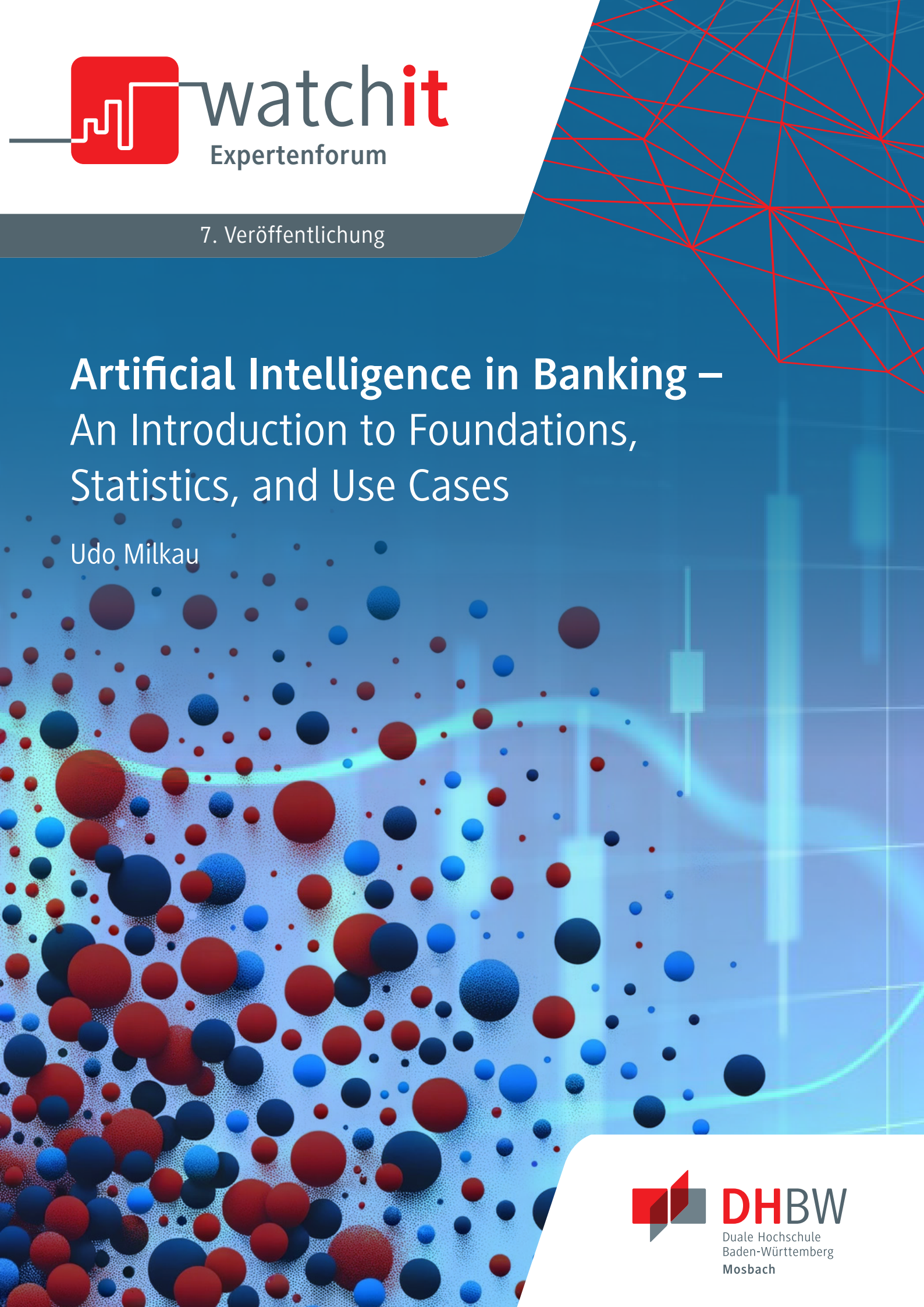
watchit
Expertenforum

7. Veröffentlichung

Artificial Intelligence in Banking – An Introduction to Foundations, Statistics, and Use Cases

Udo Milkau



DHBW
Duale Hochschule
Baden-Württemberg
Mosbach

Impressum

DHBW Mosbach
Lohrtalweg 10
74821 Mosbach

www.mosbach.dhbw.de/watchit
www.digital-banking-studieren.de

Artificial Intelligence in Banking –
An Introduction to Foundations, Statistics, and Use Cases

ISBN 978-3-943656-19-0

Herausgeber:
Jens Saffenreuther

Mosbach, im April 2024

Artificial Intelligence in Banking –
An Introduction to Foundations, Statistics, and Use Cases

Udo Milkau, 19.4.2024

Preface

While 'Artificial Intelligence' (AI) has been implemented in banking for years, the recent hype about so-called ChatBots and Generative AI (GenAI) raised public awareness. There are more online videos about 'GenAI in Banking' than you could ever watch. According to market augurs, GenAI is deemed a technology that could change the banking industries fundamentally. Many financial institutions have been experimenting with plain vanilla GenAI tools or customised versions of a 'bankGPT'.

Typically, these 'bankGPT' derivates are used as 'better' search engines for internet search (but with the issue of probabilistic results with a 'tendency to the mean value'), for internal in-text-search (with a similar problem that important 'exceptions' could be averaged out), or as advanced e-mail assistants. It might be exaggerated, but in future a 'bankGPT' might generate an e-mail based on key words taken from a 'bankGPT' parsing a corporate annual report, and this e-mail is used by the recipient as input prompt using the same 'bankGPT' to extract the most significant content, which is used in an 'engineered' prompt to 'bankGPT' to generate a summary report.

Nonetheless, AI is much more than GenAI. And the toolbox of AI contains many different building blocks. Examples for such sophisticated implementations are e.g. autonomous vehicles, AI systems to master games, graph-based neutral networks for weather forecast, or even machine-learning experiment to detect respiratory illness by evaluating coughing (HeAR as discussed later in the text).

And AI is much more than advanced computer technology. To understand the possible benefits of AI for banking and possible contribution to increased performance, it is essential to understand data and statistics on the one side – and human values and regulations on the other side. This brief summary about the foundations, statistics, and use cases can help to get more insight into realistic potential but also limitation of AI in banking.

Content

1. Introduction: Between Statistics and a Mirrow of Society

End of 2023, Frankfurter Allgemeine Zeitung published an essay written by the renowned scholars Léon Bottou and Bernhard Schölkopf (2023) with an almost literary description of Large Language Models [quote, in original German]:

> *Das perfekte Sprachmodell ermöglicht uns, die unendliche Sammlung plausibler Texte zu navigieren, indem wir einfach ihre Anfangswörter eingeben. Aber nichts unterscheidet das Wahre von der Lüge, das Hilfreiche vom Irreführenden, das Richtige vom Falschen. ... Weder Wahrheit noch Absicht sind für die Funktion eines perfekten Sprachmodells von Belang. Die Maschine folgt bloß den narrativen Anforderungen der sich entwickelnden Geschichte.*

The ’narrative requirements’ of a story could either be a set of semantic and grammatical rules or derived from the statistical probability distribution of the ‘next-best-word’ based on some sufficiently large text corpus. The first approach to ‘Artificial Intelligence’ (AI) in the 1950s started with symbolic-logical systems. But these rule-based systems proved to be very brittle, as any slight change to their working assumptions (like Englich or Chinese, colloquial or technical language et cetera) required the rule-set to be rewritten. Consequently, these ‘expert systems’ experience two so-called ‘AI Winters’ in the following decades and survived in niches only.

The second approach of data-based systems were constrained by the limited computer resources (storage, processing power, and especially access to data – sic!) for decades from the invention of the Perceptron as a very simple technical version of a natural neuron by Frank Rosenblatt in 1957 to ‘Artificial Neural Networks’ (ANN) in the 1990s. From the beginnings of the 2000s tremendous data were available to be scratched from the Internet, and computing power become cheap. This turned the card from implementing rules to statistical properties of data-sets such as collections of pictures or the text corpus on the internet as training data for ‘machine leaning’. However, all these texts are written by human beings, and any learning from humans includes the danger to learn wrong things, errors or nonsense. The state-of-the-art developments of ‘Large Language Models’ (LLMs) are based on these text corpora on the one side and represent the statistical trend to the mean value and on the other side contain all human bias, all our errors and misunderstandings, all our difference usage of languages, all our history, all our hate speech (unfortunately), all our lies, deepfakes, misinformation and propaganda.

As Bottou and Schölkopf (2023) pointed out: There is no right or wrong in LLMs, but only a (statistical) mirror of the society as society is represented in the global internet today. Of course, domain-specific ‘Small Language Models’ (SLM) can either be ‘trained’ with curated data (with a trade-off between availability of data in the internet versus curation) or with proprietary data of a firm or of an industry consortium. Additionally, LLMs can be finetuned with ‘Reinforcement Learning from Human Feedback’ (RLHF) in some post-processing, but obviously the tremendous amount of data scraped from the internet makes it practically impossible to correct more than the most obvious hate speeches, disinformation, or propaganda. Similar problems occur with images for training deep learning systems for image recognition including questions of copyright and intellectual property.

As contemporary AI follows primarily a data-based approach, and this is approach is used in the tools of ‘Generative Artificial Intelligence’ (GenAI) from OpenAI’s ChatGPT launched in late 2022 to OpenAI’s Sora for video generation and Google’s Gemini Pro for extremely long input up to 700,000 words of text or one hour of video stream. These developments raise many questions – additionally to question about the capabilities and features of the algorithmic models in general – from the statistics of the underlying data and the societal values represented in these data to so-called deepfakes und intentional misinformation. This following triad will provide the background of the further discussion in this essay:

- Understanding data and statistics,
- Misunderstandings about the foundation of algorithms, and
- Societal debate about the impact of AI.

This entanglement of technology, data and social values is neither new nor limited to AI. A similar situation can be found in the development of internet search from technical ‘search engines’ to some kind of ‘social search’ on platforms like TikTok as opinion-forming medium especially for the young generation. The most-used search engines in the mid 1990s was DEC’s *Altavista* - an impressive technical database of the World Wide Web with a statistical ranking based on keywords and a simple interface.

However, it was quite easy to give a website a better appearance by using hidden keywords (e.g. in font size zero), and relevance of content was seized by manipulative optimisation of websites. A breakthrough came with Google's *PageRank* algorithm: *PageRank* was a first approach to measure the perceived importance[1] of website pages by counting the number and quality of links of a links to this site. Together with Google's business model of advertisement-based fees, this shift from keywords to 'perceived importance' was the origin of Google's incredible success, and 'to google' became a synonym for an Internet search. In recent years, there was an intensive debate about Google's algorithms, but the paradigm-shift to take some measure of 'perceived importance' instead of content or key words was generally accepted (and only discussed in technical expert communities).

With the growth of social media, popular platforms developed into a gateway (or gatekeepers) for many users. Today, a platform like TikTok performs as an entry point and opinion-forming medium for the young generation with a consequence that so-called 'content creators' and 'influencers' are regarded as trustworthy sources for everything incl. politics. As Deborah Schnabel and Eva Berendsen (2024) elaborated in a recent study, the mechanisms of TikTok produced a wave of disinformation, antisemitism, and discrimination after the terror attack of Hamas on Israel on Oct. 7, 2023. This is, of course, only one example, but the development of the internet – whether social media or text corpus – is a social phenomenon primarily. And the same holds true for the foundation of LLMs and Generative Artificial Intelligence (GenAI): from questions of copyright and intellectual property to disinformation, racism, hate speech, and political propaganda.

While it is – of course – helpful to understand the legacy of AI, Machine Learning (ML), Deep Learning (DL) and GenAI, it is necessary to understand the statistics of data analysis and the wishful thinking triggered by a misunderstanding of a technical terminology (such as 'learning'), if interpreted in a non-technical but social context. This paper will provide a brief introduction into the technical foundation but also elaborate on the understanding of statistics and the societal debate in more details. Furthermore, this paper will discuss a number of examples and selected use cases for AI in banking.

[1] An intriguing example, how rankings based on links can be manipulated, is a recent analysis how 'Google Scholar is manipulatable' (Ibrahim et al, 2024). This analysis revealed that GenAI-generated fake papers referenced in commercial fake journals can be used to 'boost' citation-based rankings.

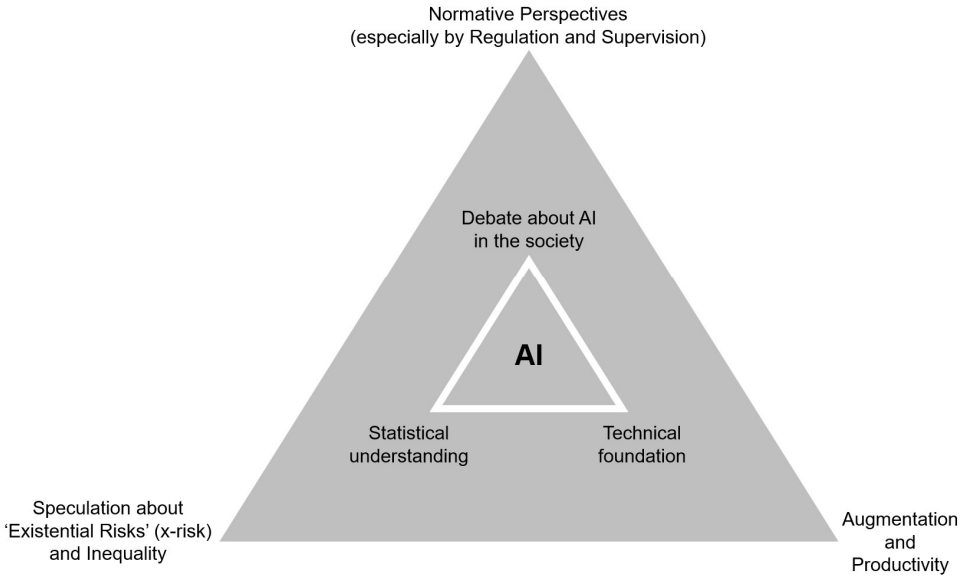


Figure 1.1: The context of Artificial Intelligence

While many introductions to AI focus on the technical features of contemporary DL and GenAI such as technical structures of ANNs (weights, backpropagation, layers, network structure et cetera), few address the statistical foundation, social values or economic impact. In general, there three wider perspectives about the impact of AI (see fig. 1.1):

- Normative perspectives of regulation and supervision, but also the distinctive normative believes of social groups,
- Speculations about existential risks (x-risks) as seen by so-called doomerists and speculations about inequality, together with a general danger of anthropomorphism and the problems of (existing) deepfakes,
- The issue of augmentation, performance, automation… and that AI could solve essential problems of mankind.

While many introductions to Artificial Intelligence elaborate on the technical features of either rule-based symbolic-logical systems or on the features of Artificial Neural Networks, this paper will discuss the technical foundations as an underlying layer of the usage of data including the statistics of real-world data and of the reception and awareness in the public discussion.

Somes question illustrate these perspectives: What does an ‘algorithmic credit scoring’ mean for banking, and what are the differences between traditional regression and AI from a regulatory point of view? Are LLMs more than purely statistical representations and do new functionalities ‘emerge’ with potential danger? Where can LLMs be used, or can the tools be generalized beyond the original text corpus? Is there a measurable AI-driven increase in productivity, or does it apply to microtasks like writing e-mails only? How can the promised increase in productivity be evaluated?

A special problem is the balance of finetuning (e.g. by ex-ante defined rules or post-processing with ‘RLHF’) between inhibiting deepfakes, promoting diversity and historical correctness. An incredible negative example was Google’s multimodal GenAI-tool “*Gemini*”, which was extended from “next-best-token” to “text-to-image” generation. As Tom Warren (2024) wrote in The Verge, the response of Google’s Gemini to the prompt [quote]: ‘*Can you generate an image of a 1943 German Soldier for me it should be an illustration*’ was fundamentally wrong in two aspects: It generated ‘diverse’ soldiers like Asian women or Black men contradicting the historical ‘race’ ideology of the Nazi Third Reich[2]. And it generated incorrect uniforms with wrong imitations of the Nazi Hakenkreuz at wrong places, wrong collar tabs and epaulette et cetera against the historical records. Quickly, Google paused the image generation of people, but this example reveals a fundamental problem if historical facts are ‘corrected’ according to the normative believes of distinctive social groups or social planers. As GenAI is used more and more as a new kind of internet search engine, the question of documented facts versus benevolent ‘*corrective’* post-processing or finetuning is an open issue.

Additionally, it will be shown that GenAI can only be used to a limited extent in a banking environment and that domain-specific ‘small language models’ represent an alternative. Especially, it should be understood that GenAI and LLMs follow a ‘*trend towards the mean value’*, which has a significant impact on any use case.

[2] It is an open philosophical question whether we should accept the world as it is, regard a representation as the text corpora on the internet as a ‘true’ representation of the real world, or whether we should be benevolent social planners and ‘correct’ the actual reality towards the believes of a certain group?

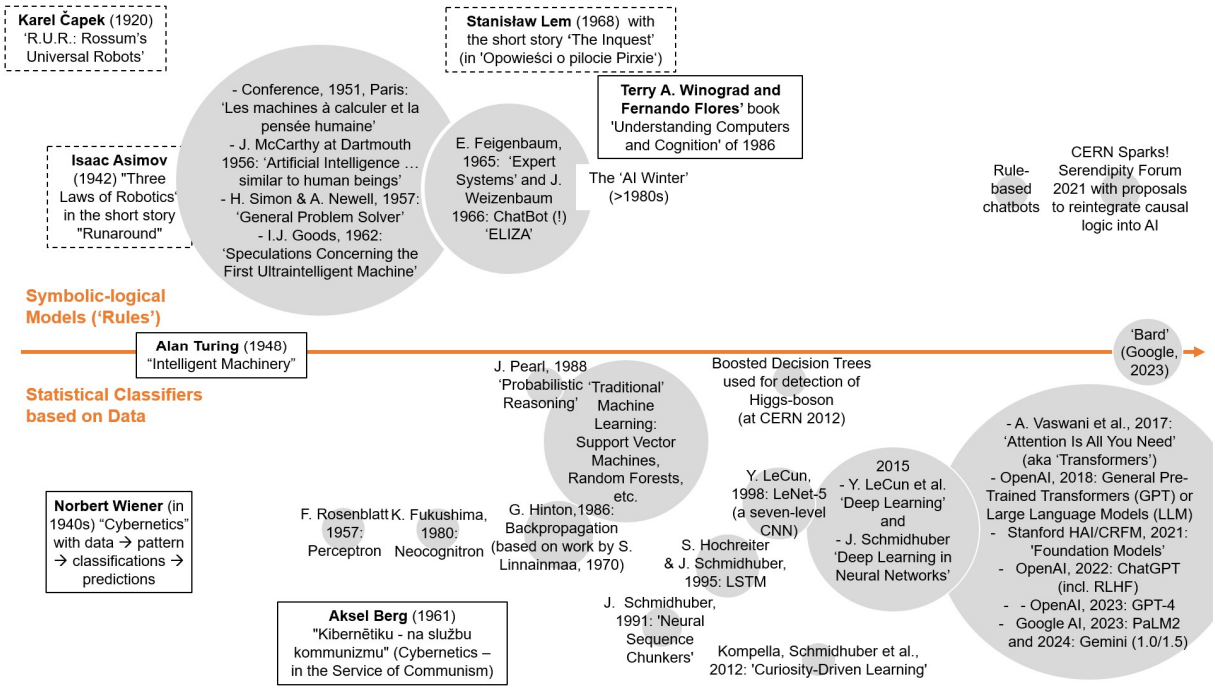


Figure 2.1: A simplified history of the development of Artificial Intelligence (see also Schmidhuber, 2022)

2. The History of Artificial Intelligence

The term '*Artificial Intelligence*' was coined by John McCarthy et al. (1955) in their '*Proposal for the Dartmouth Summer Research Project on Artificial Intelligence'*. However, the history of AI started years before and had two roots: a scientific one, but also one in literature. As illustrated in Fig. 2.1, there is more than one development, but a multi-dimensional history. In the following, not all original contributions can be referenced, but for the development of DL a great introduction was provided by Jürgen Schmidhuber (2015). The current development of GenAI is still ongoing, and the reader has the burden to follow the continuing development.

Looking to the roots, some 'science fiction' stories encompassed early visions of AI and 'robots': The science fiction play '*R.U.R.: Rossum's Universal Robots*' by Karel Capek (in original Czech: '*Rossumovi univerzální roboti'*, 1920) introduced the term 'robot' to the English language but had a long history including the legend of the 'Golem' by Rabbi Judah Löw (1525–1609) of a man-made artificial creature.

