Max Planck Institute
for Innovation and Competition

Max Planck Institute for Innovation and Competition Research Paper No. 19-13

Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective

Version 1.0, October 2019

Research Group on the Regulation of the Digital Economy

Josef Drexl, Reto M. Hilty, Francisco Beneke, Luc Desaunettes, Michèle Finck,
Jure Globocnik, Begoña Gonzalez Otero, Jörg Hoffmann, Leonard Hollander,
Daria Kim, Heiko Richter, Stefan Scheuerer, Peter R. Slowinski, Jannick Thonemann

Max Planck Institute for Innovation and Competition Research Paper Series

Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective

Josef Drexl, Reto M. Hilty, Francisco Beneke, Luc Desaunettes, Michèle Finck, Jure Globocnik, Begoña Gonzalez Otero, Jörg Hoffmann, Leonard Hollander, Daria Kim, Heiko Richter, Stefan Scheuerer, Peter R. Slowinski, Jannick Thonemann

The purpose of this Q&A paper is to provide an **overview** of artificial intelligence* with a special focus on machine learning* as a currently predominant subfield thereof. Machine learning-based applications have been discussed intensely in legal scholarship, including in the field of intellectual property law, while many technical aspects remain ambiguous and often cause confusion.

This text was drafted by the Research Group on the Regulation of the Digital Economy of the Max Planck Institute for Innovation and Competition in the pursuit of understanding the fundamental characteristics of artificial intelligence, and machine learning in particular, that could potentially have an impact on intellectual property law. As a background paper, it provides the technological basis for the Group's ongoing research relating thereto.

The current version summarises insights gained from background literature research, interviews with practitioners and a **workshop** held in June 2019, in which following experts in the field of artificial intelligence participated: Dr. Maximilian Alber (Charité Berlin), Dr. Cristian Ramirez Atencia (Otto von Guericke University Magdeburg), Dr. Hoda Heidari (ETH Zürich), Dr. Jelena Mitrović (University of Passau), Dr. Cigdem Turan (Technical University of Darmstadt) and Heiner Zille (Otto von Guericke University Magdeburg).

Since machine learning is a **dynamic** field, this document reflects the current understanding and might therefore be updated in the future. Readers are kindly invited to make further suggestions or to provide clarifications. The presentation of the members of the Research Group including their contact details, and the Group's other research outputs can be found at the following address: https://www.ip.mpg.de/link/regulation-of-the-digital-economy.html.

Terms with an asterisk (*) are defined in the Glossary at the end of this document.

Suggested citation:

Drexl, Hilty et al., Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective, Version 1.0, October 2019, available at: https://ssrn.com/abstract=3465577.

Contents

1. Overview of Artificial Intelligence and Machine Learning

Q1 What is generally understood by artificial intelligence? How does it differ from related fields? What subcategories does it encompass?

Artificial Intelligence

Artificial intelligence* is a branch of computer science.[1] It is often described as computer-based systems that are developed to mimic human behaviour.

Artificial intelligence is a **catch-all term** that covers machine learning*, evolutionary algorithms* (see Q10), and other technologies like rule-based systems.[2] However, due to the fact that different subfields often overlap, the **exact delineation is difficult** and a subject of controversy among researchers.

Delineation between Artificial Intelligence, Robotics, and Statistics

One of the closest neighbouring fields of artificial intelligence is **robotics**, which is a branch of engineering.[3] While in principle both fields are independent, overlaps emerge when artificial intelligence is embedded in robots with the aim to enable them to better react to a complex, uncertain and dynamic environment.[4]

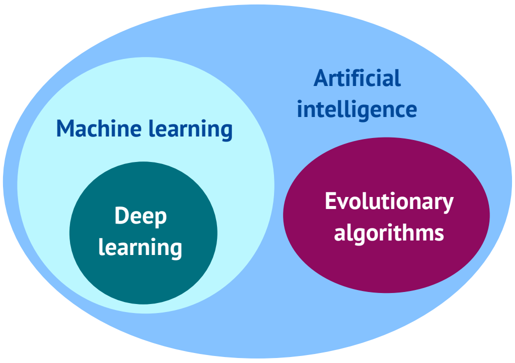
A further neighbouring field is statistics. It is disputed to what extent machine learning should be considered as an application of statistics. Indeed the purposes of both fields might not be identical: while statistics aims at analysing distributions or correlations, artificial intelligence is more about behaving intelligently or predicting (the future) correctly. Still, machine learning process relies heavily on statistical methods.

