1

Generative AI

An Introduction

Tom Melham

Artificial intelligence (AI) is notoriously hard to define, but a useful shorthand description is *the capability of a machine to imitate intelligent human behaviour*.[1] One merit of this definition is that it makes no reference to the underlying technology; this is under constant and rapid development, and far too varied to encapsulate neatly as the defining characteristic of AI. This terse definition also avoids questions about whether our machines really are intelligent. Whether they *think*. Most importantly, this definition is a salutary reminder, in the context of the astonishing performance of modern *generative artificial intelligence* (GAI) – the subject of this research handbook in its connections with law. It reminds us that this is technology that generates outputs that look like – that *mimic* – the intellectual productions of humans, from everyday speech to creative literary works.

Of course, artificial intelligence has many potential and practical applications, beyond just mimicry of human intelligence. Artificial intelligence technologies are widely looked to for practical solutions to specific problems in particular application domains, such as disease diagnosis, personalised education, electronic retail, risk management, or fraud detection in finance – and many more. The European Union Artificial Intelligence Act (EU AI Act) is therefore rightly concerned with the regulation of a very broad range of both AI applications and technologies. Accordingly, the Act defines an 'AI system' to mean 'a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments'.[2]

The contrast between these two definitions illuminates a distinction that is commonly made between the two principal aims of artificial intelligence research. The first is so-called artificial general intelligence (AGI). The pursuit of AGI aims at creating machines that have an intelligence equal to humans – that can learn intellectual skills across a broad or even an unlimited range of human intellectual activities. It can learn to play chess, but also compose a poem, solve

[1] Paraphrased from 'artificial intelligence'. Meriam-Webster.com, 2024. www.merriam-webster.com/dictionary/artificial%20intelligence, 2 June 2024.

[2] The Artificial Intelligence Act, Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending Regulations (EC) No. 300/2008, (EU) No. 167/2013, (EU) No. 168/2013, (EU) 2018/858, (EU) 2018/1139, and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828, Art. 3, para. 1.

https://doi.org/10.1017/9781009492553.004 Published online by Cambridge University Press

a puzzle, or work out some legal reasoning. There are no examples of AGI in existence today – but the optimistic pursuit of AGI is a widespread mindset among AI researchers and in the tech industry. It is sometimes explicit in published objectives or aspirations, but also commonly betrayed by a tendency to yield to our human predisposition to anthropomorphise.[3] Some researchers or commentators also posit artificial intelligence that is self-aware, conscious, or sentient – about which there is a great deal of discussion and speculation.[4] In this chapter, we will leave this controversial topic aside – though should sentient artificial intelligence ever be achieved, this will present many obvious challenges to ethics and law.[5]

The second principal aim of artificial intelligence research is applied, or 'narrow', artificial intelligence. This is artificial intelligence designed to do a specific task – one narrow, specific task that usually requires human intelligence, but which we can 'train' the computer to do, typically using a significant quantity of data and special 'learning' algorithms. A famous example is AlphaGo, which can play the complex game of Go as well as the most accomplished human master.[6] That is all it can do, but it does that extremely well.

There are also two broad technical approaches to achieving artificial intelligence, whether general or applied. The first is so-called *symbolic* or *logical* artificial intelligence. This is the kind of artificial intelligence that represents the substance of focus for the artificial intelligence algorithm – whether it be knowledge, a virtual world, some goals, or rules governing simulated actions – in explicitly *symbolic* form. Abstract, usually structured, explicit representations of entities, their properties, and the relationships between them. Symbolic AI has a long history and was the dominant paradigm of AI research from the 1950s to the mid 1990s. And it is still useful today; for example, AlphaGo incorporates significant components of symbolic AI.

The second technical approach could, in very general terms, be called *statistical AI*. This is artificial intelligence that is based on something resembling a *statistical model* of the domain of interest. A characteristic example is an AI system that generates a decision outcome for a particular case in hand, based on a statistical summary of decisions that have been made in the past. Example applications include automation of a wide range of decision-based government services.[7] This approach to AI has been in the ascendant throughout the twenty-first century, animated by the remarkable successes of *machine learning* with *deep neural networks*.[8] Statistical AI is also the approach that underpins much of the generative artificial intelligence that is explored in this handbook, and therefore the focus of this introductory chapter.

Research in generative artificial intelligence has a long history, going back – in many different forms – to the founding of the modern field of artificial intelligence in the 1950s. By the early 2020s, powered by 'deep learning' with neural networks, highly capable generative AI had been developed that can produce many forms of intellectual output – including text in natural language (such as English), but also many other forms of output such as images, videos, music,

[3] N. Epley, A. Waytz, and J. T. Cacioppo, 'On seeing human: A three-factor theory of anthropomorphism', *Psychological Review*, 114(4), pp. 864–886.