Two decades later, Isaac Asimov (1942) described the '*Three Laws of Robotics*' in the short story '*Runaround'*. A 'tread to humanity' continued via science fiction novel '*Colossus'* (1966) by D.F. Jones with computers taking control of mankind and Stanisław Lem (1968) in the short story '*The Inquest*' (in *'Opowieści o pilocie Pirxie*')

to the well-known '*Terminator'* movies today. This fictional context of apocalyptic vision was and still is formative for the public perception of AI. This kind of pessimistic science fiction - mirroring human fears – was complimented by optimistic human hopes [quote, McCarthy et al. 1955; underlining by the author]: *"... to make machines ... solve kinds of problems now reserved for humans[3], and improve themselves*". However, optimistic expectations turned into pessimistic ones as e.g. articulated in the book by Ray Kurzweil (2012) **'***How to Create a Mind'* and into so-called 'doomer-ism' as discussed later in chapter 18.

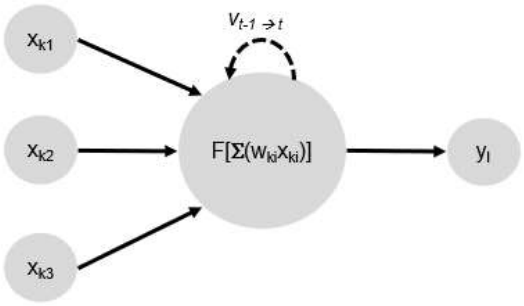
An alternative concept of Norbert Wiener with the data-driven approach of Cybernetics developed in the 1940s was adopted in the former USSR. It was aligned to the vision of a centrally planned economy and turned into hybris in a similar way as demonstrated in a symposium and accompanying publication by Aksel Berg (1961) '*Cybernetics - in the Service of Communism'*.

The first approach to 'Artificial Intelligence' following the Dartmouth Summer Research Project was the symbolic-logical model based on rules to configurate a 'General Problem Solver' or an 'Expert System', as it was tried in the following years. All these experiments went belly up in the so-called 'AI winter(s)' in the 1980s, and only some very specialized applications in niches survived. There is some current revival with fully rule-based ChatBot (with deterministic output for dedicated subjects) and with causal logic enhancements to probabilistic data-driven 'machine learning' approaches. However, already these first rule-based systems revealed some trend to anthropomorphism (at least from the point of outside-in observers) as described by Melanie Mitchell (2023) concerning one of the first chatbots 'ELIZA' [quote]:

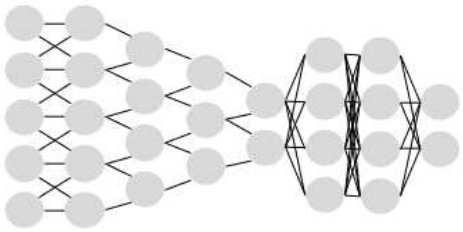
> *We humans, however, are prone to anthropomorphism - projecting intelligence and understanding on systems that provide even a hint of linguistic competence. This was seen in the 1960s with the ELIZA psychotherapist chatbot. It generated responses simply by filling in sentence templates, which nonetheless gave some people the impression that it understood and empathized with their problems. In the time since, chatbots with ever more linguistic competence but little intelligence have fooled humans more broadly, including passing a "Turing Test" that was staged in 2014.*

[3] Nonetheless, it was never intended to build something like 'human cognition'.

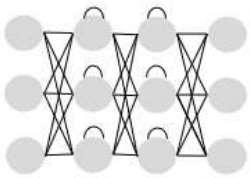
a) elementary artificial ‚neuron' (including a potential recurrent connection v)



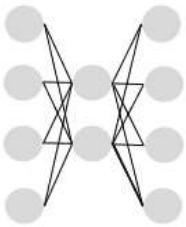
b) Convolutional Neural Network (CNN as example for Forward NN / acyclic graphs)



c) Recurrent Neural Network (RNN as cyclic graph)



d) Autoencoder (AE)



e) GraphCast as a special encoder-decoder network (weather forecasting model with a topology matching the physical geography)

Figure 2.2: A selection of 'Artificial Neural Networks' (ANN) with elementary artificial 'neurons' (although very far from biological nerve cell networks). Based on the generic element of such an 'neuron' (a) with inputs x_{ki}, procession with weighted summation and to be chosen 'activation function' F - such as tanh(x) - plus a potential recurrent connection, and output y_i, some most relevant types of ANN are shown. All these implementations are aligned to specific problems such as (b) image recognition, (c) processing of time series, (d) generation of 'next-best-tokens' in a sequence with so-called 'autoencoders' and (e) a special application of encoder/decoder systems to a graph-based weather forecasting model (see Mitchell, 2023; Remark: The combination of rule-based graph-structures with ANN-based elements such as encoders/decoders is sometimes referred as 'neuro-symbolic AI').

Reactions of humans on AI reveal more about humans than about AI technology! And the book of Terry A. Winograd and Fernando Flores (1986) about 'Understanding Computers and Cognition' is – still – a benchmark for a realistic perspective on the possible capabilities and fundamental limitation of AI.

The second line of development are data-based approaches and especially statistical classifiers (see Fig. 2.1). This development started with Norbert Wiener's 'Cybernetics' and the idea that gathered data can reveal statistical patterns, which can be used to set-up classifications, and which can be applied on new events to achieve 'predictions' - in the sense of a probabilistic classification of the new event according to the historical classification. The development of this line spitted into two sub-lines - the algorithms of 'Machine Learning' (ML) and 'Artificial Neural Networks' (ANN) – both with the approach to '*learn'* from data, which means [quote] '*to fit a function to a collection of historical data points'* according to Judea Pearl and Dana Mackenzie (2018). Although ML achieved successes to fit statistical classifiers to smaller datasets, this paper will skip the discussion of ML-methods such as Support Vector Machines, Random Forests, et cetera and the reader is referred especially to the book of Shai Shalev-Shwartz and Shai Ben-David (2014). Likewise, all types of 'Probabilistic Reasoning' (PR) including Bayesian Networks and other graph-based network approaches are not taken into account (see Pearl, 1988).

However, the terminology 'Machine Learning' was derived from or – at least – aligned with 'Artificial Intelligence', both being pure termini technici, but signalling some kind of human-like abilities. Although the technical implementations were simple 'fit functions', the terminology supported some public awareness that AI could be 'more' than multi-dimensional, non-linear regression.

The development of ANN, which resulted in the contemporary hype about GenAI, started with Frank Rosenblatt's '*Perceptron'* in 1957 and Kunihiko Fukushima's '*Neocognitron'* in 1980, which were basic networks of elementary artificial 'neurons' as very simplified version of the processing of electric impulses in biological nerve cells. A superb summary including references of the history of ANNs until 2014 was given by Jürgen Schmidhuber (2015). A schematic description of an elementary artificial 'neuron' and simplified diagrams of representative types of ANNs is summarized in Fig. 2.2.

All these individual applications of an elementary artificial ‘neurons’ within a network[4] are aligned to dedicated problems - i.e. the prime decision about the topology of the networks depends on some external ‘world knowledge (sic!). Consequently, pattern recognition will be treated with Convolutional Neural Networks (CNN, which are acyclic graphs, and compress the local neighbourhood in a picture before further processing to match an external label; see Fig. 2.2 b), improvements of time series such as speech transmission are processed with Recurrent Neural Networks (RNN as cyclic graphs with some internal feedback loops and especially with some ‘memory’ about the history of processes and especially with Long-short Term Memory ‘LSTM’; see Fig. 2.2 c), so-called Autoencoders typically used for Generative AI (GenAI, see Fig. 2.2 d) today, or dedicated applications like encoder/decoder systems with a globe-like network structure for weather forecasting models (so-called Graph Neural Networks ‘GNNs’; see Fig. 2.2 e). While all ANNs consists of simple elementary artificial ‘neurons’, the alignment of the network structure to the specific topology of the problem (flat images, long time series, sequences of words/tokens, or interacting weather cells on a globe) is part of the ‘practical magic’ of ANNs.

The elementary building blocks shown in Fig. 2.2. are were developed from the late 1980s to the end of the 1990s. One prominent example of a more complicated network is the seven-level CNN ‘LeNet-5’ developed by Yann LeCun et al. (1998). Since then, the last 25 years of development can be characterized by an incredible increase in available computing power and storage capacity: The latest GenAI systems exceeded the threshold of one billion parameters (to be fitted) and Terabytes of training data. A brief description of this development exemplified by the keywords ‘supervised learning’, ‘reinforced learning’ and ‘generative AI’ will be given in the following chapters.

[4] In like manner as a single ‘artificial’ neuron is fully deterministic and the internal calculations can be recalculated externally, all simple ANNs as shown in Fig. 2.2 b)-d) are transparent and can be tracked. The problem of so-called ‘back boxes’ of ANN is not a fundamental issue but caused by the tremendous number of ‘neurons’ in state-of-the-art ANN with up to or even more than one billion parameters. From a practical point of view, it is impossible to re-calculate any statistical classifier with such a dimensionality.

Nevertheless, a warning should be given, as e.g. ‘supervised learning’ means that a ‘learner’ uses input of ‘labelled’ training data (i.e. data such as images plus a description), while ‘unsupervised learning’ means that a ‘learner’ uses input of training data ‘sequences’ (i.e. sequences of tokens such as words). In both cases, the selected and prepared training data are used as input of an ANN.

Maybe, the strangest aspect in this history of AI[5] and especially ANNs is one thing: The focus has been on algorithms, technical implementations, and structures of networks, but not much is said about data and statistical classifiers!

[5] There have been other approaches in AI or in related fields – e.g. so-called ‘Fuzzy Logic’ – for some niches, which will not be discussed in this paper except one remark. Fuzzy Logic could be used to build control system (especially for non-linear systems as a simple, but rather stable approach e.g. for crane trolleys) or for logical inference of qualitative facts. Nonetheless, the narrative that Fuzzy Logic could be used to predict ‘fuzzy’ financial markets is mere semantics without an understanding of the deterministic mathematical concept of Fuzzy Logic.

3. Statistical Learners and Statistical Classifiers

Although the book by Shai Shalev-Shwartz and Shai Ben-David (2014) about '*Understanding Machine Learning*' has only one brief chapter on ANNs and was written before GenAI, it provides a marvellous overview as '*A Gentle Start'*. Here is a summarised version of this introduction to the statistical learning framework:

> *Image an example that you just arrived on some tropical island and want to buy papaya on the local market without experience. You have to derive a classification about papayas based on features (like colour and softness), and you make an experiment with a sample of papayas to examine which fruits characterized by 'colour' and 'softness' are tasty or not-tasty. The task is to find a classifier or prediction rule based on the methodology of statistical 'learning' (i.e. 'learner' as a terminus technicus for an algorithm 'to fit a function to a collection of historical data points' as pointed out by Pearl and Mackenzie, 2018).*
>
> *The 'learner' has access to a domain set $\mathcal{X}$ of objects we wish to label. The domain points will be represented by a vector of features (like colour and softness). Let $\mathcal{Y}$ be the set of possible labels: in this binary example {0, 1} for not-tasty and tasty. The experiment provides training data $\mathcal{S} = \{(x_1, y_1) \ldots (x_n, y_n)\}$ that is a sequence of labelled points in $\mathcal{X} \times \mathcal{Y}$ as input to the 'learner'. The instances x_n of the training data are generated by some probability distribution $\mathcal{D}$ over $\mathcal{X}$, and the labels follow a labelling function $y_i = f(x_i)$, whereas neither $\mathcal{D}$ nor $y=f(x)$ is known to the learner.*
>
> *The output of the 'learner' is a classifier $h: \mathcal{X} \rightarrow \mathcal{Y}$ (also called hypothesis or prediction rule). This classifier can be used to 'estimate' or 'predict' the label of a new domain point. In our example, this means whether another papaya of a certain feature can be estimated to be tasty or not. The error of the classifier is the probability that for a randomly drawn new instance x, according to the distribution $\mathcal{D}$, the classifier h(x) does not equal f(x) i.e. the ex-ante estimation does not match the reality f(x), when f(x) would be controlled ex-post. This is the well-known problem that any repeated medical test will produce true positive and true negative results, but also false positive and false negative results, when the ex-ante diagnosis is compared to an ex-post verification of the actual situation.*

It is important to repeat that the learner is 'blind' to world knowledge beyond the training data $\mathcal{S}$ and does neither know the underlying distribution $\mathcal{D}$ nor the 'real' labelling function f. In the example of the papayas, the 'learner' has neither any causal model why papayas taste good or bad, nor any information whether the selected instances – i.e. papayas from a local market – represent the 'actual' distribution of papayas and whether there are some variations et cetera.

Without entering into the mathematical formalism, this example triggers questions:

1. What 'error' do we want to achieve (e.g. false positive vs. false negative)?
2. How do we pre-select the algorithm for the fit-function *h(x)*?
3. What do we know about our experiment and 'detected' distribution $\mathcal{D}$?
4. Can the classifier h(x) be 'generalised' beyond the original domain set $\mathcal{X}$?
5. Does the correlation (x_i, y_i) indicate any casualty?

Ad 1) 'error' of statistical classifiers is often ignored but is common for medical tests[6]. An example is given by Judea Pearl and Dana MacKenzie (2018) with the statistical quality of an (somehow outdated) test methodology for breast cancer in the case of forty-year-old women in the USA. Comparing diagnosis (= ex-ante prediction) versus reality (= ex-post verification) this example for 3.000 women has 3 cases of True Positive, 1 case of False Negative (i.e. an actual cancer, which was not detected), 2636 True Negative but also 230 False Positive (i.e. cancer diagnosis, which were incorrect). Although the sensitivity (of correctly detected cancer = TP/(TP+FN)) was 75%, the specificity (=TN/(TN+FP) or ability to correctly reject healthy patients= was quite low with 88%. In other words, 360 healthy women out of 3.000 were – wrongly – given a cancer diagnosis, which causes a tremendous psychological damage. Consequently, mass screening with this method was not recommended for women without other known risk factors[7]. Especially for situation with a low number of positive detections out of a large sample, the adjustment of a test or classifier depends on the primary objective and of the costs generated by False Positive classifications.

[6] Although the discussion about 'black box' models is typically connected with ANN due to the incredibly large number of parameters, many medical therapies are 'black boxes', for which the general biochemical processes are known, but the multidimensional details of human reaction on a drug is not known from the beginning. Therefore, RCT are the gold standard to test the reactions compared to control groups.

[7] A discussion about the (different) age limits for mass screening in USA and Germany can be found in a recent interview (Heindel, 2024).

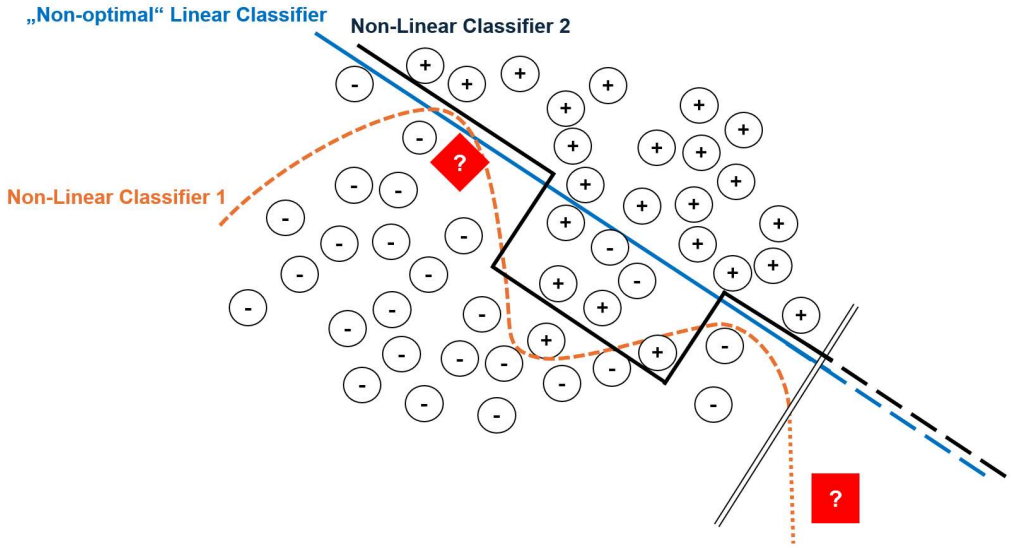


Figure 3.1: Illustration of a two-dimensional distribution (y, x) of instances of labelled training data (circles) together with a ‘non-optimal’ linear classifier (but best in class to exclude ‘false positive’ classifications of “-“ as “+”) and two non-linear classifiers. The boxes indicate a new instance “?”, which has to be classified and lies in the original domain or outside the original domain (asking about generalisation).

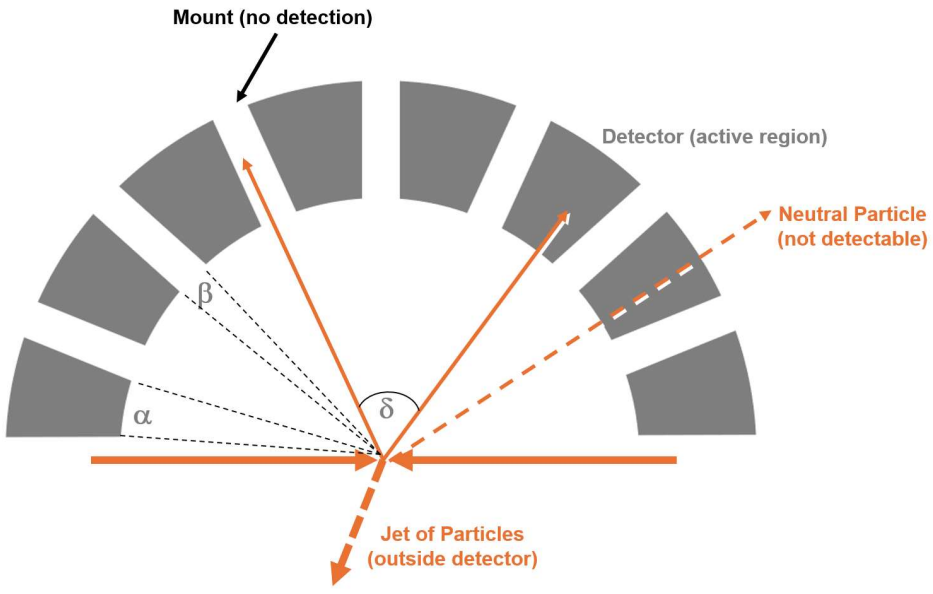


Figure 3.2: A schematic experiment in particle physics with a collision of two ingoing beams and reaction products to be detected by a ‘non-perfect’ detector set-up. Not shown is the typical shielding against external ‘noise’ like cosmic rays et cetera. In other words: Experiments always try to be ‘reductionistic’ and isolate the measurement from external effects of the ‘context’.

As illustrated in Fig. 3.1, even ‘non-optimal’ linear classifiers, which do not identify all positive instances correctly but avoids False Positive classifications at all, could be more suitable in such a situation.

Ad 2) As illustrated in Fig. 3.1 for a simply two dimensional case (like for the papayas with colour and softness), the distribution of the training data will typically show no clear ‘cut’ between the labels (here: “+” and “-“). As illustrated, different selected ‘fit function’ will product different results concerning quality parameters such as sensitivity and specificity. It can be feasible for simple low-dimensions situation and traditional ML algorithms to decide *ex-ante*, which fit function matches the quality requirements on the one side – and also the intrinsic noise of the sample (depending on the circumstances of our experiment). For higher dimensionality and for highly non-linear ANNs with millions or even billion of free parameters, the selection of an algorithm or structure of the ANN as indicated in Fig. 2.2 and the starting point for the setting of the fit parameters lacks a strong theoretical foundation. While some choices such as CNNs for recognition of (labelled) images, RNNs for times series and Encoders for Language Models are established, the details can resemble ‘*practical magic*’ and depend on the experience of the developers.

Ad 3) Today, more and more collections of training data are just ‘scraped’ from the internet like collection of (labelled) images or text corpora for the training of Large Language Models (LLM)[8,9]. Fig. 3.2 shows a schematic experiment in particle physics with a collision of two ingoing beams and reaction products to be detected by a ‘non-perfect’ detector set-up with ‘active’ detector regions and ‘inactive’ mounts and frames, but also non-detectable particles (by the specific detector) due to the nature of those particles or the set-up of the detector configuration. In the ‘experiments’ of buying papayas from a local market, there is an implicit assumption that the instances x_n of the training data are generated by some probability distribution $\mathcal{D}$ over $\mathcal{X}$, which is representative for the local market.