Machine Learning and Evolutionary Algorithms as Subfields of Artificial Intelligence

Machine learning is currently the most commonly used subfield of artificial intelligence. It deals with teaching a computer program to identify patterns in data and to apply the knowledge to new data.[5]

One of the most advanced subfields of machine learning is **deep learning*** (see Q4).

While this document focuses mainly on machine learning, it is important to also consider other types of artificial intelligence that are gaining importance, such as **evolutionary algorithms** (see Q10).



2. Machine Learning

General

Q2 What factors contribute to the current widespread usage of machine learning?

Machine learning* can be described as a **general-purpose technique** as it can be applied in **all sectors** to optimise decision-making and facilitate innovation.

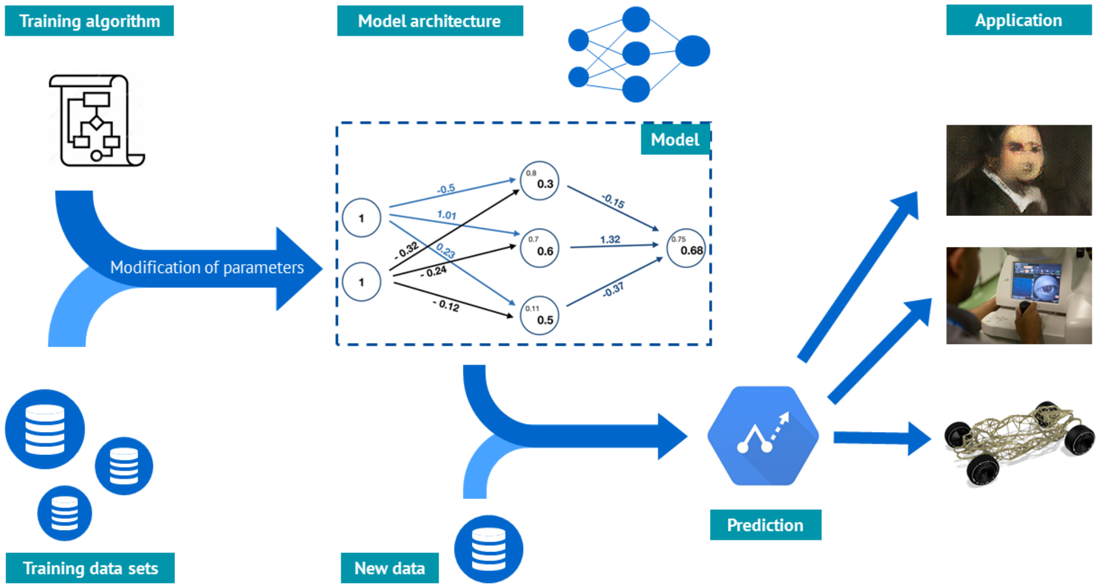
Breakthroughs in mathematical optimisation methods occurred already in the 1980s. Since then, there have been further advancements in how these methods are applied in machine learning (e.g. generative adversarial networks*, deep neural networks*).[6] The current widespread penetration of machine learning is mostly attributed to two major factors: the emergence of **large data sets** that can be used for the training process, and the increase in **computing power.**[7]

Q3 What are the basic characteristics of the machine learning process?

The entire process relies on the analysis of **data** by different **algorithms***. An algorithm is a step-by-step instruction. Generally, algorithms are encoded as **software*** to enable their readability by computers.

Machine learning* processes exist in **different variations**, depending on the data they build upon, and their task. In order to make further explanations more understandable, a basic machine learning process relying on supervised learning* (see Q5) and an artificial neural network* as model architecture* (see Q4) is presented below. Unless stated differently, this document uses this type of machine learning as its basis.