[4] P. Butlin et al., 'Consciousness in Artificial Intelligence: Insights from the Science of Consciousness', *arXiv Computing Research Repository*, arXiv:2308.08708 [cs.AI], 22 August 2023.

[5] K. Mindaugas, 'Legal framework for the coexistence of humans and conscious AI', *Frontiers in Artificial Intelligence*, 6, 20 September 2023.

[6] D. Silver et al., 'Mastering the game of Go with deep neural networks and tree search', *Nature*, 529, pp. 484–489, 27 January 2016.

[7] V. J. Straub et al., 'AI for bureaucratic productivity: Measuring the potential of AI to help automate 143 million UK government transactions', *arXiv Computing Research Repository*, arXiv:2403.14712 [cs.CY], 18 March 2024.

[8] Y. LeCun, Y. Bengio, and G. Hinton, 'Deep learning', *Nature*, 521, pp. 436–444, 27 May 2015; T. J. Sejnowski, *The Deep Learning Revolution* (The MIT Press, 2018).

https://doi.org/10.1017/9781009492553.004 Published online by Cambridge University Press

or computer code. But it is no exaggeration to say that the introduction of the *transformer* model by researchers at Google Brain in 2017 revolutionised the field of generative AI for production of natural language outputs in particular.[9] The invention of transformers initiated a period of very rapid development of AI for natural language tasks, exemplified by a series of language 'models' of unprecedented capability and versatility being developed and released by researchers, technology companies, and creators of open-source AI software. We provide a brief sketch to introduce the transformer architecture later in this chapter.

A ground-breaking advance was the introduction of the 'Generative Pre-trained Transformer 3' (GPT-3) by the US technology company OpenAI and released for early testing in spring 2020.[10] GPT-3 was significantly larger than all other transformer models to date; it was trained on around 500 billion so-called tokens, which for our purposes here can be thought of as words of text. Most of this text was harvested by computers from the internet, with some filtering applied. GPT-3 represents its learning as a predictive statistical model with around 175 billion *parameters* – numerical values embedded within the model software that determine the model's operation and outputs.

This scaling up in size gave GPT-3 unprecedented capability for language processing, very significantly beyond what existing models could do at the time. GPT-3 and its successors produce written text in ordinary English – and other languages – of such a high quality and apparent meaningfulness that it is often difficult to tell if it has been written by a human or a machine. The following years saw an explosion of research and development for such so-called large language models – as well as huge commercial interest worldwide.

It is entirely natural to suppose that there may be many opportunities for the use of such generative AI in legal applications. Legal knowledge, whether black letter law or arcane legal expertise, and the articulation of law in textual form – as texts in natural language – are in some sense central to the enterprise of law. And the importance of text to legal knowledge aligns well with the strength of generative AI embodied in large language models. The emergence of highly capable generative AI also poses many legal issues and questions, including consequences for intellectual property, contracts and licenses, liability, data protection, use in specific sectors, potential harms, and of course ethics, policy, and regulation of the technology.

This research handbook discusses many of these issues in depth, and this chapter is intended to support these discussions by providing a very basic sketch of generative artificial intelligence from a technical perspective. The focus here will be on generative AI for natural language, though other forms of generative AI – for example for images or videos – also exist, are increasingly capable, and have specific legal issues.

It is, of course, impossible to provide a full account in this chapter of the highly technical, mathematical foundations of generative AI. Nor would it be helpful to delineate the state of the art at the time of writing. Technical progress in generative artificial intelligence – at least in certain directions – has been both rapid and voluminous, and any account of the frontier would quickly be out of date. We therefore focus in this chapter on a few basics, at a largely non-technical level.

Machine learning, usually underpinned by the AI computer technology of *artificial neural networks*, lies at the heart of much of the generative artificial intelligence that is the topic of this

[9] A. Vaswani et al., 'Attention is All You Need', in 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, pp. 6000–6010, December 2017.

[10] T. B. Brown et al., 'Language models are few-shot learners', *arXiv Computing Research Repository*, arXiv:2005.14165 [cs.CL], 22 July 2020.

https://doi.org/10.1017/9781009492553.004 Published online by Cambridge University Press

handbook. Artificial neural networks are introduced a bit later in this chapter. But broadly speaking, machine learning is a way of using certain computer algorithms[11] to 'learn' from data and to generalise beyond the data to perform tasks without explicit, step-by-step instructions or programming.