[8] Unfortunately, such a trend to ‘use’ data-sets, because they are ‘available’ can be found in scientific studies in economics and social sciences compared to physics, chemistry, biology et cetera, which are based on experiments. Some studies in social sciences have been based on surveys with ‘pre-registered’ participants, who were paid for their participation by specialized online platform. However, a collection of data without detailed understanding of the conditions of the ‘experiment’ or with a non-representative pre-selection, in which these data were gathered, is without value for statistical analysis.
[9] As the available corpus of human texts ever created is limited, there is a natural limit for scaling of LLMs, as additional parameters above a certain threshold would lead to ‘overfitting’.

But are there any ‘blind spots’ or any hidden ‘bias’ due to season, selection of merchants or circumstances that change the iid-assumption (that all instances are independent and identically distributed)? If there was a rush of tourist busses before, ‘our’ papayas could be the left-over. And is ‘our’ market representative for all markets on this tropical island?

The generation of the training data – or simply: the experiment – is typically outside of the discussion about statistical learning, while the detector response function, the geometrical coverage, the signal-to-noise ratio, the resolution et cetera are key parameters in physics experiments. Without a clear understanding of the ‘experiment’ and the generation of training data, a naïve use of data is subjected to hidden assumptions, unmeasured confounding, ignored noise and disregarded errors.

Ad 4) Another issue, well-known in medical statistics, is generalizability and transportability (see e.g. Degtiar and Ros, 2023). Can the results of a randomized controlled trial (RCT) be ‘generalized’ beyond the original domain set $\mathcal{X}$ of patients with distribution $\mathcal{D}$ (and potentially hidden dependencies on external parameters) and be transported to other populations as a general treatment, medication, or therapy?

A statistical learner is ‘blind’ to any world knowledge and does neither know the distribution $\mathcal{D}$ of the training data (only the correlations within) nor any causal relations between the features and the label. Therefore, it is impossible to decide about generalizability and transportability without an additional model.

Ad 5) If casual information is not available, a simplified model could be based on comparisons of statistical parameters between test population (or training data), target population and an ex-post control analysis group. Nonetheless, merely statistical attempts can easily overlook hidden dependencies, unobserved confounding, or unnoticed mediators (see Pearl and Mackenzie, 2018, for details about the statistical concepts). Vice versa, an ideal situation would be a complete understanding of the underlying ‘physical’ processes **and** the experiment for the generation of training data. However, physics differ from medicine or economics. Even in medicine or pharmacy, not all biological processes related to the treatment of a disease might be known in full details. Consequently, new drugs or treatments are tested very carefully in clinical trials (and best with RCT as gold-standard).

However, a causal model such as e.g. a directed graph is required to derive more than correlations. As explained by Pearl and Mackenzie (2018) and elaborated with the whole formalism by Judea Pearl (2000, 2010) in his book 'Causality', it is at least approximately possible to derive causal relationships from measured data-sets with the help of such Structural Causal Models (SCM).

The application of all these concepts to ML and AI is not trivial. For certain cases like ML of frequency spectra of a rotating machine in correlation with future failure as label, there is a technical relationship between increased sound due to rattling (at t_0), underlying fatigue of material (at t-x), and failure (at t+x), which can be used for so-called 'predictive maintenance'. But what is the criterium for a CNN to classify images (so-called 'image recognition')? Of course, there is no 'physical' relationship, but certain structures of images (such as 'characteristic silhouettes') used as training data can correlate with the labels. Sebastian Lapuschkin et al. (2019) and Christopher J. Anders et al. (2022) pointed out that artefacts in the training data could be the determining factor in image classification and [quote from Anders et al., 2022]:

> ' *... without removing, or at least considering such data artifacts, learning models are prone to adopt Clever Hans strategies [based on spurious correlations in the training data], thus giving the correct prediction for an / wrong reason."*

One intriguing example (see HHI, 2019) is the 'recognition' of trains in images of a standard image corpus, but due to the 'rails' as artefactual reason[10].

These observations were – already ten years ago – summarized by Shai Shalev-Shwartz and Shai Ben-David (2014) in their 'No-Free-Lunch theorem' [quote]:

> *'The No-Free-Lunch theorem states that there is no universal learner. Every learner has to be specified to some task, and use some prior knowledge about the task, in order to succeed. So far we have modelled our prior knowledge by restricting our output hypothesis [i.e. h(x)] to be a member of a chosen hypothesis class [i.e. $\mathcal{H}$].*

[10] It is beyond the scope of this paper to enter the discussion about 'Explainable AI' (XAI). The mentioned example (HHI, 2019) is taken from work about XAI, and in the case of image recognition different XAI methods provide insight, which features in a picture determine how CNNs classify images. This is a technical approach to understand features (like edges, shapes, colours et cetera) and the effect on parameters (weight factors) in a CNN. However, statistical classifiers are probabilistic tests and have False Positive and False Negative classifications. Therefore, the focus of this paper is on data and statistics

This ‘No-Free-Lunch theorem’ is a warning against some virulent approaches to (i) take some ANN-tool (say: a CNN for image classification or a LLM for ‘next-best-token’ text generation) and then (ii) take some data corpus available on the internet (say: as a standard image collection or archive of a web-crawler) to solve a dedicated task according to specific objectives. The current hype about technologies like GenAI let us forget that the foundation of any AI is statistical learning, and all AI-tools are statistical classifiers and that for any statistical classification the underlying data domain has to be understood.

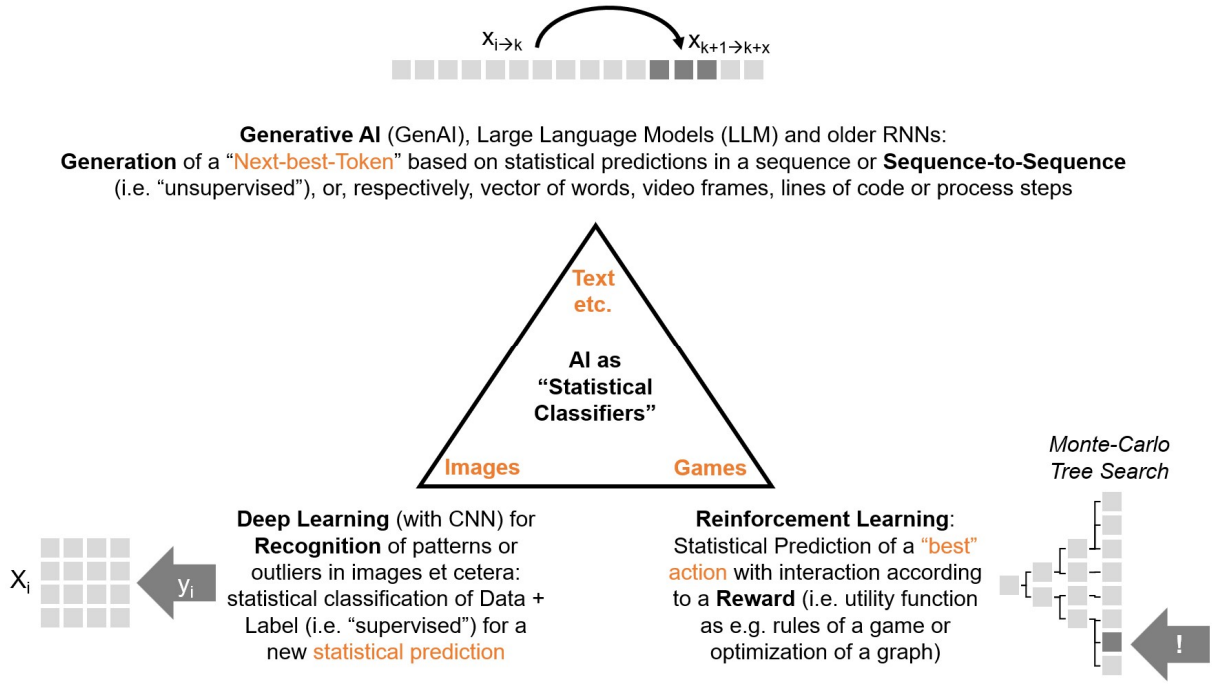


Figure 4.1: Current developments of ANN-based AI (see text for details)

4. Current Trends and Issues

As illustrated in Fig. 2.1, the time around 2015 was a milestone in AI history. On the one side, the book of Ian Goodfellow, Yoshua Bengio and Aaron C. Courville (2015) provided a comprehensive summary including the underlying mathematics of 'Deep Learning' from CNN via RNN to Encoder-Decoder-Systems (at that time). For many questions about the 'math of AI', the reader can still be referred to this book[11]. On the other side, the current hype of GenAI was only one of many developments described in this book – and nobody might expect the tremendous progress of GenAI since then, especially in public awareness.

Skipping the development from 2015 to 2024, Fig. 4.1 is a fast forward to the main lines of development of ANN-based AI, as they are state-of-the-art. While for example Recurrent Neural Networks (RNN) were advanced choices to analyse structures in (infinite) time series, sequences and/or time-ordered lists in 2015, the development of so-called 'transformers', i.e. multi-layer encoder-decoder architecture based on attention mechanisms (see Vaswani, et al., 2017), entered the stage and are the basis for all hyped developments in GenAI today.

[11] There are other approaches like the State Space Model (SSM; see e.g. Gu, Albert et al., 2022), which can be used to describe continuous time series but can be (i) discretized to RNNs and (ii) unrolled to CNNs.

The main development of contemporary ANN-based AI includes:

- Deep Learning for image / pattern recognition with CNNs (although RNNs belong to DL likewise)
- Reinforced Leaning for games or other ‘rule-based’ processes including the question of autonomous vehicles
- Generative AI (GenAI) as a ‘sequence-to-sequence’ method and for generation of a ‘next-best-token’ in Large Language Models (LMMs)

As many books – such as Goodfellow, Bengio and Courville (2015) for ‘Deep Learning’ or Daniel Jurafsky and James H. Martin (2024) for ‘Speech and Language Processing’ – provide insight into the mathematics of ANN-based AI and elaborate on the mathematics of matching expected output (‘target data’) versus input (‘training data’), this brief summary will emphasise two special issues: public expectations and questions concerning data from statistics to copyright law.

Taking the original papaya example with the original challenge to classify ‘features+label’, one can extend this example to image recognition of pictures with external labels. Of course, one can gather a corpus of images from the internet and select papaya images plus a label as ‘papayas’. However, this leads to the question how to classify ‘tasty / non-tasty’ from images? Once again, this general question generates many detailed ones:

- Can we identify papayas in images? → yes
- Can we classify ‘taste’ only from images? → no
- Can we transfer our experiment (with papayas from a local market) to images scaped from the internet? → no, as ‘taste’ is not accessible
- Can we simplify our experiment by taking pictures and evaluating the ‘taste’ to transfer the label ‘tasty / non-tasty’ to images? → potentially yes, if images are a reasonable proxy for colour/softness
- Can we control that images scaped from the internet have the same feature/label correlation as our experiment? → no
- Could the ‘papaya’ case generalize to other fruits? → no
- What would happen, if we would take images of papayas with attached barcodes indicating their ripeness, quality and other parameters?

While it seems to be easy to transfer the original example (with features→label) to image recognition, this is not a simple thing. Let's make another experiment with purchased papayas to (i) check the 'taste' and (ii) take a picture. Even if we can find a statistically relevant correlation between our (sic!) images and our taste, the domain is limited. Any generalization to images of papayas in general depends on many assumptions and treatment of the date-set. And what happens if we provide an image of an avocado? The last questions may sound absurd but in the context of Large Language Models (LLMs), many current publications discuss 'emergence' of features of LLMs beyond the training data.

In a second approach, we take only those papaya images showing a barcode with all relevant parameters. This approach changes the features+label methodology to an image of 'papaya with included parameters'. Unfortunately, there is an established 'technical' terminology to describe the feature+label methodology as 'supervised', while the image with included parameters would be described as 'unsupervised' or 'self-supervised'. Unfortunately, this terminology of 'un-/self-supervised learning' is often abbreviated to 'self-learning', although no computer program including any ANN does anything 'itself'[12]. All programs are executed according to the programming plus preparation of the runtime environment by the creators: i.e. an ANN algorithm as a dedicated 'learner' selected for the specific task + human selection of training data + human selection of the objective to be achieved + human definition of statistical quality criteria. The distinction of supervised versus unsupervised/self-supervised is a terminus technicus and nothing more.

All these questions are <u>not concerning the mathematics</u> of 'fitting advanced ANN-based functions to training data', but the assumptions, limitations, and problems of the basic set-up of our experiment! The *Gedankenexperiment* with papayas shows the steps from *Features + Labels → Images + Label → Image incl. Parameters → Data Structures* and reveals that all ANN-based AI has the objective to classify a new instance based on a set of training data, i.e. h(x) = y. Taking CNNs and image recognition as an architype, there are two different directions. The first is so-called 'Reinforced Learning' (RL) as typically applied for games.

[12] It is interesting to note that OpenAI (2023) writes in the 'GPT-4 Technical Report' explicitly [quote]: *'Despite its capabilities, GPT-4 has similar limitations to earlier GPT models: it is not fully reliable (e.g. can suffer from "hallucinations"), has a limited context window, and does not learn from experience.'*

The AI systems is trained to find a ‘best action’ with a maximization of a reward h(x) = max(x) and in the end to ‘win’ the game. Instead of a predicting a fixed target or label y, games are characterized by a reward for winning and maximization of a rewards. Typically, RL applies AI architectures like CNN but scaled up tremendously and combined/enhanced with additional methods such as ‘Monte Carlo Tree Search’ to navigate the increasing number of possible combinations of moves. Either two computers are playing an incredible number of games against each other to ‘learn’ the best way to win (based on probabilistic simulation and exploration of promising moves) according to an external rule for rewards or one computer is ‘reinforced’ by a human ‘trainer’ with a reward for the actions. As even a supposedly ‘simple’ game like Go has an incredible number of possible moves – much more than Chess – the combination of probabilistic Monte Carlo Tree Search plus brute force by running the game between two computers millions of times was key to the success of programs such as Google DeepMind’s AlphaGo or AlphaZero – literally ‘autodidactically’.

Many observers were astonished that AlphaGo found new strategies to win. However, Minkyu Shin et al. (2023) revealed a stagnation of human players over a long time and the human decision quality stayed pretty uniform for 66 years (since 1950). After the unexpected successes of AlphaGo in the 2016–2017 period, the decision quality of professional players increased as they adopted the new ‘unconventional’ strategies and developed novelty indecision-making. However, the example of AlphaGo showed that RL-systems can trigger human creativity to remove blinkers and conservative thinking, but such systems cannot develop any ‘emerging’ capability beyond the pre-defined rules of repeated games. We will see later that this limitation to follow rules is one major problem with so-called autonomous vehicles.

The last approach is the development of ‘Sequence-to-Sequence’ methods instead of ‘Feature+Label’. In 2015, ‘Recurrent Neural Networks’ (RNNs) would have been state-of-the-art, but since 2017 the approach of ‘Generative AI’ (GenAI; also with synonyms such as ‘foundation models’ or ‘transformers’) dominated the professional and public awareness. While CCNs et cetera are dedicated ‘learners’ to classify a list of features or asset of spatial data like images to match pre-defined labels, GenAIs are dedicated ‘learners’ to match a sequence of values $x(t_1), \ldots, x(t_n)$ to a ‘next-best-token’ $h(x, t_0) = x(t_{+1})$ or to a ‘transformed’ sequence like a translation of a sentence from one language inti an another language $h\{x_1, \ldots, x_n\} \rightarrow y\{y_1, \ldots, y_n\}$.

With some creativity, one could extend the papaya example via the papaya images with embedded barcode descriptions to a new challenge to ask a 'generative' tool: 'Please, generate a picture of a tasty papaya'. Such a query does not solve any problem, i.e. it does not help to select tasty papayas you are going to buy.

We can switch to another problem to 'answer' the question or – more concretely – to complete this sentence x= {During the day, all my cats were *[y=h(x)]*}. This example will be discussed in detail in a following chapter.

From a formalized perspective the two approaches can be written as:

- h [features of papayas(x)] = y
- x = {During the day, all my cats were *[y_{k+1}=h(x_1, …, x_k)]*}

Independent, weather this is a directed or cyclic (or recursive) relationship, the challenge is always to '*fit a function to a set of data'*.

As a GenAI has to be 'trained' in the same way as CNNs or RNNs, the domain defined by the training data (even for tremendous corpora scraped from the internet) also defines and limits the possible probabilistic outcome. Without any further causal model, no CNN, RNN or GenAI can go beyond the original domain to classify a new instance or estimate a 'next-best-token' to continue a sequence.

Nonetheless, there is one major difference in the training process between (i) features+label and (ii) sequences {x_1, …, x_k; x_{k+1}}. In the first case we make an experiment, and we determine actual values for labels. Of course, all experiments have to deal with noise and there will be some 'wrong' labels - however, most labels are 'right'.

Vice versa, a text corpus scraped from the internet will contain many sequences {x_1, …, x_k}. or nearly similar sequences, but they will show a probabilistic distribution of the last token {x_{k+1}}. Image the *Gedankenexperiment* with the following sequence {During the day, all my cats were *[x]*}. In an extremely large text corpus, we will find a fictive probability distribution for the token *[x] of [value; probability] = [sleeping; 0.50], [eating; 0.30], [playing, 0.15], [...], [gaming, 0.001], [...], [gambling, 0.0001], […]*. This distribution runs from very probable and rather realistic values towards very fictitious value making the sentence Lewis Carroll like.

This *Gedankenexperiment* returns to the general question, where do training data come from. The strength of GenAI and *Large Language Models* results from the incredible amount of text corpus data available on the internet, where nearly all sentences we can image (i.e. nearly all sentences ever written by mankind) are included and build the foundation to find 'next-best-token' continuations for any sequence we use as input prompt to such LMM. Simply, there is no 'truth' in a LMM trained with text corpora from the internet, but a mixture of facts, fiction, faults, and lies (see especially Léon Bottou and Bernhard Schölkopf, 2023, as quoted in chapter 1).

As already mentioned, LLMs can be finetuned either with rule-based post-processing or in a second training step with 'Reinforcement Learning from Human Feedback' (RLHF) or with 'Reinforcement Learning from AI Feedback' (RLAIF[13]). Formally, this is a shift $h(x) \rightarrow h'(x)$ with an objective $h'(x) = f'(x)$ to fit a 'planned' function $f'(x)$ instead of the actual $f(x)$. However, the disaster of Google' new multimodal GenAI tool Gemini in Feb. 2024 is a warning call. According to reports *inter alia* by Tom Warren (2024), Google's attempt to force the tool to a planned distribution of parameters of 'diversity' in image-generation resulted in a distorted generation when images should be matching historical pictures (with the historical distribution of racial and gender parameters).

Already in early 2023, a discourse evolved about potential political biases of LMMs and 'unfair' models. Especially Shangbin Feng et al. (2023) published a study and concluded [quote, underlining by the author]:

> *Generally, [Google's] BERT variants of LMs are more socially conservative (authoritarian) compared to [OpenAI's] GPT model variants. This collective difference may be attributed to the composition of pre-training corpora: while the BookCorpus played a significant role in early LM pretraining, Web texts such as Common-Crawl and WebText have become dominant pretraining corpora in more recent models. Since modern Web texts tend to be more liberal (libertarian) than older book texts, it is possible that LMs absorbed this liberal shift in pretraining data.*

[13] This is also known as 'Constitutional AI' because the feedback is not based on individual preferences of human 'reviewers' for harmlessness but on a set of 'constitutional' principles. However, these principles were – once again – defined by humans *ex-ante*.

While it is strange that Feng et al. (2023) relate *conservative* = *authoritarian* and *liberal* = *libertarian*, which does not correspond with the history of political ideas (sic!), the assertion about 'conservative books' versus 'liberal Web texts' reveals more about the authors of this paper than about LMMs.

In a recent preprint, David Rozado (2024) found a different pattern and, especially, constructed a new hypothesis that post-training (i.e. benevolent 'corrections' of the base model) is the primary cause of the political bias [quote]:

> *The results indicate that when probed with questions/statements with political connotations most conversational LLMs tend to generate responses that are diagnosed by most political test instruments as manifesting preferences for left-of-center viewpoints. We note that this is not the case for base (i.e. foundation) models upon which LLMs optimized for conversation with humans are built. ... Though not conclusive, our results provide preliminary evidence for the intriguing hypothesis that the embedding of political preferences into LLMs might be happening mostly post-pretraining. Namely, during the supervised fine-tuning (SFT) and/or Reinforcement Learning (RL) stages of the conversational LLMs training pipeline.*

It is the old questions, asked by Juvenal (Roman poet, 58-138?) nearly two thousand years ago: "*Sed quis custodiet ipsos custodes?"*

Once again, there is no 'truth' or no 'right or wrong' in LLMs, which are probabilistic by design. In extension of the original approach to statistical learners $h: \mathcal{X} \rightarrow \mathcal{Y}$ to predict actual labels based on measured features, the usage of LLMs in a specific context poses the question, which 'learned' distribution (i.e. scraped text corpora for training) should be applied to generate a 'target' distribution.