Machine learning consists of **several stages**. First, a model architecture is programmed; second, a **model*** is developed through the training process based on a **training algorithm*** and **training data sets***; third, the model is applied to **new data** to generate a certain **output** (e.g. a prediction, see Q6).



Example: A machine learning model might be used to recognise cats in pictures. In the case of supervised machine learning, the model is trained on a data set containing labelled data (i.e., each picture is accompanied by the information whether there is a cat in the picture), allowing it to become more accurate. Once the training is completed, the model should in principle be capable of recognising from an unlabelled picture whether a cat appears in it (output). This model could finally be implemented in a self-driving car, allowing it, for instance, to brake when confronted with a cat (application).

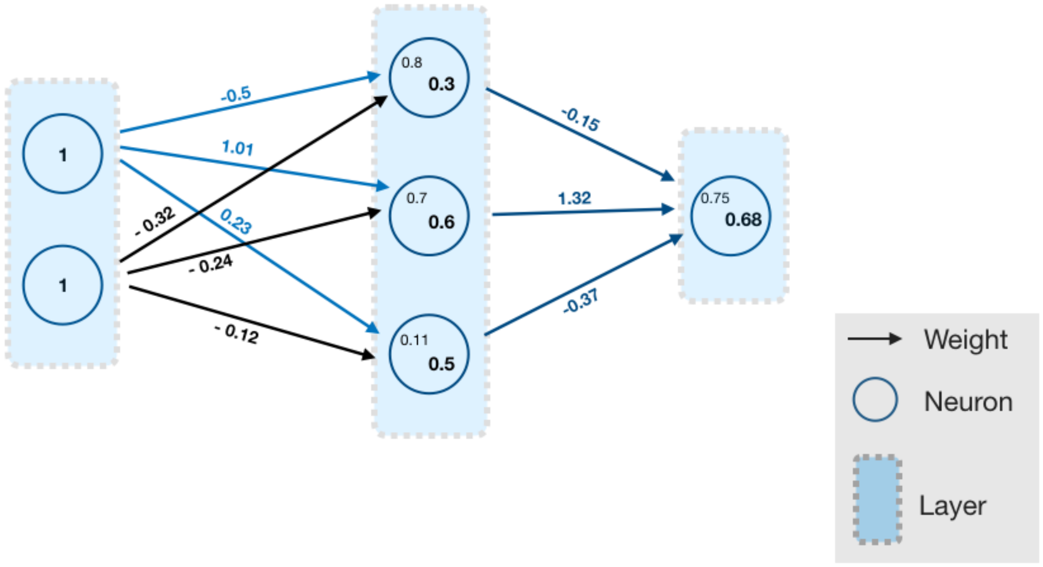
Training Process and Model

Q4 What is a machine learning model and what are its components?

Model

The trained machine learning* model* is the immediate output of the training process. It is an algorithm* based upon a **(nonlinear) mathematical function** that generates output based on the learned patterns in the training data*.

One type of models are **artificial neural networks***, the structure of which imitates the functioning of a human brain. These models rely on an **architecture*** which is usually established by a programmer prior to the training process and is composed of layers of **neurons*** connected by **weights***. Each neuron is a mathematical function which transforms inputs (the numeric value of the upstream weights) into an output (the numeric value of the downstream weights). The model is composed of the sum of all the functions entailed in the neurons.



When the number of layers is high, the neural network* is described as **deep (deep neural network)**, whereby the scientific community does not agree upon a clear delineation.

Trainable Parameters and Hyperparameters

Two types of parameters might be distinguished in the machine learning context, namely **hyperparameters*** and **trainable parameters***. Trainable parameters evolve during the training process, whereas hyperparameters are fixed before the training process and do not evolve.

The **weights*** are **trainable parameters**. They are numeric values that are first randomly allocated and then optimised during the training process.

Alternatively, a model trained for a different, yet similar task can be used as a starting point instead of randomly allocating the weights (**transfer learning***). This allows a quicker optimisation compared to starting from scratch, but a training process remains unavoidable.