Typically, algorithms for machine learning are statistical in nature and the result produced is some kind of mathematical or statistical model. Such a model is an abstraction, encoded within the computer, of aspects of the system that is typified by the data, laid out using mathematical entities and structures, and concepts. The model is then typically used to make 'predictions', for example a classification, for new examples or situations not explicitly given in the data. Internal to such a model there will be some – often very many – numerical values (numbers of some kind) that are the encoded result of the learning. Creating the model from the data proceeds by a process called 'training', which iteratively adjusts these parameters to minimise the errors it makes in its predictions or to optimise some so-called objective function, which measures how good the output is.

There are three main kinds of machine learning.[12] First, there is *supervised learning*, in which a model is trained with a (typically large) number of pairs comprising an input and a known, desired output. For example, to train an image recognition model one might take a large collection of digital photographs and manually 'label' them with what they are a photograph of. The training data in this case would consist of photographic images paired with their labels. (Obviously, manual labelling in this way can be expensive, time-consuming, and possibly error prone.) As training proceeds, a predictive model of a statistical nature is developed that represents a 'function' from inputs to outputs – a kind of rule or process that assigns an output to each possible input. It is hoped that the model will generalise beyond the training examples to produce the correct, or expected, outputs for new inputs that were not present in the training data. These are the 'predictions' and using the model in this way is sometimes called 'inference'. Models represented by artificial neural networks are a very prominent example of supervised learning and often used for classification.

With *unsupervised learning*, a model is also built from training data – but in this case without the provision of an expected outcome for each input. Algorithms for unsupervised learning work without human guidance and build models that encode patterns that are detected in the data, such as similarities or differences. A typical application might be to cluster inputs into groups or to detect anomalous inputs. Some machine learning applications combine supervised and unsupervised learning – or are trained through *self-supervised learning*, in which some structure within the input data itself is leveraged to provide the desired or 'correct' outputs, rather than relying on the desired outputs being provided by human effort – such as manually labelling a set of digital photographs for training an image recognition model by supervised learning. GPT models are, in part, trained through self-supervised learning.

Finally, in *reinforcement learning*, the machine is situated within a simulated, artificial 'environment' and can monitor certain features of the state of that environment. It also has the capability to execute actions within this simulated environment that can change its state. A system of artificial 'rewards' – usually encoded numerically – is set up by the designer of the learning algorithm. As actions are taken by a simulated AI agent, rewards are generated by the

[11] An *algorithm* is a finitely described, unambiguous sequence of precise instructions to solve instances of a well-defined class of problems or to perform a calculation to produce a well-defined, specified output. Algorithms are implemented for use as computer software.

[12] A. Burkov, *The Hundred-Page Machine Learning Book* (Andriy Burkov, 2019).

https://doi.org/10.1017/9781009492553.004 Published online by Cambridge University Press

learning algorithm in response to the actions taken. The magnitude of these rewards will align with progress towards or achievement of a desired goal. The aim of reinforcement learning is to develop a 'policy' for the optimal action to take in each state – a policy that, if followed, will maximise the cumulative reward. Reinforcement learning is used to develop AI that executes a series of decisions, such as would arise in playing a board game or controlling a robot.

A structure very commonly used in today's machine learning models is the *artificial neural network*. This structure is loosely inspired by the structure and function of the interconnected neurons in a human brain – hence the name 'neural' network. These networks are behind many of the amazing capabilities in modern AI systems and lie at the heart of generative AI such as the transformer models.

There are many ways in which a neural network can be structured, but typically they consist of a (usually large) number of elements called 'artificial neurons' that have some arrangement of connections between them. Each neuron provides a numerical value as its output called its 'activation level', which is computed to be some non-linear function of the activation levels of the neurons connected to it. (Non-linear means a change in the input does not produce a strictly proportional change in the output.) The activation levels travelling on the connections are moderated by numerical parameters called 'weights', which are adjusted by a (typically supervised) machine learning algorithm that trains the model on data.

The neurons in an artificial neural network are commonly structured into a series of layers, with the connections going from one layer to the next – though many other kinds of structures are also possible. The input to the model is presented at the first layer, and the output taken from the last. This arrangement, in essence, realises a function from inputs to outputs. Such structures can have many hundreds of layers, and thousands or even millions of neurons overall – the connections of which are all individually governed by a weight parameter that is determined by the process of training. These are the 'parameters' mentioned earlier in our discussion of GPT-3. The term 'deep neural network' refers to such structures with multiple layers, and 'deep learning' refers to the special algorithms and methods that have been developed to train such networks effectively.[13]