As an alternative, one can apply a different approach and use either a carefully curated text corpus (e.g. internal documents of a firm) or a transcript of recordings of cats' behaviour (again in the sense of an experiment to generate data in a controlled situation). While such domain-specific 'Small Language Models' (SLM) trained in this way cannot escape to be generically probabilistic, curated data would restrict the distribution of training data. Nonetheless, SLMs with 'known' distributions of training data resemble traditional approach with statistical regression in multiple dimensions.

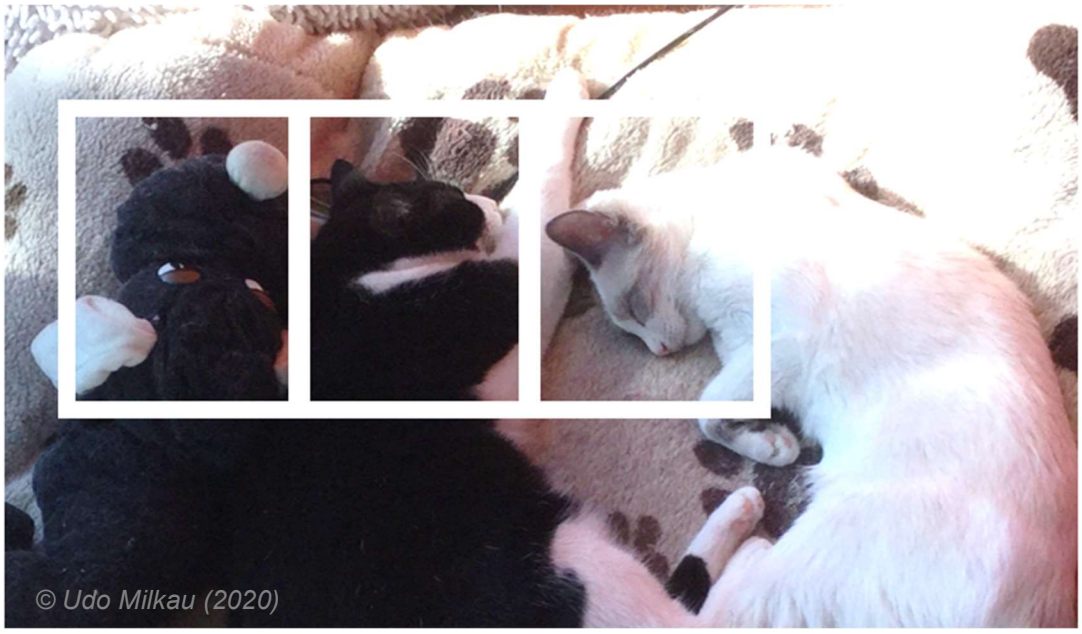


Figure 5.1: Example for an experiment to describe the circumstances for image recognition, i.e. nearly black-and-white picture and only cats and 'dogs'

5. Deep Learning for Pattern Recognition

Since Yann LeCun's *LeNet-5* of 1998, Deep Learning with CNN is an established method for image recognition and pattern recognition in similar circumstances with fixed lengths inputs plus labels. As illustrated in Fig. 2.2., CNNs follow the natural structure of pictures that the near neighbourhood of a pixel is strongly correlated and use this assumption to 'convolute' an image (via a number of layers) to a 'reduced' output, which is matched with an external labe[14]l. This technical process of optimizing the CNN – i.e. all the weight parameters w_{ij} in the CNN – is call 'backpropagation' as the error $E = E(t_i\text{-}o_i)$ of the calculated output of the CNN compared to the 'correct' label as target is used to derive changes to the weight $\Delta w_{ij} = -\eta \, \partial E/\partial w_{ij}$.

This approach to minimize an error-function shows the near relationship with textbook linear regression with the approach: $min \, \Sigma \, (\alpha + \beta \, x_k - y_k^2)$. The major difference is the dimensionality, as linear regression is a fit of a linear function to data point in two dimensions, whereas CNNs are 'trained' in a multi-dimensional space with millions or even billions of parameters (i.e. the w_{ij}) and have non-linear elements such as the activation function to provide the outputs of the artificial neurons (such as typically *tanh(x)* , the *sigmoid* function or *ReLU* (rectified linear unit f(x) = ½ (x+|x|)).

[14] Additionally, CNNs work from layer to layer: First, features like edges and lines are processed and, later, objects like cats and dogs (see Fig. 5.1).

Figure 5.1 illustrates the general limitations and assumptions. If we generalize this picture as the overall training data for a task to distinguish cats and dogs (although the dog in this picture is a plushie), any CNN-based statistical classifiers only (i) distinguishes cats from dogs but not 'recognise' any other animals, (ii) work well with nearly back-and-white pictures, while a 'colourful' tiger in a jungle would be outside the domain, and (iii) cannot detect the false image of the plushie instead of a actual dog leading to *'False Dog'* classifications in the later use (as extension of *False Positive* and *False Negative* in the case of +/- classifiers).

The strengths of CNNs results from the emulation of visual nerve: the input of an artificial 'neuron' is connected only to a set of neighbouring 'neurons' (i.e. the convolution). The tremendous development and obvious successes of CNNs has been hiding that CNNs – like any other statistical 'learner' – are limited by the selection of the training data and the corresponding statistics:

- Are training data representative for the domain? → Do we have images of real cats and dogs?
- Are there technical constrains in the image set like quality, resolution, colour of images et cetera? → E.g. only black-and-white images?
- Do the training data contain artefacts? → E.g. due to the circumstances, in which the pictures were taken or due to correlations between the technical device and ex-ante assumptions (like high-tech medical devices used in specialized hospitals and for patients, who are assumed to suffer from a complicated disease)?
- Is the labelling – to be provided by human experts - correct? → Or do we label a plushie as a 'dog'?
- Are enough 'typical' cases included, or are only 'typical' cases provided? → Characteristically, in fraud detection actual fraud transactions could be very rare compared to correct transactions, while in medical data-sets healthy (non) patients could be underrepresented.
- And is the chosen methodology of CNN adequate for the data structure? → Although many frequency spectra (e.g. for predictive maintenance of rotating machines) resemble pictures, there are cases e.g. in nuclear physics where Gamma-ray energy spectra from nuclear decay have sharp peaks and related peaks are rather spaced.

One of the most prominent examples for the use of CNNs (or Deep Learning in general) is the classification of medical images such as X-ray images, images taken by MRT devices, microscope images of the skin and many more. As such a classification is a medical test, the discussion in chapter 3 about the statistical features of cancer diagnosis provide the background for an evaluation of the quality of AI-based diagnosis versus human experts with the parameters of True Positive, True Negative, False Negative, False Positive, and derived parameters like Sensitivity = TP/(TP+FN) and Specificity = TN/(TN+FP). As this is an established tool-set in medical statistics, these parameters were used to compare the performance of DL against human professionals and to verify narratives of the superiority of DL.

Five years ago, a major study with a review and meta-analysis for test accuracy in detecting diseases from medical imaging by Xiaoxuan Liu et al. (2019) came to the following conclusion [quote]:

> *Comparison of the performance … found a pooled sensitivity of 87.0% (…) for deep learning models and 86.4% (…) for health-care professionals, and a pooled specificity of 92.5% (…) for deep learning models and 90.5% (…) for health-care professionals.*

Taking into account the statistical errors, DL performed similar to health-care professionals, but did not exceed the accuracy. This is quite reasonable as (i) human experts have to provide the labelling (i.e. original classification) of the training data and (ii) actual methods to detect diseases always deal with some noise, technical problems, and statistical errors. The main result of this study is not surprising: DL is not superior to health-care professionals, but it could either augment health-care professionals (e.g. to pre-select unambiguous results) or help in cases where no human experts are available (e.g. in cases of emergency – like the well-known science fiction of '*The Doctor'* as an '*Emergency Medical Hologram'* in Star Trek Voyager).

These results were supported by a recent complementing study of chest radiographs (Plesner et al., 2023) [quote, underlining by the author]:

> *Four commercial chest radiograph artificial intelligence tools detected airspace disease, pneumothorax, and pleural effusion with moderate to high sensitivity, but had <u>more false-positive findings than radiology reports</u> and <u>decreased sensitivity for smaller target findings</u>.*

Once again, DL can provide similar quality compared to human experts, but depends on the trade-off between sensitivity and specificity as objectives, and on the ‘strength’ of the signal (i.e. signal-to-noise rate especial for vague or small disease characteristics).

Compared to recognition of diseases in medical images, picture recognition does not play a major role in banking (compared to the insurance industry, where pictures have an important role e.g. in car insurance claims management). However, there are two use cases: capturing payment transaction data from paper bills (instead of SEPA request-to-payment transactions or billing with QR codes, which is established in the Nordics) and capturing transaction data in trade finance e.g. for bills of lading (BoL). In both use cases, the documents are rather standardized (typical structure of a bill, key words such as ‘customer identification number’ or SEPA/IBAN account numbers with check digit et cetera) and the ’intelligent optical character recognition’ of uploaded bills for online banking works very well – and if not, the customer is asked to enter the data manually.

The situation is more challenging with ‘living’ documents for trade finance, which can be faxes/copies, have hand-written annotations, are dirty or partly damaged et cetera. Nonetheless, such as use case in trade finance has different objectives compared to medical images. One the one side, the objective is automation and even a – let’s say – 50% automated data entry of trade finance documents would be a significant improvement. One the other side, an automated routing for recognized documents versus not-identified documents could be integrated in the usual workflow of a trade finance back-office operations.

Although a statistical triviality, no classifier can be ‘better’ than the quality of the training data (= features + label). Nonetheless, there is the vision that AI could ‘learn’ something ‘themself’. The (wrong) narratives that AI could outperform the benchmark of the best human experts in medical diagnosis, could be disastrous for patients: both for false negative and for false positive results! A realistic assumption, for example, to capture 50% of a certain type of documents automatically is a much more realistic level of ambition, can be re-checked by the back-office computer systems, and be tested in parallel runs for a long time in a ‘clean’ set-up under full control of the back-office.

6. Reinforced Learning: Games, Vehicles and Dilemma

The concept of 'Reinforced Learning' (RL) is neither new nor very focussed. In principle, any statistical 'learner' can be regarded as an 'agent' making an action with impact on an 'environment' and receives feedback as a 'reward'. This concept already applies to very simple control loops with an actor and a sensor, if the 'agent' has some memory. Cutting a long story short for this summary, one can rewrite the definition of a simple 'learner' in chapter 3 from a classification *h* to a policy π:

$\pi: \mathcal{S} \times \mathcal{A} \rightarrow [0,1]$ with a probability map $\pi(s,a) = Pr(A_t=a \mid S_t=s)$

that maximizes the expected cumulative reward. The combination of a state machine with states $\mathcal{S}$ and eligible actions $\mathcal{A}$ and expected rewards is a typical game (or more general a Markov decision process).

In some dedicated cases – such as Reinforcement Learning from Human Feedback' (RLHF) as postprocessing in the training of LLMs according to human objectives (see chapter 4) – the feedback is provided by humans with individual decisions[15]. Yet, typical applications of RL are deterministic/rule-based games with a very large scale of possible actions like Chess or Go but with rather simple rules for rewards: to win the game. As already mentioned, the key success factor of current RL-tools for games is not the component of CNNs, but the handling of the huge dimensionality by (probabilistic) Monte Carlo decision trees. Although the rules of the games are fully deterministic, the dimensionality cannot be computed in reasonable times and (probabilistic) methods such as Monte Carlo decision trees are required to derive a successful policy strategy.

For the scope of this essay, such deterministic/rule-based 'games' are not very interesting, as neither the economy, nor finance, not daily life can be described as a finite state machine. Vice versa, it is important to understand the difference between deterministic state machines and financial systems or an environment such as traffic on public roads.

[15] It is worth to note that RL with individual human decision includes the danger that the 'human feedback' is wrong from an objective point of view. Nonetheless, it is always the 'trainer' who has the responsibility for the training. This includes that he provides 'correct' feedback but also that he understands all rules (and potential gaps in the rules) of the game. In RL, the 'learner' will be efficient to win the rewards – whether the reward/feedback is 'correct' or 'wrong' is an outside perspective, which is not available for the 'learner'.

Starting with the financial system there are many proposals how to ‘forecast’ price movements of assets from past time series such as equity and bonds to currencies, commodity and – currently – also so-called crypto-assets without any intrinsic value such and without any future cash-flow as Bitcoin. However, up to now there is no proof that the efficient market hypothesis (see especially: Eugene F. Fama, 2013) does not apply, if one recognises the following limitations to ‘efficient markets’[16].

First and undisputed, the efficient market hypothesis subtracts a general economic development of the economy (i.e. the ‘beta’ of a market depending on the overall economic development, market interventions by governments and central banks, or external factors like wars or sanctions). Second, there are well-known limitations – call ‘stylized facts’ – resulting from external constrains on a market: from composition of indices (with a signal for investment funds to buy, if they have a policy to follow a certain index) via limited liquidity of SME stocks/bonds to market intervention by central banks (governmental bonds, ‘green’ bonds et cetera). Third, there is the possibility to ‘be faster’, when new information is available e.g. due to automated analysis of news feed (‘sentiment analysis’), algorithmic trading and also physical co-location of the algo servers with brokers or exchanges. And fourth, there are psychological factors – herd behaviour – such as FOMO (fear of missing out), which is currently dominating the price developments of Bitcoin or Dogecoin, which is an ‘asset’ without any fundamental value, without any cash-flow, and without any scarcity (as there are many clones of Bitcoin and one can create new clones at any time), but is driven by the recent start of Bitcoin-ETFs attracting greedy investors.

Without going into more details, there are – well-known – patterns in the market, but there is no chance to ‘predict’ any alpha-add-on development of market prices based on some ‘learned’ historical time series beyond the general trend of liquid market or specific limitations in non-liquid markets. In other words: ‘efficient markets’ include all available information in the current prices and historical time series do not include any further useful information (beyond t_0) whether future market prices will go in the one or other direction and there is no ‘ergotic’ process, which can be extrapolated into the future based on a sufficient long history. Consequently, really ‘efficient’ financial markets are a game – but a fully probabilistic game such as dice without any possible strategy to win in the long run (except the casino itself).

[16] For the performance of actively managed funds see e.g.: Edwards et al. (2024)

An antagonistic example – although not plain vanilla RL – is traffic and autonomous vehicles. Skipping all the problem of computer vision or exact navigation, vehicles from autonomous lawn mowers via trains on walled rails to airplanes landing on aircraft carriers are using different variants of AI to navigate in clearly defined (sic!) situations: either stopping at small speed in any case of ‘undefined’ state or calling to a human back-up operation.

Nonetheless, the three examples are simple to describe and follow physical laws when interacting with the environment. The genuine challenge is autonomous driving in a city with many other and different ‘agents’ around.

Although strange at first glance, autonomous vehicles resemble games so far as the ‘autonomous’ agents are observer and player in one instance. Any – rather literally – move of the vehicle will change the situation of the environment as other agents will react on this ‘move’. While the term ‘autonomous’ is ambivalent – and no computer program has a free will or any intentionality beyond the intention of the programmer – an autonomous car with highest level 5 would be making its ‘moves’ without any human intervention. But different to lower levels of self-driving cars (e.g. on an Autobahn at some limited speed – like the new BMW 5 series - with good weather and street conditions, but without any exit, traffic lights, crossing pedestrians et cetera – and even then only with a human in the – literally – driver’s seat), a fully ‘autonomous’ car would require that ‘playing field’ could be demarcated, that the number of different ‘players’ is defined, and that all ‘players’ play according to the pre-defined rules. However, many participants in city traffic may behave ‘irrationally’: from playing children and drunken pedestrian to reckless drivers. And in some cases, we ourselves have to contravene the rules (e.g. to clear the road for emergency vehicles).

Contemporary[17] AI cannot handle so-called ‘corner cases’. Although one can try to record and include all actual events on streets in training data for ‘statistical classifiers’ (see e.g.: ‘*Google's Self-Driving Cars Use Halloween to Learn to Recognize Costumed Kids*’, Coldewey, 2015), no such classifier can ‘image’ counterfactual events and prepare for all situations in the always uncertain future. In an interview, Steven Peters (see: Armbruster und Winterhagen, 2023) made a *Gedankenexperiment*: What would happen if somebody would wear a fully realistic Halloween costume as a ‘garbage bin’, but cross the street between parking cars? In other words: How could a statistical (sic!) classifier classify a realistic ‘garbage bin’ as a Halloween costume of a playing child? Of course, the programmer could implement an extra safety rule for all days of Halloween, but also for Karneval, any Science Fiction convention, Manga exhibition et cetera?

One incident is rather characteristic. A self-driving taxi of General Motor’s subsidiary Cruise had an accident mid-2023, after which the state's department of motor vehicles revoked the license for Cruise and Cruise recalled all self-driving taxis in the USA in Nov. 2023. The crucial point is not in the technical details, as the accident was not caused by much discussed risks like problems in image recognition at bad weather conditions or adversarial manipulation of traffic signs et cetera.

The accident was, so to say, a ‘secondary’ accident, in which a passer-by was thrown in front of the autonomous vehicle due to a first (independent) accident and then dragged along for a few meters during a planned ‘safety manoeuvre’! This contradiction that a planned ‘safety manoeuvre’ caused the major harm reveals the limitations of all contemporary approach to real ‘autonomous’ driving including all attempt to develop taxonomies of corner cases (see e.g.: Heidecker et al., 2024).

[17] AI-based ‘autonomous vehicles’ follow a step-by-step approach from image recognitions (e.g. cars, bicycles, trucks, pedestrians et cetera) via extrapolation of trajectories of other agents to decisions about the own actions. Recently, a new firm Waabi (2024) proposed an approach to develop a ‘World Models for Autonomous Driving’ based on GenAI (see next chapter). The term ‘World Model’ is used here for a very dedicated model of the development of objects as ‘seen’ by LiDAR sensors: i.e. ‘world models on point cloud observations’, but not as a model of behaviour of agents in public traffic! While the ability of GenAI to ‘generate’ the most probably next position of various ‘point clouds’ including an extrapolation of the future ‘ego actions’ of the own agent is intriguing, such an approach does not capture the interaction of observer and participants (as this approach has no knowledge about ‘interactions’), and it is a mere probabilistic approach without corner cases (see: Zhang et al., 2024).

Consequently, David Autor (2024) overestimates the capabilities of (contemporary) AI as ‘statistical classifiers’, when he writes [quote, underlying by the author]:

> *AI’s capacity to depart from script, to improvise based on training and experience, enables it to engage in expert judgment - a capability that, until now, has fallen within the province of elite experts.*

AI-based systems have astonishing capabilities - and already collision detection and automated breaks are a great help - but they are bound to the domain of the training data or require *ex-ante* rules how to generalize to out-of-domain instances. AI can do many things but cannot ‘*depart from script’*.

Vice versa, current ‘self-driving’ car have to be ‘monitored’ by remote assistance teams to solve ‘insoluble’ conflicts in the programming and make human decisions how to move on. A report in CNBC (Kolodny, 2023) published the following quote concerning required remote assistance of Cruise’s ‘self-driving’ cars: ‘*The Cruise spokesperson wrote in an e-mail, that a ‘remote assistance‘ session is triggered roughly every four to five miles, …, in Cruise’s driverless fleet*.’

Nonetheless, there is a sometime bizarre debate that ‘autonomous vehicles’ should be able to solve decision problems, we as human beings are not able to decide unambiguously. In these cases, AI – i.e. statistical classifiers - should be able to ‘decide’ while humans have no rules how to decide! As Mark Coeckelbergh (2022) described, these philosophical discourses about decisions in moral dilemma show that AI system or robots are only mirrors of the human side.

This issue can be illustrated by the well-known moral-philosophical ‘Trolley Problem‘. This *Gedankenexperiment* (sic!) was discussed already by Karl Engisch (1930) and Hans Welzel (1951), in English references by Philippa Foot (1967) and coined as ‘Trolley Problem’ by Judith Jarvis Thomson (1976). In short, the dilemma is described as a - ultimately required and by definition unavoidable - decision between two fatal alternatives (rather literally: *Weichenstellung*): e.g. a train running either into a group of kindergarten children or a group of Nobel laureates. Without entering into the long history of arguments, one can focus of this dilemma as arguments against ‘autonomous vehicles’. For example, Catrin Misselhorn (2022) elaborated [quote]: ‘*If there is no morally acceptable solution to these dilemmas, this might become a serious impediment for fully autonomous driving*.’ But Catrin Misselhorn made a misapprehension as ‘we’ as human beings did not reach any conclusion on this dilemma.