The **architecture***, on the contrary, is a **hyperparameter**, which means that it (currently) needs to be developed by a human before the training process and does not evolve during it. The determination of the optimal structure relies on heuristic* methods and human know-how. Research is nonetheless being conducted to allow the modification of the architecture through the training process.

Online and Offline Models

A model can be online or offline. In the case of the currently predominant **offline models*** (also called static models), the optimisation process and the actual application are separated. In the application phase, no modification of the weights takes place.

If the model is designed as an **online model*** (also called dynamic model), its optimisation never ends, because even when being applied, its output is used to continually modify the weights. This requires feedback on the correctness of the output.

Q5 How is a model trained? What kind of human input does this necessitate? Which are the different training methods?

Training Process

In the training process, **training data*** is 'fed' into the model*. Based on this, the **training algorithm*** (sometimes also called optimisation algorithm) optimises **trainable parameters***.

A training algorithm includes a **loss function*** that reflects the accuracy of the model. A loss function is a mathematical function which evaluates the magnitude of error of a model. The smaller the loss function, the better the model is. The training algorithm will, therefore, seek to minimise the loss function by finding the best combination of trainable parameters.

The modification of the **weights*** is hence a mathematical operation. Therefore, under identical parameters, an identical model would be developed. Slight changes of the parameters (like for instance regarding the order of feeding the training data, or in the initial allocation of the weights, see Q4) would lead to the creation of a model that is slightly different but likely capable of producing outputs of similar accuracy.

Development, Availability, and Choice of Training Algorithms

Training algorithms* vary in originality from standard to uniquely developed. Many established training algorithms are standard, and available online in **open-source** libraries in the form of pre-written software*. However, for some problems, new algorithms* have to be developed, necessitating human and financial investments.

The choice of particular training methods requires know-how and currently relies on certain **heuristic methods***. Heuristics is an approach to problem-solving relying on experience and intuition rather than a pure scientific methodology. Heuristic methods are often used due to the lack of sufficient computing power or the absence of exact methods for the solving of certain problems. If computational power was limitless, it would be possible to test every possible solution to find out which is the best one. Research is being conducted to replace heuristic methods by exact ones. However, even though available computing power is continuously increasing, experts doubt that in near future all heuristic methods could be replaced by exact ones.

Supervised, Unsupervised, and Reinforcement Learning

Three training methods can be differentiated regarding the type and the way training data is used.

In the case of **supervised learning***, the model is 'told' during the optimisation process what the training data it is confronted with represents. Hence, it knows whether the prediction it made in the optimisation process was right or wrong. For doing that, it necessitates labelled data, for which additional investments (for instance human involvement) are required. Supervised learning is currently the most common form of machine learning*.

Example: Cat recognition (see above, Q3).

Unsupervised learning*, on the contrary, relies on unlabelled data. Here, a model is trained to identify similarities, parallels and/or differences in data, the main use case being clustering. Since labelling of the training data is not required, this type of training requires less human participation. On the other hand, a more extensive human interpretation of the output is required.

Example: An algorithm analyses data about customers to build groups based on their potential purchasing power to customise the offers.*

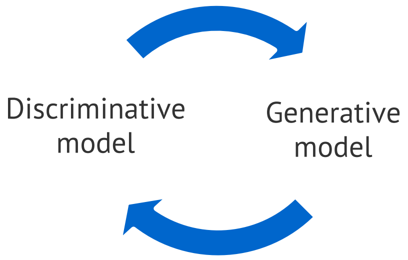
Reinforcement learning* does not rely on pre-existing data sets, but rather gathers data from simulations or games. The algorithm determines the rules based on continuous feedback on the actions it takes during the training.

Example: This methodology was used to train an algorithm to play Go.[8] *The computer played the game against itself multiple times and gained feedback based only on the final score of each game. Even though the algorithm was never taught about Go strategies, it was, based on this feedback, able to improve its skills and outperform human players.*

Generative Adversarial Networks

A special type of machine learning systems is **generative adversarial networks***, which refer to the interplay between two models: a discriminative and a generative one.

