyer uses ChatGPT. *The New York Times*, 27 May 2023.

About the Author

Shalom Lappin is Professor of Natural Language Processing in the School of Electronic Engineering and Computer Science at Queen Mary University of London, Emeritus Professor of Computational Linguistics in Informatics at King's College London, and Scientific Researcher in CLASP at the University of Gothenburg. He is also a Fellow of the British Academy and a Member of Academia Europaea.

https://doi.org/10.1017/S1062798725100379 Published online by Cambridge University Press