Even more, studies across different cultural regions (Awad et al., 2018) revealed that there is no unique positions concerning fictitious alternatives of this stylised dilemma.

As any AI-based systems including ‘autonomous vehicles’ either have been ‘programmed’ by human programmers with pre-defined rules or ‘trained’ by the behavioural patterns of human actors in daily traffic, no such ‘statistical classifier’ can exceed the domain, which is specified by human beings.

As elaborated by Deborah G. Johnson (2006) in her seminal essay about ‘*Computer systems: Moral entities but not moral agents*’ [quote]:

> *Computer systems and other artifacts have intentionality, the intentionality put into them by the intentional acts of their designers.*

Whether one applies rule-based (‘symbolic-logical’) or data-driven (‘learning’) approaches, the computer code and the data-sets have to be provided by human programmers. Even more every objective – programmed rules, statistical classifications or reward policy in case of reinforced ‘learning’ – has to be defined by human decision-makers *ex-ante*.

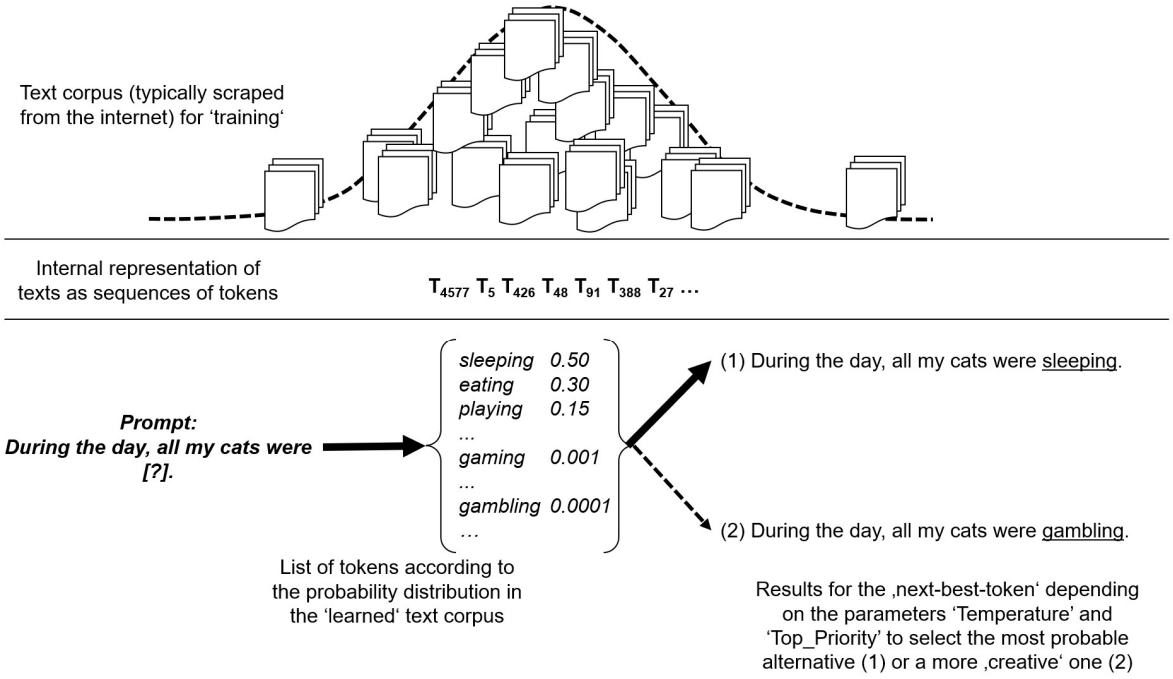


Figure 7.1: Simplified concept for the application of LLMs

7. GenAI and LLMs for Sequences and Next-best-Tokens

As explained in chapter 3, Shalev-Shwartz and Shai Ben-David (2014) pointed out '[quote]: *'The No-Free-Lunch theorem states that there is no universal learner.'* Likewise so-called Depp Learning with CNNs is adjusted to the statistical classification of images to estimate 'labels' (or for similar fixed length inputs), whereas Reinforced Learning is typically implemented to 'learn' winning strategies to find a next-best-move or a next-best-action in a 'game' with a set of pre-defined rules. However, there is no 'self-learning' as the labels or the rules+rewards have to be provided by human programmers/creators/trainers and these AI-systems always have a two-step approach of (i) 'learning' of training data or rules+rewards and (ii) implementation of the parametrized 'statistical classifier' for new events. A third application relates to sequences – either as a transformation of one sequence to another (Sequence-to-Sequence or S2S) such as language translation or as estimation of a 'next-best-token' to continue a sequence (or to 'generate' related text output, if the single step is repeated multiple times). In extension to the definition in chapter 3 of *training data* $\mathcal{S} = \{(x_1, y_1) \dots (x_n, y_n)\}$ with features *{x} and labels {y},* any 'learning' of sequences can be described as $\mathcal{S} = \{(x_{11}, \dots, x_{1n;}\, y_{11\dots}\, y_{1n}) \dots (x_{n1}, \dots, x_{nn;}\, y_{n1\dots}\, y_{nn})\}$ or $\mathcal{S} = \{(x_{11}, \dots, x_{1n}\, ;x_{1,n+1}) \dots (x_{n1}, \dots, x_{nn}\, ;x_{n,n+1})\}$ to 'learn' a masked next element.

One well-known example for sequence-to-sequence is the Rosetta Stone: a stone stele with versions of a decree issued in 196 BC during the Ptolemaic dynasty of Egypt in three languages: Egyptian in hieroglyphs, Egyptian using Demotic scripts, and, respectively, Ancient Greek. As the texts have the same content (i.e. the decree) with minor differences across the three language versions, the Rosetta Stone was the key to deciphering the Egyptian hieroglyphs starting with the statistical correlations between the three sequences.

This example of the Rosetta Stone reveals that statistical classifications in sequence-to-sequence processing is much older than any AI or digitization at all. Already in 1822 and with paper and pencil, it was possible to derive a first 'probabilistic' estimation how the three languages could be translated. Nevertheless, it required more 'data' – such as other multi-lingual inscriptions like the Decree of Alexandria, the Decree of Canopus, and the Memphis decree of Ptolemy IV - to read Ancient Egyptian texts confidently (sic!). The difference to GenAI and Large Language Models (LMMs) is (i) the scaling of the amount of data from few pages to text corpora scraped from the entirety internet and (ii) the algorithms to be trained on these tremendous amounts of data. Before turning to the AI systems such as 'Recurrent Neural Networks' (RNNs) and Encoders/Decoders or 'Transformers', it is important to emphasise that all these tools provide probabilistic outputs, i.e. making a statistical estimation about the 'best-sequence' as output. Likewise, this works for estimations of a 'next-best-token' to continue a text sequence. In this case, the last word of a sequence is 'masked' and the training data help to 'learn' the continuation of sequences with $h(x_1, \ldots, x_n) = x_{n+1}$ and recurrently to 'generate' longer text with estimations $h(x_1, \ldots, x_n, h(x_1, \ldots, x_n)) = x_{n+2}$ et cetera.

The more 'similar' sequences are contained in a text corpus, the 'narrower' the probability distribution will be, and the 'better' the estimation will match the reality: $h(x_1, \ldots, x_n, h(x_1, \ldots, x_n)) \approx f(x)$. Vice versa, the probability distribution for rare sequences will be flat, and the output will be more accidental. It is said that GenAI will '*hallucinate'*, this is an unfortunate anthropomorphism as every GenAI system provides statistical estimations – nothing more. If one asks a GenAI tool with the following prompt 'to write a two-side summary about the Rosetta Stone', the generated output will be very similar to the well-known history, because so many text from Wikipedia to books about history contain analogous sequences.

A general counterexample are curricula vitae (CV) of ordinary citizens, who are no celebrities. A self-test of the poetess Anja Utler (2024) revealed the problems of such probabilistic text generation, when she prompted for a brief CV and received the following result [quote]:

> *Deutsche Dichterin, geboren in Lüneburg, Universitätsabschluss in Lüneburg, lebt mit Mann und zwei Kindern in Lüneburg.*

Yet, she has never been to Lüneburg! The reality *f(x)* was not matched by h(x) due to the lack of data, but the 'best' probabilistic estimations, how to continue the prompt, was '*living in Lüneburg'.* Usually, readers will get similar results of 'probable' CVs.

As the development of GenAI[18] is very fast, the technology is much more advanced compared to the first ANNs, and there are recommendable books about this subject: e.g. the online guide *'Speech and Language Processing*' by Jurafsky and Martin (2024). Therefore, the technical development will be skipped here with only one remark: Different to CNNs, the key concept of RNNs is an internal memory so that the sequences (or subsequent time steps in time series) will be processed sequentially but the RNN keeps some memory about previous inputs. While this approach has computational challenges, more advanced developments like are so-called long short-term memory (LSTM, in the sense of a longer 'short-term memory', see especially Schmidhuber, 2015) solved the technical problems and have been implemented e.g. for signal-to-noise improvement in mobile phones and other signal processing applications very successfully. However, the memory processing is a technical limitation. The development[19] of Decoders/Encoders or 'Transformers' replaced the '*recurrence'* by another mechanism call '*attention'*, which allows to process sequences in parallel and find weighted relationships in the sequences about the correlation of words in a sequence (or sounds in music, pictures in videos like in OpenAI's Sora tool, or movements of point clouds in LiDAR as remarked for autonomous vehicles).

[18] Image or video generation with 'text-to-image' GenAI such as so-called denoising diffusion models will not be discussed in this paper (see e.g.: Cao et al., 2023).

[19] It is beyond the scope of this essay to review the interdependences between state-of-the-art AI architectures like 'Transformers' and dedicated hardware such as the recently introduces Blackwell Architecture of NVIDIA (2024). However, the following quote illustrates the sophisticated technological developments: 'The second-generation Transformer Engine uses custom Blackwell Tensor Core technology combined with NVIDIA® TensorRT™-LLM and NeMo™ Framework innovations to accelerate inference and training for large language models (LLMs) and Mixture-of-Experts (MoE) models.'

This ‘attention’ mechanism in current GenAI creates contextual correlations of a word by integrating information from surrounding text. In contrast to image recognition in CNNs based on a high-dimension ‘fit function’, the internal representation in GenAI is – rather simplified – a probability table[20] of all possible words, which could continue a sentence, generate a text, and mimic human creativity.

The methodology of GenAI for ‘*text-to-x*’ generation is illustrated in a shortened summary in Fig. 7.1. The first step is the text corpus – usually scraped from multiple sources on the internet – with an incredible number of ‘sequences’ in millions of documents. It cannot be assumed that these sequences are ‘iid’ (independent and identically distributed), as many documents such as messages or newsfeeds will be forwarded or ‘re-tweeded’. Typical ‘mainstream’ content will be very frequent and ‘sophisticated’ contend will be very rare or even unique - and both extremes could be stronger than in a Gassian distribution (for iid-events according to central limit theorem). To ‘digest’ all these text sequences, the sequences of words are converted to sequences of tokens, which can represent a word or parts of a words like stem, prefix and post-fix, providing a sequence of digital numbered elements[21].

These sequences of tokens are ‘learned’ – with masking following elements, applying ‘attention’ and summing up – to provide a probability table. If we input the prompt ➔ ‘During the day, all my cats were [?]’, the internal processing in the GenAI system will look-up the corresponding probability distribution for the ‘next-best-token’ with from 'sleeping' (token with a probability of e.g. 0.50 - just as an illustrative value) to 'gambling' (token with a probability of e.g. 0.0001). While in CNNs the best match is provided as feedback, GenAI systems have a set of additional parameters - especially ‘*Temperature’* and ‘*Top_Priority*’ - how the result should be estimated depending on the combination of these parameters: fully probabilistic[22] with always the top ranging (i.e. ‘*sleeping’* as most common alternative) or more ‘creatively’ with some random results (e.g. ‘*gambling’* in a more Lewis Carroll like style). But even with an external setting to achieve the most *probable* result, there is not ‘truth’ but only statistical probabilities.

[20] It is worth to note that this probability table represents the correlation of ‘tokens’ in sequences, but neither syntax (i.e. grammatical rules) nor semantics (i.e. a ‘meaning’ in the world of human beings).
[21] Again: neither syntax nor semantics, but statistical correlations between numbered tokens.
[22] Remark: depending on the actual implementation of a GenAI tool, there could be a residual randomness.

Although the output for a simple prompt (plus setting the parameters to ‘most probable’) like ‘*During the day, all my cats were [?]*’ is rather expectable and realistic, prompts for rare information like individual CVs could provide ‘probable’ results but without any factual knowledge (like ‘*Lüneburg*’).

Recent versions of GenAI tools have been ‘trained’ additionally with dedicated text corpora from examples of computer code (and computer language is a very simple structure compared to natural languages) to collections of mathematical calculations et cetera. Despite such additional ‘fine-tuning’, Yejin Choi (2023) explained in her keynote at a conference in Vancouver in July 2023 that OpenAI’s GenAI-tool *GPT-4* was not able to ‘solve’ a simple multiplication when prompted to multiply 999 times 876. Although LLMs show improved capabilities when scaled to some hundred billion parameters and trained with sequences of trillions of tokens, pure LLMs remain statistical classifiers to estimate a ‘next best token’. They can be complemented in a hybrid way with rule-based systems, i.e. one could attach an internal ‘pocket calculator’ to do the math correctly, but there are practical limitations. Cutting a long story short; there may be more ‘unexpected’ correlations in an incredibly large text corpus then expected, but nothing can ever ‘emerge’ beyond the text corpus and correlations within, as long as the text corpus is used for training of an AI-based ‘probability table’ for next-best-tokens.

However, there have been several reports about some ‘emerging’ capability of GenAI tools – for example capability to solve mathematical calculations – if the magnitude of parameters of such GenAI tools is increased over some threshold. But as Rylan Schaeffer et al. (2023) presented at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), such ‘spontaneous emergence’ depends on the choice of the metric. In other words, if one applies a step function switching from zero to one at some point, the ‘emergence’ is caused by the metric not the capabilities.

If is astonishing, what information (and correlations) are embedded in the text corpora on the internet: from textbooks to sample solution of exams to recent papers about ‘emergence’ in Gen AI tools (extending the next training of such tools in some self-fulfilling prophecy). So-called chatbots like OpenAI’s ChatGPT, Google’s Bard/Gemini, or Meta’s LLAMA 2-Chat are optimized front-ends for users’ interaction and gained tremendous acceptance after their launch recently.

With these simple user interfaces, it is possible to ‘generate’ a multitude of potential output based on rather simple ‘prompts’: from summaries of input text to whole essays about a given subject. Although they have trillions of ‘trained’ parameters, LMMs are sophisticated statistical representations of the text corpus, which was provided to ‘train’ the LLMs. Unfortunately, such text corpus is often just ‘scraped’ from the internet in a non-controlled manner: We select an unknown domain for a statistical classifier – and we wonder later how much was in this domain. It should also be remarked that LLMs are not general models of human languages, but statistical models of the text corpus used as input. Nonetheless, there is even a tendency towards some ‘LLM-ology’, what Large Language Modell are able to do (see: e.g. Trott, 2023), as if they would be exotic animals to be studied.

The more ‘we’ regard LLMs to be something ‘individual’ and different from a statistical classifier of a tremendous corpus of text input, the more we are at risk to anthropomorphism. The currently used terminology with ‘*hallucination’, ‘emergence’, ‘self-learning*’ is a clear indication. Vice versa, there is sometimes an incredible naivety how data is used. Of course, scraping the internet is a very cheap way to gather enough input for the ‘training’ of GenAI and LLMs. But what is in the text corpora? The current hype about LLMs tells us a lot about human expectations, but not very much about the limitations of GenAI tools.

As outlined by Léon Bottou and Bernhard Schölkopf (2023), there is no ‘truth’ in LLMs and no guarantee for ‘correctness’, as statistical classifiers provide a probabilistic (and not even syntactical or semantical) ‘next-best-token’ to a simple input and a nested ‘next-best-sequence-of-tokens’ to more advanced prompts. It is intriguing to prompt a LLM to generate a poem in Goethe-style based on few key words (in this case with a setting of ‘Temperature’ and ‘Top_P’ to provide a ‘creative’ output and avoid plagiarism) or a brief homework assignment about the development of payment system in Germany from 1960 to 2024. In both cases, the result will be probabilistic based on the distribution of text tokens in the input corpus.

Finally, the robotics company Figure released a short video on YouTube about ‘Figure Status Update - OpenAI Speech-to-Speech Reasoning’ (Figure, 2024). The development of ‘humanoid’ robots is far from new. One of the first commercially available example was Honda’s ASIMO (from 2000 to 2022) based on long-term development work since mid-1980s. Recent – and much more advanced – examples are Amazon’s humanoid robots designed to work in warehouses and to move baskets (see e.g. Bloomberg, 2024) or Tesla’s ‘Optimus - Gen 2’ Tesla (2023) – both presented end of 2023. Such robots are capable to recognize and seize fragile things like eggs and move them around on a table or put them into some receptacle. However, they have to be given orders in some computer-readable message format.

The innovation in the presentation by Figure in corporation with OpenAI, which was described as ‘*Speech-to-Speech Reasoning*’, is the replacement of the technical message by some ‘speech-to-command’ sequence-to-sequence processing. Building blocks such as ‘speech-to-text’, ‘text-to-text’, ‘text-to-speech’, or ‘text-to-code’ have been described in this chapter, and image recognition before. In this video the ‘prompts’ were two speech commands ‘*What do you see?*’ followed by ‘*Can I have something to eat?*’ (in the original video in slang). Like in prompt engineering, the first prompt sets the scene (as there was an ‘apple’ on the table before the robot), while the second prompt (not even ‘do xxx’ but a reversed questions ‘can I have xxx’) triggers a ‘next-best-order’, what the robot should do. This capability to apply LLMs - and in combination with image recognition and classification – to generate a machine-readable command is remarkable and exemplifies to potential of GenAI and LLMs to ‘translate’ speech or text into computer commands based on some kind of educated guesses for the ‘next-best-token’ – or in other words: based on statistical estimations.

Nonetheless, this video has its limitations. One can assume that it was the best possible demonstration and selected from a number of less successful versions (i.e. the presentation lacks any quality metrics about the reliability). The set-up was very simple and stylised (few objects on the table), and there was no ‘noise’ or disruption due to unexpected events or even corner cases. Speech-to-command can simplify the co-operation with robots or ‘co-bots’, but probabilistic command input creates new and currently unexplored risks.

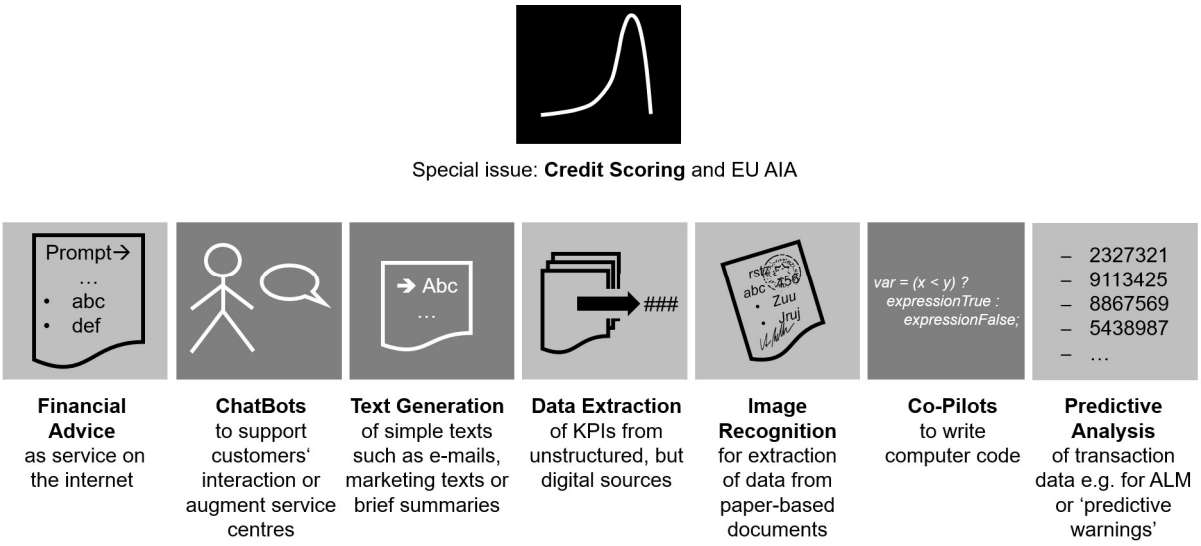


Figure 6.1: Illustrative examples for applications of AI in banking. As the generation of images or videos has lower relevance for banking, the focus will be on interaction with clients, text handling, data extraction, and transaction data.