The **discriminative model** is trained to detect whether a piece of data is part of a real data set or was generated by an algorithm*, whereas the **generative model** is trained to create outputs imitating those of a real data set. Thereby, the outputs of one model serve as the inputs for the training of the other one. The reaction of the latter to these inputs serves as a feedback for the former. The loss function* of the generative model hence depends on its capacity to mislead the discriminative model, whereas the loss function of the discriminative model depends on how well it can identify whether the inputs were created by the generative model.



Importance of Data

In general, **training data is the most valuable element** of the machine learning process. The better the training data in terms of quantity, quality and variety, the more accurate the calculation of the trainable parameters and hence the more precise the output.

Training data is **not stored** within the model during the training process, and once the training process is completed, the model is fully usable independently of the data. Therefore, the developer of a model can commercialise this model without the need to disclose the data used for training.

Machine Learning Outputs

Q6 What are the possible outputs of a machine learning model, and to what extent are they explainable and interpretable?

Need for Differentiation between 'Direct Output' and 'Application'

A distinction should be made between the direct output of a model* and the potential practical applications of this output. Such distinction might appear formal but is nevertheless necessary: in particular, from an intellectual property law perspective, **distinct legal issues** can arise with regard to direct outputs and their applications.

Direct Output

The final output of a machine learning* model depends on what the model is trained for. It can be e.g. a **correlation between data points, clustering** (e.g. defining groups of customers based on their characteristics), or a **prediction** (e.g. the probability that a cat is in the picture).

The **accuracy** of the output depends on the quality of the model and, thus, on model architecture*, training algorithm* and training data* (see Q3, Q4).

If a trained machine learning model is confronted with the same (new) data, it will produce the same output. Even though the output is in principle **deterministic** and **traceable**, it is often not human-explainable due to the complexity of the calculations, especially in the case of artificial neural networks (the 'black box' issue).

Application

The output can then be put into practical use. Examples of applications might fall within the **areas of art** (e.g. paintings, text conception and translation, music creation) and **technology** (for instance the functioning of a self-driving car, optimisation of a car design, development of medical treatments, virtual assistants). For instance, the goal of an application in technical fields can be to develop a new product (such as an antenna), or to guarantee the correct functioning of a device (such as a self-driving car). Further, machine learning outputs can be used to improve **business processes** (such as targeted marketing).

Technically, an application may take different forms: the output of the model might need to be decoded or not, and be used as such or combined with other technical or knowledge-based elements.

Any application requires **human input**: a machine learning model will not independently start using its output to do something. However, the required degree of human contribution can vary considerably from one application to another.

Reverse Engineering of Machine Learning Components

Q7 Is it possible to reverse engineer different components of the machine learning process?

Reverse Engineering

In the machine learning* context, **reverse engineering** can be defined as the possibility to extract or deduce certain elements of the machine learning process through access to other elements (e.g. to deduce the model* itself by analysing predictions made by it, or to deduce the training data* from the model).

The possibility to reverse engineer machine learning elements is controversial. Research is currently conducted in this field. Its purpose appears to be primarily to explain the machine learning output, i.e. to understand the factors leading a given model to a concrete output.

Currently, complete reverse-engineerability of all the parts of the machine learning system does not seem realistic. Some elements could potentially be reverse-engineered,[9] but since there are often too many 'moving parts' (see Q4, Q5), this seems very challenging. The possibility to reverse engineer varies from case to case and depends on the complexity of a model and on how much knowledge one already has about the training process. For example, knowing that a linear model was used, and having access to the model function (but not the parameters) makes it easier to reverse engineer the model.

In some cases, reverse engineering might not be viable from an economic point of view, as it might be easier to design an alternative model that produces a comparable outcome.

Teacher-Student Learning

The factual exclusivity of a model could potentially be circumvented by developing a so-called **student model***. A student model is trained to reproduce the outputs of the teacher model when confronted with the same input. The model will therefore not be the same, but could produce similar outcomes.

Machine Learning 'Intelligence'

Q8 Which parts of the machine learning process are pre-determined by humans?