It is beyond the scope of this chapter to go into more technical depth about machine learning or artificial neural networks. For a little more detail, in a form accessible particularly to legal scholars, chapter 2 of *The Law of Artificial Intelligence* can be recommended.[14] But it is worth just mentioning here, in the context of this handbook, the notorious issue of 'interpretability' or 'explainability' for statistical AI implemented with (deep) neural networks.[15] Deep neural networks are highly complex models, and the 'knowledge' represented within them may be distributed across millions of numerical parameters, making it very difficult if not impossible to understand or explain their behaviour – how and why they get the outputs they produce – in human terms. This issue is the subject of much ongoing research, but for the time being, deep neural networks are commonly regarded as the 'epitome of black box techniques'.[16]

Large language models (LLMs), such as GPT-4, and their applications, such as ChatGPT, are constructed using a certain assemblage of (deep) neural network components. As mentioned earlier, GPT stands for Generative Pre-trained Transformer, a type of large language model based on the transformer architectural structure introduced by Google Brain in 2017.

[13] LeCun et al., 'Deep learning'.
[14] M. Hervey and M. Levy, *The Law of Artificial Intelligence* (Sweet & Maxwell, 2021).
[15] R. Marcinkevičs and J. E. Vogt, 'Interpretable and explainable machine learning: A methods-centric overview with concrete examples', *WIREs Data Mining and Knowledge Discovery*, 13(3), e1493.
[16] ICO and the Alan Turing Institute, *Explaining Decisions Made with AI* (ICO, 2020).

https://doi.org/10.1017/9781009492553.004 Published online by Cambridge University Press

Transformers are complex and a little ad hoc – and there is a great deal of current empirical and theoretical research aiming to understand, at a deep level, why they work so well. But a key feature that contributed to their success, in practice, is that they are much better suited to massively parallel computation on modern computers for machine learning than previous generative AI for language based on forms of 'recurrent' neural networks. This is what made it possible to train them with such vast quantities of data.

The learning algorithms used to build LLMs usually belong to the 'self-supervised' class of methods. These do not require the training data to be annotated or labelled, so it's possible for the models to be trained on truly vast quantities of textual data available on the internet and in digital repositories. There are usually some additional fine-tuning steps after the initial training. Supervised learning with further data may be applied to improve the model's output for the purpose we wish to use it for, such as responding to instructions. And reinforcement learning with relatively small amounts of human feedback can be used to make the model's responses more aligned with human intentions and human values.[17]

A GPT model, once trained, generates text – a sequence of words in human language – by producing a 'good' or 'reasonable' next word to continue a block of text that the model is given as its input. (In fact, language models do not work directly with words, but with numerically encoded textual elements called 'tokens' that can be words, sub-words, or even individual letters.[18]) Language models based on the transformer architecture work by computing a *probability* for each possible next token – in essence a numerical measure of how likely each potential token is to appear immediately after the block of text that was entered as the input, based on the internal statistical model derived from the vast quantity of training data used. The token that is presented as the output is then chosen according to these probabilities, possibly moderated by a so-called temperature setting, which can be used to increase the variety and unexpectedness of the outputs by boosting the likelihood of less probable words being selected. This process is iterated, with the output token appended each time to the input text and fed back into the model, to generate sentences and paragraphs.

It's important to appreciate that the system doesn't just look up the sentence it produces in a large database of sentences. There is no database of sentences inside the model. The system predicts each next word, one by one, using a statistical model of how such an initial fragment of text might *typically* be continued. One can think of it as a staggeringly sophisticated form of autocomplete or predictive text.

The probability distribution of next tokens output by transformers is computed from a relatively large amount of previous context presented as input – not simply the last word or last few words of the initial text. This allows them to capture so-called long range dependencies in the input sequence – where the sense of one word is influenced by the appearance of a word or words much further back. Crucially, these models also have a sophisticated mechanism for weighing the relative *importance* of the words in the preceding context to the next word to be produced. This is the so-called attention mechanism.[19] This is a technically complex topic, but in simple terms the function of the attention components is to focus the model's prediction for next tokens on the parts of the input text where the most relevant information appears.

[17] T. Shen et al., 'Large language model alignment: A SURVEY', *arXiv Computing Research Repository*, arXiv:2309.15025 [cs.CL], 26 September 2023.

[18] N. Rajaraman et al., 'Toward a theory of tokenization in LLMs', *arXiv Computing Research Repository*, arXiv:2404.08335 [cs.CL], 12 April 2024.