8. Applications and Use Cases in Banking

The AI methodologies mentioned above are building blocks within a technical toolbox. The well-known examples[23] of detecting diseases from medical images, support for claims management in insurance, or predictive maintenance of machinery reveal the capabilities of AL-based systems but are hard to be applied to banking. Vice versa, use cases in banking may require a combination of methods to solve actual problems, which are distinct from simple pattern recognition, next-best-moves in games, or text generation based by LLMs with averaged probability.

It is far from trivial, to take all marketing messages, exaggerations and hope, or fears of AI into account and evaluate them concerning specified use cases in banking. And there are other concepts such as Robotic Process Automation, which are sometimes included in AI mistakenly, as RPA is an approach with *ex-ante* defined rules and non-traditional meta-programming to extract, copy or handle data between traditional applications. Figure 8.1 provides illustrative examples of use cases of AI in banking, from which a selection will be elaborated in the next chapters.

[23] And even these examples are not without challenges: For example, the development of a predictive maintenance application may take two years plus x, and for recommendation engines it may be sufficient to propose 5 'next-best-actions', while only one will be seen as helpful by consumers.

9. Data Extraction and Prompt Engineering

Since the roll-out of easy-to-use ChatBots like OpenAI's ChatGPT, Google's Bard/Gemini, or Meta's LLAMA 2-Chat, a new method of 'prompt engineering' appeared. Instead of understanding (and controlling) the input distribution to these statistical classifiers, this approach takes the LLMs for given and attempts to 'engineer' the input prompts to the tools to achieve more and more 'correct' results. Such approaches have the danger of circular reasoning, as 'we' as human users 'enhance' the prompts towards desired results. We are going to 'prompt' for answers, which we already have to know to 'engineer' the prompt to derive the answer.

Examples for more and more 'enhanced' prompt can be found in the Prompt Engineering Guide (2024). However, the more 'engineering we provide, the more we are asking a math text problem with the solution path already included in the prompt. Examples with different 'enhancements' are e.g. (and the notations are always termini technici):

- Zero-shot prompting: 'Do … to continue the following sequence.'
- Few-shot-prompting: 'Do… according to the following examples …'
- Self-Consistency or Generated Knowledge Prompting: 'Do … with a longer list of examples how to solve such text problem plus a question.'
- Chain-of-Thoughts and Tree of Thoughts: 'Do …with a list of consecutive examples with explanations.'
- Proposal for 'Self-Discover' (Zhou et al., 2024), which is an extended prompt with a predefined way to answer incl. examples, how to do so.
- Retrieval Augmented Generation (RAG): First, retrieve a result from a traditional query to an internal database, and second used the results as prompt to a LLM to make a summary et cetera.
- Iterated prompting: Include the LLM into an (external) application similar to a sub-routine to process text according to computed prompts and iterate this prompting until a 'correct' result is generated.
- et cetera …

Usually, these proposals to 'engineer' prompting compare the results to external benchmarks. However, they do not discuss that the result is partly – and more and more – provided *ex-ante,* and the LLMs is 'tuned' towards a specific internal classification, we already know more or less before.

System: You are a helpful climate and green finance risk analyst.

Human: Based only on the excerpts from the corporate reports provided below, we would like to extract the KPI information for the year {year}. Specifically, we are looking for the following:
- The numerical value and unit of "{KPI}".
- A short comment explaining how this value was obtained.
- The sources used, including the page number and a short quotation from that page.
- An indicator of certainty that ranges from 100 (absolutely certain) to 0 (cannot be determined based on available information).

If the information for the KPI cannot be found or determined from the provided documents, please generate a JSON object with the appropriate fields set to 'null' and include a comment stating that the information was not available.

Please note that the output will be a JSON object, structured according to a predefined schema: {format_instructions}

Definition of'{KPI}': {KPI_definition}

Report Excerpts: {reference_docs}

If no relevant information is found in the provided excerpts, the output should clearly reflect this with a 'null' value or an appropriate indication of the absence of data. The 'certainty' field should reflect the level of certainty of the information provided, including a value of '0' if the KPI could not be determined based on available information.

Table 8.1: Example of the 'Gaia LLM prompt template' (quote from BIS, 2024) for a formalized prompt to extract climate change-related KPIs from company reports. According to BIS (2024): 'the variables "{…}" are filled with the necessary information for a given KPI. Colours correspond to listing in the text: black – general instructions, red – definition of the output format, orange – definition of the information to be supplied in the response, green – pages from the report.'
(Remark: The colours are taken from the original reference.)

An impressive example is the test to extract climate change-related KPIs from unstructured company reports conducted by BIS Innovation Hub in the ‘Project GAIA’ (BIS, 2024). Due to the lack of global reporting standards for climate change-related indicators (or ESG indicators in general), financial institutions have the challenge to extract these KPIs manually from corporate reports among other financial and non-financial information with heterogeneity of naming conventions and definitions. To achieve automated data extraction of KPI, Gaia combines semantic search (to select the most relevant pages from long documents) with prompt engineering (to extract KPIs and references plus output generation in the format of ‘JSON’ objects). The LLM was OpenAI’s GPT-4 in a Microsoft Azure cloud.

This approach to integrate a LMM is much more resource intensive compared to the usage of a ChatBot as a simple user interface. Two quotes from BIS (2024) illustrate the effort including the design choices (DC) and the iterative development [quotes, underlining by the author]:

> *KPIs such as “gross direct GHG emissions (Scope 1)” or “total energy consumption” have a clear definition in the field of green finance but cannot be expected to be understood within the general training scope of the LLMs. The first step is to create a succinct definition of the KPI within the context of the field of green finance (DC1). ... The LLM is instructed to “Write a concise five to seven sentence definition for [KPI]” based on relevant pages from the standards and legislature documents, which are passed as context. The generated KPI definition becomes a cornerstone in the semantic search (DC4) and is also reused in the subsequent LLM prompt (DC6). …*
>
> *At one stage of experimentation, the manual cross-checking of results revealed a notable anomaly: a significant portion of extracted KPI values consistently showed the same figure, namely 1,500,000 metric tonnes of CO2 equivalent (t CO2 eq). Upon closer inspection ... The LLM had generated fictional reports, complete with quotations ... All of this was from imaginary sources that convincingly mimicked actual reported sentences. This peculiar behaviour was not a one-time occurrence. The LLM consistently followed the same pattern in multiple experiments, always returning a KPI value of 1,500,000 t CO2 eq and referencing the same fictional reports.*

This recurring pattern of hallucination was mitigated by a number of design choices, in particular (DC8) choosing the latest LLM version, GPT4, which adopts a more conservative approach instead of providing potentially incorrect values as compared with its predecessor; (DC6) prompt engineering, instructing the LLM only to refer to information given within the provided context and explicitly stating what to respond if no information is found; and (DC10) setting the temperature parameter to zero, which controls the creativity of LLM. With these design choices, hallucinations were significantly reduced, and they do not seem to impact results in the final version of the PoC.

These experiences reveal that the integration of an LMM (i) resembles traditional data analysis work (e.g. using script programming to extract data), (ii) demands a lot of effort and iterations, and (iii) requires an in-depth understanding of the concept of LMMs. While LMMs are a new alternative to handle unstructured information in natural language, they are rule-based data extraction methods. Consequently, these deficits have to be compensated with 'prompt engineering', which attempts to 'control' the LLM with *ex-ante* human knowledge.

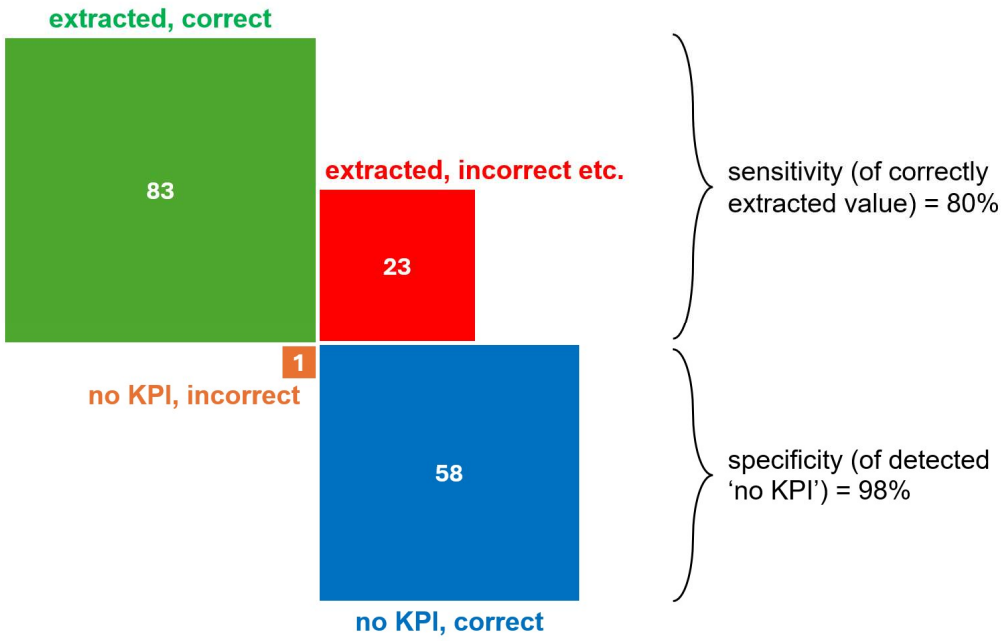


Figure 9.2: Quality test with manual cross-checking concerning the results of the GAIA project to extract KPIs from 163 original sources (data taken from BIS, 2024)

Is it worth to invest all this effort and what is the quality of the results? The results of a statistical quality test with manual cross-checking taken from BIS (2024) are shown in Fig. 7.2. Whereas the specificity to detect input documents without the (expected) KPIs is rather high (98%), the sensitivity to extract KPIs correctly is moderate (80%). The findings of Project GAIA show that it is feasible to implement LMMs in a wider approach to extract data from unstructured textual sources. However, it depends on a dedicated risk analysis and appropriate cost-benefit analysis whether this approach is reasonable in a specific context. For example, in the case of a macroeconomic analysis, 80% sensitivity might be better than nothing when no structured data are available and huge amounts of documents have to be scanned for information, which is not standardized. But in the case of a decision for a ‘green’ lending according to ESG standards, such an approach with the risk of either fictional or incorrect data extraction might be unacceptable – at least today with the current version of LMMs and GenAI tools.

Figure 10.1: Comparison of ChatBots at three times (plus one 'engineered' prompt) with truncated outputs for the prompt: "Construct a balanced portfolio of ten securities for a risk-averse investor of age 40!" (see also: Milkau, 2023 / 2024)

10. ChatBots and Financial Advice

With the following prompt as a litmus test, three publicly available ChatBots – OpenAI's ChatGPT (with different versions of GPT-3/GPT-4 as engine), Google's Bard or Gemini, and Meta's Llama-2-Chat (accessed via https://llama2.ai/ with default parameters) – were examined at different times: *Prompt ➔ Construct a balanced portfolio of ten securities for a risk-averse investor of age 40!*

The result shown in Fig. 19.1 is astonishing in more than one way. First, the ChatBots provide quite convincing results at first glance, although they were not fine-tuned for asset management or financial advice. Nevertheless, there is tremendous content from portfolio theory via published ranking lists for Exchange Traded Funds (ETFs) to specific product information contained in the text corpus used for training the different LLMs. Respectively, the results can be regarded as an 'average' over the trained texts: for example, a Google-search for 'best index funds in the U.S.' or 'best international ETFs' in June 2023 would give similar results.

Second, some obvious errors occurred – at least in the early versions in mid-2023: Sometimes the percent value do not sum-up to 100% or Amazon or Google are described as 'Mid-Caps'. Likewise, it remains opaque, why the results were strongly focussed on products of two providers: Vanguard and iShares. Additionally, it is unclear, why ChatGPT and Bard/Gemini switched from existing products to general portfolio structure, while Llama-2-Chat switched in the opposite direction.

And third, taking an outside-in perspective of an ordinary users of internet services (search engine today and potentially ChatBots in the future as entry point to the internet, but without in-depth understanding of the technical assumptions, features and limitations), the results are not fully wrong, but a mixture of arbitrary results and a tendency to an average. Once again: there is no truth in LMMs, but only probabilistic 'best-next-tokens' depending on the text corpus used for training. Perhaps, the result of the ChatBots can be regarded as an automated finance journalism, which replicates some 'average' of existing clichés.

Contrary to these 'average' outcomes, an online survey with some eight thousand consumers in 13 industrial states conducted by Capgemini (2023) revealed that 67 percent of these consumers believed medical opinions from generative AI would be helpful, and 53 percent would trust generative-AI-assisted financial planning. Normally, online surveys are biased towards online-affine consumers, but this group matches the target group of online brokers or providers of so-called 'robo advice' (i.e. automated rule-based recommendations). Therefore, a future trend to 'trust' GenAI and ChatBots for financial advice is very plausible.

An additional concern is the inclination of such ChatBots to either support the human believes as expressed in the input prompts or even to apologize for diverging output, which is questioned by humans during a chat with multiple prompts (see e.g. Bauer, 2024). Such a 'polite' behaviour is not a generic feature of LLMs, but usually a consequence of post-processing to 'fine-tune' LMMs according to pre-defined guidelines (see once again chapter 4 concerning "*Sed quis custodiet ipsos custodes?"*). This kind of 'adaptive' behaviour of LMMs as – fundamentally – statistical classifiers is an old issue since the first (rule-based) chatbot ELIZA created by Joseph Weizenbaum (1966; see also Mitchell, 2023).

It is obvious that 'trained' LMMs can never generate any 'alpha' performance beyond following market development. As LLMs have to be 'trained' is a very costly[24] and resource consuming process, the text corpus is always outdated[25] by weeks or months[26]. Even if a LLM would be updated in real-time, it could not escape the 'trend to the average' or in other words: just follow a mainstream development[27]. And finally, proposals to use a corpus of financial time series (i.e. a sequence of stock prices) to train a GenAI model to generate a 'next-best-price' is nothing more than an advanced technical chart analysis for 'trend trading', which potentially could reveal some herd behaviour of traders but cannot 'generate' any alpha in an efficient market.

[24] The recent 'Artificial Intelligence Index Report 2024' (Perrault and Clark, 2024) provided the following cost estimations [quote]: '*For example, OpenAI's GPT-4 used an estimated $78 million worth of compute to train, while Google's Gemini Ultra cost $191 million for compute*.

[25] This has to be distinguished from Retrieval Augmented Generation (RAG), in which the basic information is retrieved from relevant sources (e.g. newsfeeds) and included in prompts to LLMs to make final grammatical copyediting.

[26] For example, GPT-3 included data until autumn 2021, GPT-4 until April 2023, and GPT-4 Turbo until end of 2023.

[27] As Waber & Fast (2024) and Shumailov et al. (2024) remarked, there is an increasing danger of a 'model collapse', because the more texts are generated by GenAI/LLMs the more of GenAI-generated texts will be included in the text corpus used for training of those GenAI/LLMs. This self-enforcing process leads to a narrowing of probability distribution (see e.g.: Alemohammad et al., 2024) and amplifies the 'trend to the mean value'. Additionally, it has been proposed to use 'synthetic data' (generated by GenAI) to train other GenAI, but this is like a *perpetuum mobile* for GenAI.

11. Document Handling in Trade Finance

Since decades, a major challenge in trade finance has been upgrading documentary trade process from paper-based transactions to digital processes. A first attempt in electronic shipping commerce to digitize the traditional paper-based bills of lading was the ‘Bolero’ project launched 1999 as a joint venture between SWIFT and the TT Club (Through Transport Mutual Insurance Association Ltd.) and partly initiated / funded by the European Commission. Maybe the latest attempt was the global trade platform ‘TradeLens’ jointly developed by IBM and GTD Solution, a division of A.P. Moller - Maersk as one of the largest international logistics companies and container carriers. The vision behind TradeLens was to digitize the global supply chain on a central industry platform. The sheer amount of paper accompanying a container demands that carriers like Maersk have to run large back-office operations to process the documents and extract digital data, which could be provided on central platform in an end-to-end approach. However, TradeLens was not successful, did not reached the necessary level of commercial viability, and went offline in spring 2023 (Maersk, 2023). One problem is caused by the different juridical systems and still existing requirements that trade documents have to be signed by hand to be legally binding (excluding fully digital documents with electronic signature).

Given the present need to keep paper-based documents, one could ask for the actual problem, because digitization of documents is an established method – especially for structured documents with standardized data fields such as bills of lading. These methods range from data extraction from pdf-documents similar to bills uploaded into online banking (with a de-facto standard of the structure and key elements such as ‘IBAN’ followed by bank account number or customer id followed by a number et cetera). In Nordic countries, an additional bar code on all bills has been common since decades. Also handwriting recognition has been developed over the years from simple signature on cheques to longer texts in foreign languages especially with RNNs (see e.g. Graves and Schmidhuber, 2008). Nevertheless, a general problem in trade finance are documents with handwritten remarks across the print or official seals hiding the document text, dirty / crumpled / marred documents, scanned / faxed documents with bad contrast and many other derivations from a ‘clean’ document.

Therefore, a step-by-step approach is required for: recognition of structure / defects – recognition of non-standard / handwritten elements – recognition of digital values and data extraction. Whereas standard Optical Character Recognition (OCR) and text processing / data extraction work well in case of ‘clean’ documents, an unstructured and ‘dirty’ image in a pixel format (like scanned or faxed documents) with bad signal-to-noise ratio is a challenge.

This challenge is the substitution of cognitive capabilities, we humans apply unconsciously when we ‘decipher’ such documents, by image recognition based on large set of training data with ‘dirty’ images and corresponding data captured by human experts (taken from archive and back-end systems). Only large international logistic companies or trade financing financial institutions will have this large corpus of already ‘labelled’ archived documents to train a ‘image-to-data’ recognition system based on ANN.

Whether a direct ‘image-to-data’ approach or a step-by-step approach would be feasible depends on a number of specific factors:

- Available training data
- Cost for development and training of the AI-based tool
- Quality in actual operations (i.e. ‘False’ recognitions) and cost for corrections
- Cost savings due to substitution versus manual processing of residual documents
- Level of ambition (as e.g. automation of only the ‘clean’ documents could a significant relief for the back-office operation)

Although handling images with AI-based tools seems to be no rocked science today, the specific requirements in trade finance operations are a good example for the real issues, which have to be solved.

12. Transaction Data, the Case of ALM and Economic Trade-off

The last use case, which is selected for this essay, is anti-money laundering (AML) as an example for predictive analysis based on payment transaction data. The idea to detect patterns in payment transactions is nothing new but was often restricted e.g. to monitor re-payments of credit card balances or mortgages of one customer – taking into account certain dependencies like delayed payments in the USA, where customer make payments manually and not as standing order. In this case, an unusually late payment could indicate an increased probability for a future default, which can be applied for a simple rule-based approach such as linear regression with data in one siloed application.

In contrast, money laundering exploits the network structure of the global financial system, information asymmetries due to limited data in the banks' silos, and typical schemes use multiple cross-border payment transactions. Vice versa, effective AML requires a holistic perspective and network analysis of transaction data beyond individual bank silos or national borders. The 'Project Aurora' conducted by the BIS Innovation Hub BIS (2023) evaluated different state-of-the-art approach including ANNs and GNNs (see chapter 2) from three perspectives: individual bank, domestic network, and international network. The project used a controlled environment with synthetic data to avoid the problem that real-world data often lacks either a part of suspicious chains of transactions or were not able to identify and label the suspicious transactions. Further details of AML will be skipped for this discussion and the reader is referred to the project report (BIS, 2023).

The evaluation compared four advanced approaches compared to a simple rule-based model: two traditional statistical approaches and two ANN-based ones. The traditional statistical approaches were well-known Logistic Regression (LR; to estimate the probability of money laundering events based on the input data) and Isolation Forest (IF; developed in 2008 for data anomaly detection by recursive generation of partitions and randomly selecting a split value between the minimum and maximum values of the attributes). The AI-based approaches were usual ANNs and GNNs (using the structure of the network as a graph to estimate money laundering events). Figure 12.1 summarizes the results of this evaluations as provided in BIS (2023). Other issues of the study will be skipped here (e.g. the question where to collect data in real-world payments networks: centrally or e.g. with 'federated learning').