The human input in machine learning* mainly subsists in choosing or developing a training algorithm* (can require creativity to develop a new algorithm), setting the hyperparameters* (often involves trial and error; research is conducted to use machine learning to define hyperparameters), data labelling (mundane work), and developing the model architecture* (often a heuristic* process, see Q4, Q5).

Even if models* appear 'intelligent', they generate output by merely relying on probability calculations (see Q5). They are not autonomous (i.e., they do not 'reason' on their own) and need to be fine-tuned by machine learning experts.

Q9 To what extent is a machine learning model a 'black box' that cannot be explained by a human?

Artificial intelligence in general and machine learning* in particular are often described as a **'black box'**, i.e. even though experts can in general explain how a model* functions, they cannot explain precisely why it generated a concrete output based on a given input. Research is being conducted in this area (the 'Explainable AI' movement).

In general, the explainability of machine learning varies depending on the complexity of a model and the training techniques used (see Q4, Q5). The problem usually arises with regard to deep neural networks*, as unlike computers, humans are not capable of processing such large amounts of data.

Explainability and interpretability of machine learning is currently one of the major research areas.

3. Evolutionary Algorithms

Q10 What are evolutionary algorithms? Do they differ from machine learning?

An **evolutionary algorithm*** is an optimisation method that attempts to identify the best solution for a given problem out of multiple, autonomously generated alternatives. Evolutionary algorithms rely on **Darwinian principles**, as natural evolution has proven to be a powerful optimisation process.

In order to find a solution, an evolutionary algorithm does not need training data*, as in the case of artificial neural networks* (see Q5). Instead, an initial **population** of possible solutions with different characteristics is first **generated randomly**. Second, the evolutionary algorithm evaluates the quality and/or fitness of each solution in that population and **selects the best-suited ones**. Then the selected solutions are modified using mechanisms such as **reproduction**, **mutation**, and **recombination**. This process generates a new population which is again evaluated. The process continues until an optimal solution is found.

Evolutionary algorithms can be used in machine learning*, e.g. to find the best model*. However, their scope of application goes beyond the creation of models, since they can be used for other tasks (e.g. an evolutionary algorithm was used by NASA for the development of an antenna).[10]

Glossary

Algorithm: a step-by-step instruction. In the machine learning context, an algorithm is an instruction coded as software and directed at a computer. (Q3)

Architecture: a structure of a model usually established by a programmer prior to the training process. A type of architecture are artificial neural networks. (Q4)

Artificial intelligence (AI): a catch-all term that describes a branch of computer science dealing with the development of systems that behave in a way similar to human intelligence. (Q1)

Artificial neural network (ANN): a type of a model, the structure of which consists of layers of neurons connected by weights. (Q4)

Deep learning (DL): a type of machine learning based on the use of a more complex architecture (composed of a higher number of layers), often referred to as deep neural networks. (Q4)

Deep neural network (DNN): see Deep learning. (Q4)

Evolutionary algorithms: an optimisation method based on the principles of evolution trying to identify the best solution for a given problem out of a given population. (Q10)

Generative adversarial networks (GAN): a special type of machine learning systems relying on an interplay between the so-called discriminative and generative models. (Q5)

Heuristic (methods): an approach to problem-solving relying on experiences and intuition rather than on a pure scientific methodology. (Q5)

Hyperparameter: a feature of a model that is fixed before the training process and does not evolve. (Q4, Q8)

Loss function: a mathematical function which evaluates the magnitude of error of a specific model. (Q5)

Machine learning (ML): an automated process of identifying patterns in available data and then applying the knowledge to new data. Machine learning is currently the most commonly used subfield of artificial intelligence. (Q1)

Model: an algorithm based upon a (nonlinear) mathematical function that generates output based on the patterns learned from the training data in the training process. (Q4)

Neuron: a mathematical function as a part of a model which transforms inputs (the numeric values of the upstream weights) into an output (the numeric value of the downstream weight). (Q4)

Offline model: a model in which optimisation process and application are separated: in the application phase, no modification of the weights takes place. (Q4)