[19] Vaswani et al., 'Attention is all you need'.

https://doi.org/10.1017/9781009492553.004 Published online by Cambridge University Press

In practical use, the input text that is provided to the model to be continued is called a 'prompt' and often consists of a series of statements and questions that help guide the AI towards producing an output of the desired form. The prompt given to a model has a big effect on what the output will be, and the whole practice of so-called prompt engineering emerged very rapidly after large language models were widely released for use. This will not be discussed here, but there is a very large amount of practical guidance on effective prompting published on the internet, and numerous technical articles reporting experimental demonstrations.[20]

For evaluating LLM applications and their effectiveness on specific tasks, a very common approach is *benchmarking*. A 'benchmark' is a fixed, published test, often consisting of a series of questions and expected answers, together with a scoring system. The evaluation of a model on a benchmark produces an overall score, typically a percentage. A wide range of benchmarks have been devised, including specialised benchmarks for legal tasks, such as LawBench,[21] which tests for legal knowledge, and LegalBench,[22] which tests for legal reasoning capability – to give just two examples.

While benchmarking LLMs is useful, there are many limitations and inadequacies in the benchmarks we have today. A multitude of benchmarks exists, and the development of many of them has been researcher-led, possibly with community contributions. They are therefore not to be considered the same sort of thing as the stringent, consensus-based, standardised testing frameworks we have in other technology sectors, such as aviation or automotives.

The question of risk is a natural one in a handbook dedicated to generative AI and law. There is a vast amount of contemporary discussion of the ethical and safety issues around generative AI, including discussion of risks to society, such as mass production of fake news, or even the risk of creating artificial superintelligence that could lead to the extinction of humanity.[23] But we will mention here only two issues that have a technical aspect and should always be borne in mind when using generative AI.

First, there is no guarantee that the text being generated by a large language model is in fact 'true'. Language models can and do produce outputs that appear to us to have the hallmark of confidence, but refer to entirely non-existent things, people, or events. This phenomenon is known as 'hallucination' among AI researchers.

Then there is also a potential risk to data security and confidentiality. Users of generative AI must always bear in mind that the inputs they send are visible to the organisation providing an LLM. Although their prompts may not be incorporated directly into the model's learned parameters, they may be stored and used for developing the LLM service in the future. Moreover, large language models can exhibit the phenomenon of 'memorisation', whereby texts used in their training can be extracted near-verbatim from the model.[24] This has obvious legal implications for both privacy and intellectual property.

Of course, researchers and the technology industry are active in addressing all these problems – and many more. *Retrieval augmented generation* aims to reduce hallucination by using document retrieval from a designated database to anchor the results of subsequent output

[20] P. Sahoo et al., 'A systematic survey of prompt engineering in large language models: Techniques and applications', *arXiv Computing Research Repository*, arXiv:2402.07927 [cs.AI], 5 February 2024.

[21] Z. Fie et al., 'LawBench: Benchmarking legal knowledge of large language models', *arXiv Computing Research Repository*, arXiv:2309.16289 [cs.CL], 28 September 2023.

[22] N. Guha et al., 'LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models', *arXiv Computing Research Repository*, arXiv:2308.11462 [cs.CL], 20 August 2023.

[23] National Artificial Intelligence Advisory Committee, The Potential Future Risks of AI (October 2023).

[24] J. Wang et al., 'Pandora's White-Box: Precise training data detection and extraction in large language models', *arXiv Computing Research Repository*, arXiv:2402.17012 [cs.CR], 24 June 2024.

https://doi.org/10.1017/9781009492553.004 Published online by Cambridge University Press

generation, to ground these results in known text.[25] There are hybrid systems that combine generative AI with other AI technologies, including symbolic AI such as decision trees, to address some of the limitations of statistical AI alone. Private or in-house systems based on *open-source models* can help with security and confidentiality. And so-called small language models are emerging that can be remarkably capable for specific applications – and are less costly and energy-hungry than a large language model.[26]

This chapter has given only a partial sketch of the technology of generative AI, focussing only on large language models. It would be impossible within the bounds of this handbook to give a comprehensive survey or detailed technical explanation of this fascinating – and very fast-moving – technology, which has raised pressing legal questions, promised potential applications, and posed regulatory challenges. Many of these are discussed in depth in the subsequent chapters of this wide-ranging and timely *Cambridge Handbook on Generative AI and the Law*.

[25] V. Magesh et al., 'Hallucination-free? Assessing the reliability of leading AI legal research tools', *arXiv Computing Research Repository*, arXiv:2405.20362 [cs.CL], 30 May 2024.

[26] T. Schick and H. Schütze, 'It's not just size that matters: Small language models are also few-shot learners', *arXiv Computing Research Repository*, arXiv:2009.07118 [cs.CL], 12 April 2021.

https://doi.org/10.1017/9781009492553.004 Published online by Cambridge University Press