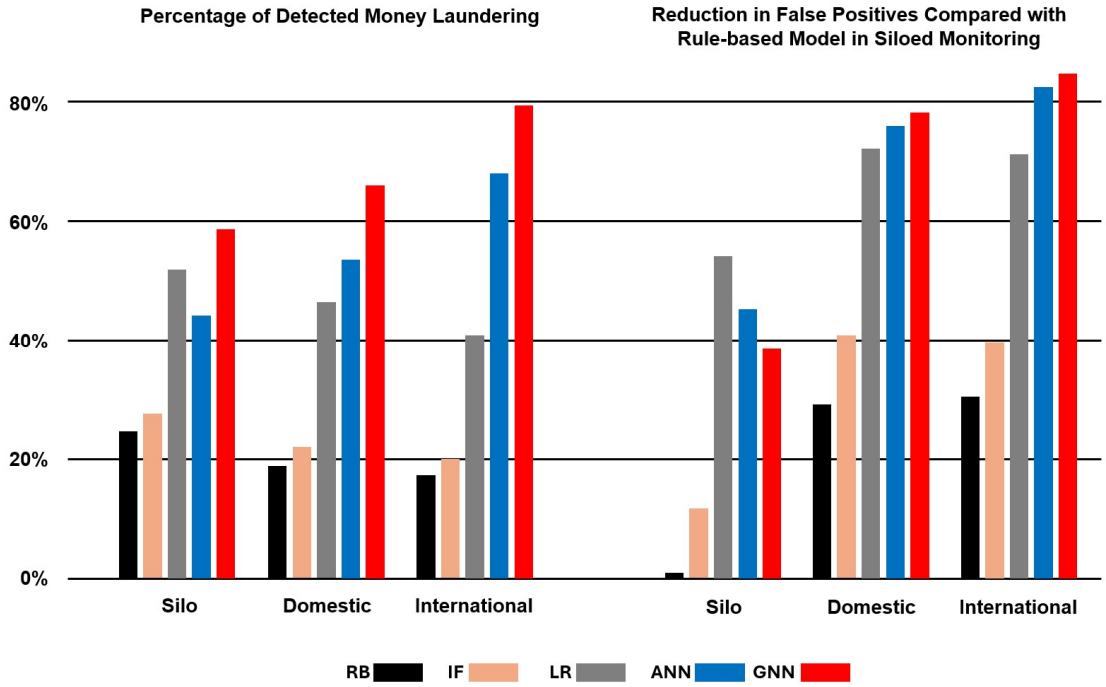


Figure 12.1: Comparison the tested methods for the three scenarios with individual Silos, Domestic systems, and International Networks (Data from BIS, 2024); RB: Rule-based Models, IF: Isolated Forest[28], LR: Logistic Regression, ANN: Artificial Neutral Network and GNN: Graph Neural Network

Beside valuable insight into new way to ALM, these results expose a number of issues of the implementation of AI in banking:

- The difference between siloed data and whole networks shows the importance of alignment between the objective, the data, and the method. While siloed data in banks' internal systems have restricted value, simpler approaches such as LR may have advantages if data are limited.
- As in reality individual banks not allowed to share transaction data, problems may be caused by legislation but not technology.
- While LR has moderate results, but ANN or GNN have better results, the issue of (i) training data and (ii) real-time data may be more important.
- As GNNs are matching the original problem (mapping the network rules to a graph-based approach; sometimes called 'neuro-symbolic AI'), this hybrid approach provides best results, but requires in-depth understanding of various sophisticated technical models.

[28] See also other reports e.g. db (2024) for a similar approach 'Black Forrest'.

This example points out that in a three-dimensional space of availability of data versus sophistication of method versus experience ‘simpler’ methods could be more practical than sophisticated ones.

Additional questions arise – not so much for ALM, but e.g. for fraud detection in credit card transactions – when detections quality is evaluated together with commercial costs. If for example the False Positive predictions are too frequent (due to a high threshold to reduce the False Negative ratio), the commercial effect to lose annoyed customers - and the associated profit - could ask for a trade-off between statistical prediction quality and commercial costs. In some cases, it could be recommendable to ‘accept’ a lower prediction quality and compensate customers for infrequent external fraud *ex-post* than to lose the customers at all.

Today, the debate about AI in banking is rather focussed on technologies. In practise, any implementation has to start from objectives and align the implementation to economic trade-offs:

Objectives (problems to be solved) and human value (as guidelines)

↓

Available data and available experience / resources[29]

↓

Selection of technical methods or combination of methods

↑

Statistical quality measures and commercial impact

↑

Economic trade-offs

Especially, the issue of ‘available data’ includes the crucial problem of ‘negative’ data points. Banks have huge amounts of transaction data but comparable few(er) documented cases of money laundering, transaction for tax evasion, external fraud (due to phishing et cetera) or even internal fraud. And as shown before in this chapter, they may see only parts of chains of transactions (also compared to international credit card networks). While it is possible to use AI to give ‘predictive warning’ for a corporate customer entering into financial problems and delaying payment patterns, huge amounts of transaction without context could lack statistical significance.

[29] Including the costs, if for example a bank wants to train a GenAI model on premise.

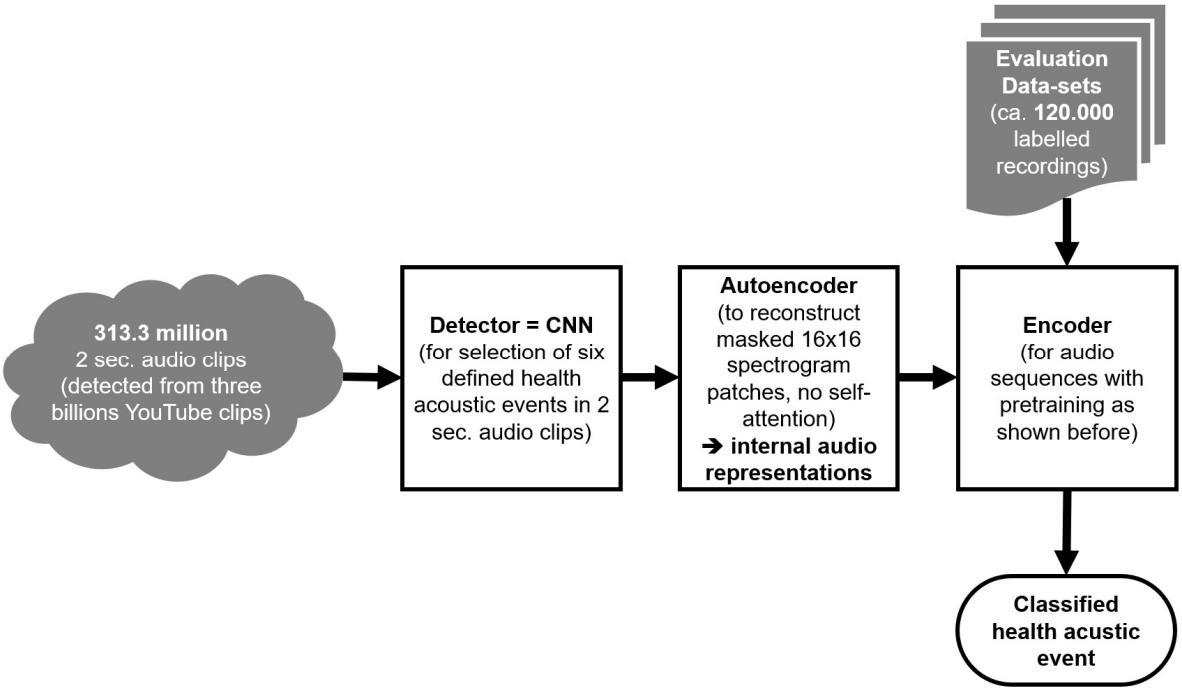


Figure 13.1: Example from health care research about a HeAR system in a combination of CNNs (for preselection of sequences), unsupervised Autoencoders (for constructing internal representations) and supervised Encoders with labelled data (Baur et al., 2014). HeAR scored 0.645 and 0.710 for COVID-19 detection and 0.739 for tuberculosis, although only a rather small set of labelled data was used.

13. A Remark about Autoencoders for Transaction Detection

The problem to have huge amounts of 'non-labelled' data but few data with a diagnosis is also prevalent in health care research. Recently, a team led by Google scientists (Baur et al., 2014) developed an innovative multi-step AL system to bridge this gap. The specific idea to use simple sound sequences as a biomarker for respiratory diseases is not new and could be very helpful for mass screening for Covid-19 or tuberculosis. As explained by Baur et al. (2014), the objective is [quote]: '*Health-related acoustic cues, originating from the respiratory system's airflow, including sounds like coughs and breathing patterns can be harnessed for health monitoring purpose*s.'

As an experiment to bridge the gap between millions of available acoustic sequences (on YouTube, sic!) and fewer combinations of a sound sequence and a diagnosis, the HeAR system is a combination of a CNNs (for preselection of sequences: coughing, baby coughing, breathing, throat clearing, laughing, and speaking), an unsupervised Autoencoders for constructing internal representations of sequences by correlating them into ‘clusters’, and a supervised Encoders with labelled data taken from medical sound data-sets (see also Fig. 2.2 b and d). For the technical details the reader is referred to the original paper.

The reported performance on cough inference tasks (0.645 and 0.710 for COVID-19 detection and 0.739 for tuberculosis, see Baur et al., 2014) are rather promising for such an early-stage experiment and indicate potential for future real-world applications.

Even more, this concept to combine huge amounts of unlabelled ‘transaction’ data with fewer labelled data could be worth to be tried with financial transactions as discussed in the previous chapter 12.

Yet, this is a very sophisticated and multi-step approach with different types of AI (i.e. CNNs, unsupervised Autoencoders, and supervised Encoders), which illustrates how much expertise with different AL-based approaches is required to ‘squeeze out’ the transaction data in an innovative way to get new insight.

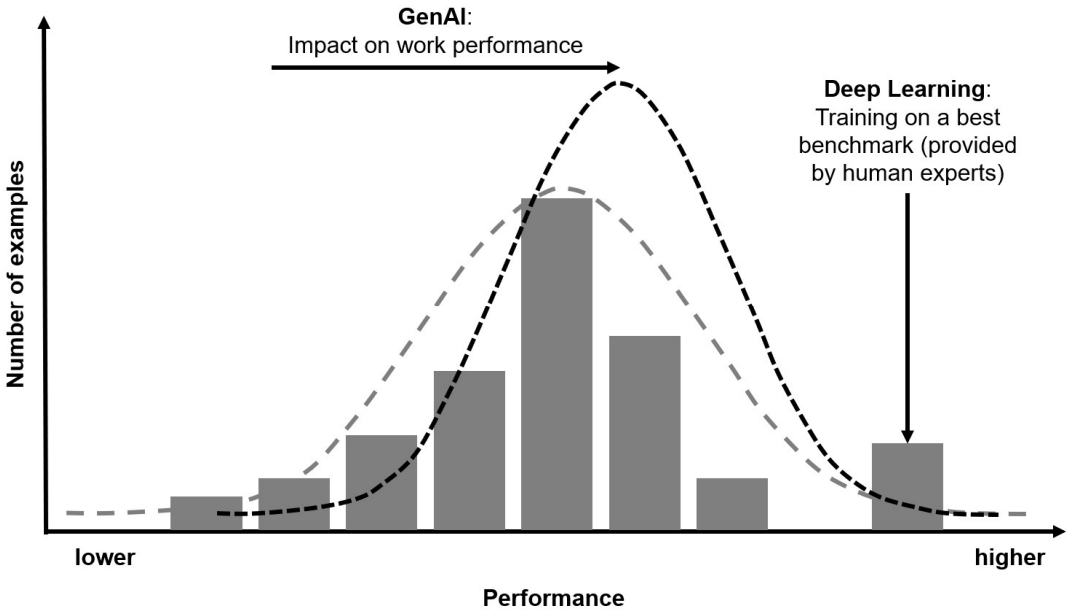


Figure 14.1: Impact of GenAI on performance or labour productivity with a 'tendency to the statistical mean' compared with Deep Learning (trained on a best benchmark)

14. Productivity, Augmentation and Performance

Since the public introduction of ChatBot like ChatGPT end of 2022, an enormous hype about the possible impact on human employment dominated the public discourse. A study by the consulting firm BCG (Bellefonds et al., 2023) claimed that 50% of the time in call centre operations could be 'optimizable'. And McKinsey & Company predicted that GenAI could add additional $200 billion to $340 billion of the industry's annual revenues with three major use cases: assistance for frontline employees, code development, and generation of marketing content (Chui et al., 2023).

However, there are few quantitative studies and there are different measures:

- Statistical measures of quality (like sensitivity or specificity – especially in image recognition; see chapter 3)
- Commercial benchmarks (weighting the statistical quality versus economic trade-offs; see last chapter)
- Performance or labour productivity (see Figure 14.1)

One of the few quantitative studies, published early in 2023, was conducted by Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond (2023) on the impact of 'GAI assistants' on labour productivity.

This study found a 14% productivity increase on average for customer support / call centre agents augmented by GenAI measured by issues resolved per hour, but with a spread according to the skill level (see list of references in Milkau, 2023). Less-skilled / inexperienced workers were enabled to resolve around 34% more issues per hour, while the impact on experienced / highly skilled workers was very limited. According to the study [quote]: '*the greatest impact on novice and low-skilled workers, and minimal impact on experienced and highly skilled workers*.' This does not exclude that many simple customer questions can be handled automatically before any human agent is needed. First level solutions from interactive phone systems to rule-based text-chatbots can solve the magnitude of customer calls like '*I need a new password*', '*I changed my address'*, or '*What is my current account balance'*. But for the second level - when human agents are needed - the study is consistent with two other findings that GenAI can augment lower performance within a certain job profile but has limited impact on skilled staff.

Another study about the impact of AI on taxi drivers by Kanazawa et al. (2022) examined the productivity gain by the use of an AI-based 'taxi driver assistant', which estimates best routes with a high demand. The 'AI assistant' reduced the time spent on cruising by 5.1% using the full sample, but with all gains concentrated on low-skilled drivers narrowing the productivity gap between them and high-skilled taxi drivers. And a study by Shakked Noy and Whitney Zhang (2023) with an online experiment, which exposed preregistered college-educated professionals randomly to ChatGPT, found [quote]: '*The generative writing tool increased the output quality of low ability workers and reduced time spent on tasks for workers of all ability levels*.'

As a typical use case for GenAI is software programming, Sayan Chatterjee et al. (2024) evaluated '*The Impact of AI Tool on Engineering at ANZ Bank'*. They compared two groups of developers with mixed experience and aligned tasks with and without support by GitHub Copilot, which is a code completion tool developed by GitHub and OpenAI. Although the composition of the two groups was not normalized and juniors had simpler tasks to complete compared to senior programmers, the main results are shown in Fig. 14.2. Similar to Fig. 14.1, a shift of the distribution is apparent with a strong time reduction for longer tasks compared to shorter ones.

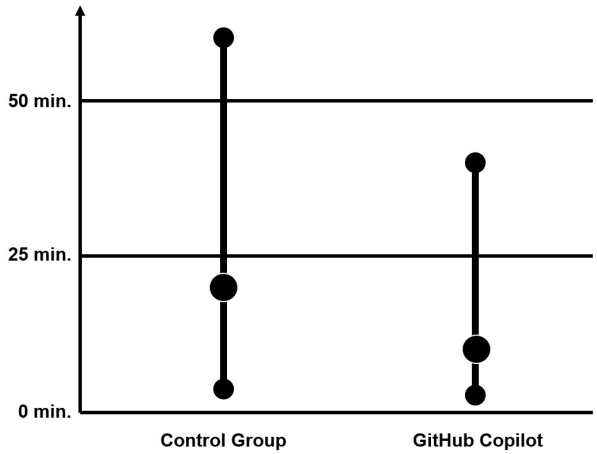


Figure 14.2: Comparison of total time spend of two groups of software developers with and without GitHub Copilot according to Chatterjee et al. (2024). As 200 developers were included in the study, 172 could solve the problems and 9 results were clear outliers, 163 developers were compared in the figure.

Although Chatterjee et al. (2024) did not compare their result with a benchmark like some Computer-aided Software Engineering (CASE, as already started in the late 1960s and peaking in the 1990s but declining with the transition from mainframe to client/server architectures) or some graphical code assistant system to browse software libraries. Because much 'software engineering' is rather basic programming work, this work can be optimized with more re-use of standard software building blocks. Consequently, all kind of 'co-pilots', which support 'best practises', can increase the 'average quality' and/or reduce time spend for programming lines of code again and again. As computer code[30] is written in a very formal language, dedicated or finetuned LLM are a new return of CASE.

As illustrated in Fig. 14.1, LLM are statistical tools to find 'next-best-tokens' or 'next-best-lines-of-code' based on a probabilistic approach of extremely large corpora of texts or computer code. The outcome is always a 'best average', which has a strong impact on lower performance (augmentation with a shift to the mean value) but limited impact on top performance. Vice versa, Deep Learning with 'labels' provided by human experts can emulate this expertise and substitute resources, where experts are missing or needed for other tasks.

[30] Especially for data handling and analysis from Excel via Python to SAP data (with SAP Joule) such 'co-pilots' can write queries, but the user – still – has to understand the data and the question to ask.

Any use case depends on the specific circumstances. For example, developing countries with a weak or even missing health care infrastructure have much to gain from introducing fully automated, AI-powered and portable medical device for diagnosis. However, such devices do not solve a number of questions: Who will pay for the technology? Who will maintain and transport the devices? What do the patients do with a best-of-class diagnosis, when there are no doctors to apply the recommended therapy. And maybe most important: Is there a generalizability and transportability of AI-based devices, which were trained on patients' data from industrial states (and perhaps in best-of-class hospitals), to target populations in these developing countries like in Africa?

Other questions could be asked in industrial states concerning the economic structure of health systems, health insurance and professional: What is the commercial benefit and what are the incentives[31] to deploy costly AI-based systems. According to the Economist (2024) [quote]: "*The world could lack 10m health-care workers by 2030, around 15% of today's workforce. And administration accounted for about 30% of America's excess health-care costs, compared with other countries, in 2022*."

As the demographic development causes more and more vacant jobs in Europe (and especially in health care et cetera), AI-based tools to either augment or complement human work will be helpful to compensate for missing resources. In the end, it will be a trade-off between missing resources, additional (current) costs and reduced (future) costs, where and when such tools will be implemented. This is a general question, which applies in adapted forms to all industries including financial services and banking.

[31] As discussed in Obermeyer et al. (2019), the U.S. health system and especially insurers rely on an established industry-wide algorithms (rules-based, i.e. not AI!) to predict patients with costly future needs based on costs in the past and enrol them to 'high-risk care management programs', to reduce future costs by additional support today. It has to be understood (see Milkau, 2021) that this is a purely commercial objective (to reduce more costs in the future), but not an attempt to improve patients' health primarily. As a collateral effect, this widely used algorithm does not include patients, who do not or are not able to use the health care system with insurance coverage. Therefore, poorer patients, who cannot spend time to see a doctor or do not have a protection in case of a sick note and will not use the health care system are not supported – simply for commercial reasons. As – in a second step – minorities in the USA are poorer in average, there is an – *ex-post* – bias towards less support to these minorities, which is attributed by Obermeyer et al. (2019) as a '*significant racial bias'*, but clearly without any intention to discriminate while searching for economic advantages in an unequal society. Of cause, it is easier to blame an 'algorithm' to show racial bias then to address the structural problems of a whole society.

15. Domain-specific Models and Copyright Questions

Every statistical classifier is ‘domain-specific’, as such a classifier always depends on the domain of the training data (see chapter 3). Therefore, the term ‘domain-specific models’ is a *contradictio in adiecto* in a wider sense. In a narrower sense, the term can be applied to ‘Small Language Models’ with training data reduced to a specific subject matter domain such as banking or public services.

An illustrative example is the text assistant ‘F13’ for staff in public administration based on the Aleph Alpha’s Luminous GAI (StM.BW, 2023) developed by the German federal state of Baden-Württemberg based on the Aleph Alpha’s Luminous. This GenAI tool for public administration provides basic functionality like (i) summaries of text inputs except for confidential or personal data, (ii) generation of (short) notes from stored cabinet bills, and additionally (iii) research in a knowledge base of information for public services. Given that in Germany nearly one million vacancies in public administration are predicted by the consulting firm McKinsey & Company for 2030 this approach can help to relieve staff from rather ‘mechanic’ text writing, i.e. it provides augmentation instead of substitution.

Domain-specific ‘Small’ Language Models are still probabilistic approaches. Even for summaries or comparisons of texts it cannot be excluded that a ‘next-best-token’ might be the ‘most probable’ one but incorrect in the specific context. Training with internal / proprietary data provides more control about the text corpus, but when one prompts for an unusual answer or a rare event (without prompt engineering, which ‘prescribes’ the result based on *ex-ante* knowledge), the output does not necessarily be ‘the truth’.