Online model: a model in which the optimisation never ends, because even when applied in 'the real world', its output is used to continually modify the weights. (Q4)

Reinforcement learning: a learning process not relying on pre-existing data but rather gathering data from simulations or games, where the algorithm 'figures out' the rules based on continuous feedback on the actions it takes during the training. (Q5)

Software: digital encoding of instructions that tell a computer how to work. (Q3)

Supervised learning: the currently most common type of a learning process in which training data is labelled to tell the model whether the prediction made during the optimisation process was right or wrong. (Q5)

Student model: a model trained to reproduce the outputs of a teacher model when confronted with the same input. (Q7)

Trainable parameter: an element of a machine learning model that evolves during the training process. (Q4)

Training algorithm: an algorithm used to train the model. It seeks to minimise its loss function by finding the best combination of the model parameters. (Q5)

Training data set (training data): a set of data used for the training of an algorithm in the optimisation process. (Q5)

Transfer learning: the use of a model trained for a different, yet similar task as a starting point instead of the randomly allocating the weights. (Q4)

Unsupervised learning: the type of learning in which the model is confronted with unlabelled data and ultimately trained to identify similarities, parallels and/or differences in data. (Q5)

Weight: a trainable parameter connecting neurons in a given architecture. It is a numeric value that is first randomly allocated and then optimised during the training process. (Q4)

Endnotes

[1] For an in-depth presentation of artificial intelligence, see European Commission – Independent High-level Expert Group on Artificial Intelligence, A definition of AI: main capabilities and disciplines, https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines.

[2] In rule-based systems, the reactions are pre-programmed and not adaptive (e.g. rule-based chatbots).

[3] Robotics is defined as the science of designing and operating robots, i.e. machines that are controlled by a computer that are used to perform jobs automatically (Cambridge dictionary).

[4] See e.g. WIPO, Technology Trends 2019: Artificial Intelligence, https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf, p. 26.

[5] European Commission, Artificial Intelligence for Europe, COM(2018)237 final, p. 10.

[6] Another exemplary breakthrough is the development of word embeddings; see Mikolov/Sutskever/Chen/Corrado/Dean, Distributed Representations of Words and Phrases and their Compositionality, https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

[7] Economic literature argues that breakthroughs in relation to machine learning techniques and hence the technological side are the reason for widespread application and technology diffusion. See Furman/Seamans, AI and the Economy, https://ssrn.com/abstract=3186591, p. 4: 'performance increases are due to breakthroughs in various machine learning techniques' or 'these scientific breakthroughs are starting to find their way to commercial applications'.

[8] Silver/Huang/Maddison et al., Mastering the game of Go with deep neural networks and tree search, https://www.nature.com/articles/nature16961.pdf.

[9] See, for instance, Oh/Augustin/Schiele/Fritz, Towards Reverse-Engineering Black-Box Neural Networks, https://arxiv.org/pdf/1711.01768.pdf; Tramèr/Zhang/Juels/Reiter/Ristenpart, Stealing Machine Learning Models via Prediction APIs, https://arxiv.org/pdf/1609.02943.pdf; Veale/Binns/Edwards, Algorithms that Remember: Model Inversion Attacks and Data Protection Law, https://ssrn.com/abstract=3212755.

[10] Hornby/Globus/Linden/Lohn, Automated Antenna Design with Evolutionary Algorithms, https://ti.arc.nasa.gov/m/pub-archive/1244h/1244%20(Hornby).pdf.

Image sources (p. 5)

Christies, Edmond de Belamy, https://www.christies.com/img/LotImages/2018/NYR/2018_NYR_16388_0363_000(edmond_de_belamy_from_la_famille_de_belamy).jpg.

MIT Technology Review, DeepMind has made a prototype product that can diagnose eye diseases, https://www.technologyreview.com/f/613249/deepmind-has-made-a-prototype-product-that-can-diagnose-eye-diseases/.

Pette, Where VR Meets the Road: How GPUs Power 'Hack Rod', World's First AI-Generated Car, https://blogs.nvidia.com/blog/2016/07/26/hack-rod-car-ai/.