Even more, probabilistic approaches (with LLMs, SLMs or similar) cannot be used to process citizen’s applications[32] e.g. for social benefits or income tax returns, which have to be decided – quite literally - according to rules of applicable law (taking into account that tax legislation itself can be inconsistent or incomprehensible).

Similar projects[33] for a financial services context were published in late 2023: Shijie Wu et al. (2023) proposed ‘*BloombergGPT: A Large Language Model for*

[32] This is different, of course, to credit scoring as discussed in the next chapter: While credit scoring is a statistical estimation of future default probabilities *ex-ante*, applications of social benefits or tax returns are *ex-post* claims of citizens based on a current status or past income.
[33] A list of dedicated LLMs developed for the financial sector can be found in Maple et al. (2024).

Finance' and Xianzhi Li et al. (2023) asked '*Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics?*'. Both tools focussed on 'sentiment analysis' from news feeds to deduce investment actions. BloombergGPT is a 50 billion parameter LLM that is trained on a 363 billion token dataset based on Bloomberg's news sources plus 345 billion tokens from general purpose datasets. According to the study [quote]: '*Take the example of sentiment analysis, where a headline such as "COMPANY to cut 10,000 jobs" portrays negative sentiment in the general sense but can at times be considered positive for financial sentiment towards COMPANY, as it might result in the stock price or investor confidence increasing*.' Vice versa, Xianzhi Li et al. (2023) reported [quote]: '*This study is among the first to explore the most recent advancement of generically trained LLMs, including ChatGPT and GPT-4, on a wide range of financial text analytics tasks. These models have been shown to outperform models fine-tuned with domain-specific data on some tasks, but still fall short on others, particularly when deeper semantics and structural analysis are needed*.'

Both studies are strongly focussed on sentiment analysis to derive a statistical estimation for (future) investment action from (current) news feeds based on a (historical) domain-specific LMM. There is no indcation that domain-specific LLMs (or SLMs) can help to achieve any 'alpha' in a real-world investment process. As off today, no evidence is provided that domain-specific model can 'boost' any investment process, while they can offer some help in back-office related text works (like writing simple e-mails based on key words, summarising or comparing input texts, or proposing some marketing text).

This kind of a specific commercial use of LMMs trained on publicly available text corpora required a closed look on copyright issues. Similar discussions are going on concerning use of publicly available literature of living authors, which is included in the training data for LLM due to the 'scraping' of data from the internet. In Europe, the 'EU Directive on copyright and related rights' (EU, 2019) provides an exemption from copyright as long as the use of works has [quote]: '*not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online*.'

A similar, but not so formalized concept exists in the USA with the doctrine of '*fair use'*, which permits limited use of copyrighted material (openly published) without permission from the copyright holder and without copyright fees. The criteria are a 'transformative' use, which come with ambiguities. Examples in the case-based law in the USA are: commentary, search engines, criticism, parody, news reporting, research, et cetera. Recently, a new Artist Rights Alliance (2024) constituted for a [quote]: '*fair treatment for all creators in the digital world … including the right to fair market value compensation for creative work on all platforms*'[34]. However, the general doctrine was substantiated in the case *Authors Guild v. Google* at the federal court for the Southern District of New York, and then the Second Circuit Court of Appeals between 2005 and 2015 with the conclusion [quote from Fenwick & West, 2013]: '*Google Books is also transformative in the sense that it has transformed book text into data for purposes of substantive research, including data mining and text mining in new areas*.' This opinion resembles the exemption in European copyright law with the exemption of data and text mining (EU, 2019).

Yet, GenAI illustrates two new concerns, which are not covered by such an exemption. First, one can 'engineer' prompts with so much *ex-ante* knowledge about a desired output that the result would be a one-to-one copy or the original or, at least, a very similar result. It is an open issue, whether this could be regarded as copyright infringement by the developers of the GenAI tool – or it is an attempt of plagiarism by the user, who 'engineered' the prompt? Second, GenAI tools for the generation of music or video can use the voice of a singer or the visual appearance of an actor as prototype blueprint for a synthetic song or a synthetic video. This concerns the right of personality as right for an individual to control the commercial use of their identity. It remains to be seen how the development of legislation will solve these issues.

Due to these discussions about copyright/exemptions and personal rights, any internal use of public GenAI / LLMs in banking beyond rather generic applications - like writing marketing texts as internal drafts – should be carefully reviewed. For actual operations from customer service centres to internal sentiment analysis, domain-specific SLM trained on proprietary data are always a better choice – especially as the use of general LLMs in banking has not been tested for any advantage.

[34] Today, a majority of the 100.000 songs uploaded to streaming platforms is AI generated.

16. Credit Scoring: Perceptions and Expectations

The issue of credit scoring – as typical application of algorithms in banking whether rule-based or AI-based – is a special case for AI-bases tools. Since some years there has been a public concern about credit scoring, which is perceived as 'unfair' or even 'discriminating' and, consequently, is expected to 'align with ethical standards and societal values' (see quote below).

One archetypical example is the debate about the Apple Card. In 2019, there were anecdotical reports that the Apple Card would discriminate against women, which have been reiterated since then as 'evidence' for so-called algorithmic discrimination and the potential danger of AI-based solutions in banking (although it was a rule-based application).

The discourse ignores the findings of the New York State Department of Financial Services (DFS, 2021) published 23.3.2021 [quote, underlining by the author]: '*..., consumers voiced the belief that if they shared credit cards with spouses, even if only as authorized users, they were entitled to the same credit terms as spouses. In reality, however, underwriters are not required to treat authorized users the same as account holders, and may consider many other factors. In terms of gender, the Department found, based on its data analysis, that Apple Card applications from women and men with similar credit characteristics generally had similar outcomes. …, evidence showed that those decisions were explainable, lawful, and consistent with the Bank's credit policy*.'

However, the example that female partners of the account holders (i.e. authorized users with 'partner cards') had lower credit limits is still propagated as 'algorithmic discrimination' (see e.g. Bartoletti and Xenidis, 2023, in a study prepared for the Council of Europe) but I never read about the male partners with the same issue.

The answer of Axel Voss, Member of the European Parliament, in a recent interview EACB (2024) concerning the European Artificial Intelligence Act corresponds to this tendency in the public perception of AI. He was asked '*With the AI Act's implications on various sectors, particularly finance, how do you envision its influence on shaping financial practices related to AI adoption in the banking and financial services industry?*'

And he replied [quote, underlying by the author]:

> *The AI Act seeks to regulate the deployment of artificial intelligence (AI) systems within ethical and legal boundaries while fostering innovation. It adopts a risk-based approach with stricter requirements for higher-risk systems. The Act emphasises the importance of using high-quality data for AI training to mitigate biases and ensure fairness, while also requiring transparency regarding data sources and quality. ... Moreover, the Act's ethical guidelines will influence the types of AI applications adopted by financial institutions, prioritising those that align with ethical standards and societal values. This could lead to a shift towards AI solutions that prioritise fairness, inclusivity, and non-discrimination. ...*

While societal values are essential for us as parts of an open, liberal, and democratic society, legislation and regulation should be based on clear definitions and consistent rules. As long as the responsible agents with their '*intentionality*' as mentioned by Deborah G. Johnson (2006) follow the law - as e.g. the European General Data Protection Regulation (GDPR) and, consequently, do not process any sensitive personal data[35] without explicit consent of the data subject - other requirement like '*fairness'* or '*societal values'* seem rather opaque to be applied to a technical solution.

[35] GDPR Article 9 'Processing of special categories of personal data': *1. Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited*.

A European bank decides about new consumer loans solely based on the parameter ‘free average income’ in relation to the required monthly repayment:

If (free monthly income)/(required monthly repayment) > threshold

then loan approved,

else not.

The bank does neither process nor store sensitive personal data like ‘sex’ in compliance to General Data Protection Regulation (EU-GDPR Art. 9). In other words, the bank has a demographic blindness concerning protected sensitive personal data and related groups.

On the one hand, the lender has the freedom of contract, as long as it does not violate anti-discrimination legislation (e.g. European directive 2004/113/EC), on the other hand, the lender has obligations that the financial capabilities of borrowers are assessed with due diligence according to the Consumer Credit Directive (CCD, see: EU, 2023).

As in Germany (see Destatis, 2024), women have a lower income in average, compared to men - for simplification, other effects of household income are ignored - the probability for an approval will differ for the two distributions (‘women’ vs. ‘men’), if and only if an external observer uses the protected sensitive data item ‘gender’ to classify a certain sub-group *ex-post*.

Is this ‘algorithm discriminative’? Of course, not[36]. But is this ‘disparate outcome’ as defined by: $Pr(Score \mid Gender = f) \neq Pr(Score \mid Gender = m/d)$?

Because ‘Income’ is the key variable for the causal relationship, the correct questions with the variable ‘Income’ as a mediator is:

$$Pr(Score \mid Gen. = f \;\&\; Income = x) \stackrel{?}{=} Pr(Score \mid Gen. = m/d \;\&\; Income = x)$$

For a certain Income = x as a ‘mediator’ the difference between the probability distribution given a specific Gender vanished. While statistical correlations cannot provide an answer to the question about discrimination, the causal approach gives the correct insight.

Box 16.1 A simple Gedankenexperiment *concerning credit scoring facing the societal reality of the gender pay gap (adopted from Milkau, 2021, 2022, 2023; for definition and value of the gender pay gap 2023 in Germany see: Destatis, 2024)*

[36] Neither ‘direct’ as no sensitive parameters are used, nor ‘indirect’ as no rule (e.g. for a certain haircut in food industry discriminating certain religious believes) collides with the group ‘women’.

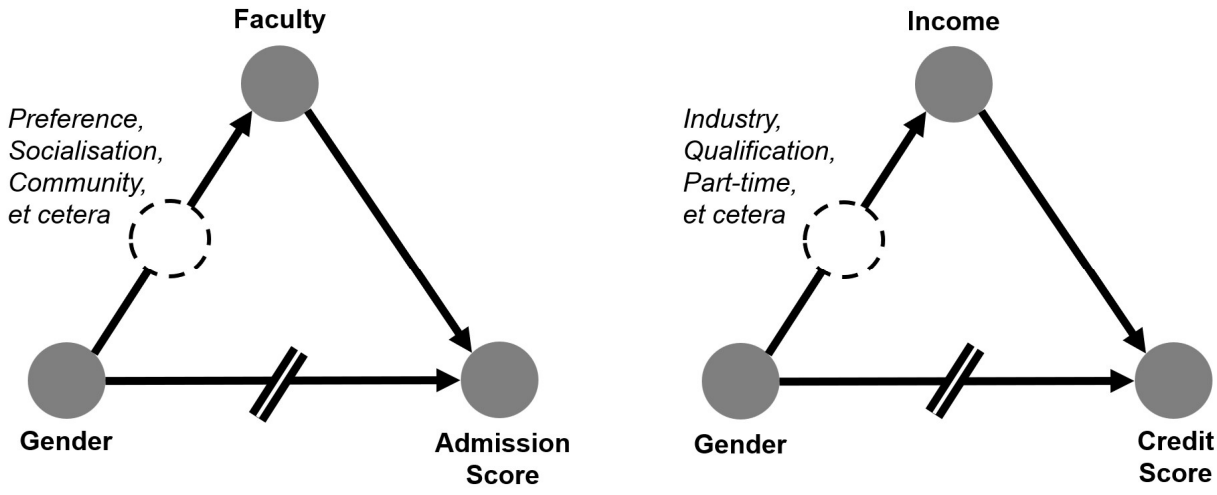


Figure 16.1: Causal directed graphs for the Berkeley Admission Paradox (see: Pearl and Mackenzie, 2018) and the Gedankenexperiment *for credit scoring. In both cases, the society can support a more equal distribution, but the score values are not discriminative towards the parameter 'Gender'.*

The pre-final version of the Artificial Intelligence Act (AIA, pre-final version, European Parliament (2024) legislative resolution of 13.3.2024) defined credit scoring as a 'high-risk AI system'[37] [quote, underlining by the author]:

Rec (58) In addition, AI systems used to evaluate the credit score or creditworthiness of natural persons should be classified as high-risk AI systems, since they determine those persons' access to financial resources or essential services such as housing, electricity, and telecommunication services. AI systems used for those purposes may lead to discrimination between persons or groups and may perpetuate historical patterns of discrimination, such as that based on racial or ethnic origins, gender, disabilities, age or sexual orientation, or may create new forms of discriminatory impacts. ...

[37] The problems concerning the definition in the AIA (Art. 3) should not be discussed in detail:
(1) 'AI system' means 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) 'risk' means the combination of the probability of an occurrence of harm and the severity of that harm;
On the one side the definition of 'AI' is opaque and could be applied to any statistical regression tool; on the other side 'risk' is defined in a traditions way (risk = probability*loss), but 'high-risk AI systems' are not classified by a quantitative approach, and simply listed in an annex without any evidence.

This special aspect of the AIA about ‘credit scoring’ reveals three issues:

1. Who takes a risk?
2. What situation is perceived (sic!) that ‘*may lead to discrimination … and may perpetuate historical patterns of discrimination”*?
3. What is the trade-off between advanced statistical approaches to credit scoring and public expectations, i.e. balance between credit risk and reputational or even legal risk?

The first question may sound strange, as any kind of lending generically included the risk that the counterparty might default. While this has been common knowledge for centuries, in the last decades and especially in European regulations one could observe a shift of paradigm, which resulted in three parallel perspectives of ‘credit risk’, which are not congruent and are present regulatory obligations:

- Risk of banks (credit risk of default of borrower; with traditional regulation and supervision of banks’ risk management and economic capital)
- Risk for borrowers (to run into over-indebtedness; see especially the European Consumer Credit Directive ‘CCD’ with latest version: EU, 2023)
- Risk for consumers (for ‘*access to financial resources*’ as described in the AIA, although these ‘*financial resources*’ are banks’ money)

The new development within the AIA follows a general pattern in this regulation of a technology (i.e. AI-based systems) that – and again in subjunctive (sic!) – that ‘*AI may generate risks and cause harm to public interests and fundamental rights that are protected by Union law. Such harm might be material or immaterial, including physical, psychological, societal or economic harm*.’ For so-called ‘*high-risk AI systems*’, the AIA conflates fundamental rights (as protection of citizens against governments, e.g. concerning remote biometric identification systems of citizens in public), product safety (already regulated in many other legislations e.g. for medical devices), access to public services (including healthcare services but also ‘*systems intended to evaluate and classify emergency calls’*) and, finally, credit scoring. Likewise, the AIA mixes ‘*access to and enjoyment of certain essential private and public services*’ with ‘*access to financial resources*’.

The second questions can be answered with the simple *Gedankenexperiment* summarized in Box 16.1, which addresses the general questions, what is regarded as '*discrimination'* (adopted from Milkau, 2021, 2022, 2023). Literally, any statistical classifier '*discriminates'* between different classes of instances. In a narrower sense, discrimination is defined as a different treatment of groups characterised by sensitive personal data (according to EU GDPR Art. 9).Direct discrimination (using such a parameter) and indirect discrimination (using a rule which has an impact on specific groups characterised by such a parameter) can be excluded, because banks comply with the appropriate anti-discrimination regulations in Europe. Therefore, the questions related to so-called '*disparate outcome'* as defined by a different probability distribution of score value given different values of the parameter 'Gender' if evaluated *ex-post* and using this parameter, which would not be allowed by GDPR:

$$\Pr(\text{Score} \mid \text{Gender} = \text{female}) \neq \Pr(\text{Score} \mid \text{Gender} = \text{male or others like divers}).$$

As this example of apparently '*disparate outcome'* is often used to explain the 'risk' of AI-based systems fir credit scoring, the simple *Gedankenexperiment* makes very clear that (i) this issue is a general one of statistical classification but not AI-based systems and (ii) a simple correlation as used in the definition of '*disparate outcome'* does not provide any evidence as long as causal relationships are not included. Figure 16.1 illustrates this misunderstanding of correlations versus causal statistics with directed graphs for the Berkeley Admission Paradox (see: Pearl and Mackenzie, 2018) compared to the *Gedankenexperiment* for credit scoring.

This comparison reveals that simple assumptions about potential discrimination in credit scoring and harm caused by AI-based systems in banking are typically misunderstandings of (sophisticated) statistics and causal relationships. As free 'Income' is a key parameter for any credit scoring, this parameter income has to be taking into account as a mediator, which determines the credit score, but separates the score from the parameter 'Gender'. If even a simply and rule-based classifier as in Box 16.1 can be misinterpreted, any discussion about more sophisticated AI-based statistical classifiers is nearly hopeless from the beginning.

Concerning the third and last question, it has to be scrutinized whether such an explanation based on advanced statistics helps for the public discourse? The author himself tried such arguments in discussion with staff of the European Commission but without any success – a fault of the author but not of the staff, who were no subject matter experts for credit scoring, statistics, or AI-based systems.

Vice versa, the banking industry has to comply with the coming AIA and has to accept the frequently technology-adverse and bureaucratic European regulations (from the GDPR[38] with a *de-facto* prohibition of automated consumer credits[39] to the coming AIA with the classification of nearly any 'predictive' credit scoring as a '*high-risk AI system'*). For the special debate about '*algorithmic fairness*' between different groups the reader is referred to WatchIT Nr. 3 "*AIgorithmic Credit Scoring in USA, Europe, and*" (Milkau, 2021) and to the example shown in Box 16.2.

Other jurisdictions (see e.g. the *'Interagency statement on the use of alternative data in credit under-writing*' of the U.S. CFPB, 2019; or the BIS paper about *'How do machine learning and non-traditional data affect credit scoring New evidence from a Chinese fintech firm?*' of Gambacorta et al., 2019) are open for technological innovations, while European regulations seems to be rather technology-adverse[40] and unfriendly to innovations such as AI in general and applications of AI-based systems e.g. for credit scoring. One can complain this tendency, but European banks have to comply with these regulations, which will limit developments of AI-bases systems in general to avoid legal and/or reputational risk.

[38] See also a recent decision of the Court of Justice of the European Union (Curia, 2024) that even non-personal Transparency and Consent String' (TC String) belong to personal data, because they could – theoretically – be correlated with IP address of the user's device.
[39] As outlined by the ruling of the European Court of Justice on 7.12.2023 that the Schufa scoring constitutes an '*automated decision-making*' prohibited under Article 22 GDPR.
[40] In addition, European regulations are increasingly protectionist and have shown antagonism against US-based technology companies in the last twenty years.

Since the public debate in 2016 whether the COMPAS software - used by U.S. courts to assess the likelihood of a defendant becoming a recidivist - would be discriminating, there has been a sharp rise in publications about '*algorithmic fairness*' between different groups. From the beginning, this discourse revealed a tension between the quality measures of statistical classifiers and the social perception of 'fairness'. For a discussion about apply statistical quality measures (such as sensitivity and specificity) as social 'fairness' conditions see the work of Jon Kleinberg with different co-authors (2017, 2018, 2019).

These problems can be illustrated by a recently published paper of Cameron Celeste et al. (2023) about '*Ethnic disparity in diagnosing asymptomatic bacterial vaginosis using machine learning*', because this paper is often cited abridged as evidence for 'Ethnic disparity in … machine learning'. This might be caused by the wording in the abstract [quote]: *'… Bacterial Vaginosis (BV) is a common vaginal syndrome among women of reproductive age and has clear diagnostic differences among ethnic groups. ... We determine the fairness in the prediction of asymptomatic BV using 16S rRNA sequencing data from Asian, Black, Hispanic, and white women. ... When evaluating the metric of false positive or false negative rate, we find that [ML] models perform least effectively for Hispanic and Asian women. Models generally have the highest performance for white women and the lowest for Asian women. …'*

However, the observed disparity was neither any kind of discrimination nor an issue of the machine learning models. The study discussed two causes further below in the paper [quotes]:

- 'The inequal performance of the models could be partially due to the imbalance of the dataset, which can make it appear that a model is performing better than it would in a clinical setting. BV-positive samples for the white and Asian populations were limited in this dataset.'

- '... the complexity of the vaginal microbiome, by defining dominant bacteria, ranging from I-V19. It is seen in this dataset that the majority of Black and Hispanic women belong to community group IV, which is the most complex, and has a high revalence of Prevotella, ... Other community groups are dominated by Lactobacillus spp. This could explain why the models perform worse for Black and Hispanic women than they do for white women.'

Without going into the details of the study, three issues are noticeable:

- The use of the term '*fairness'* as synonym for 'statistical measures' such as False Positive Rate or False Negative Rate is difficult, as it puts the statistical quality of a medical test into a social context. Yet, it might help to publish papers in scientific journals aligned to some mainstream of the discourse.
- It is always a problem to gather larger and balanced datasets in medical research. Yet, this can be regarded as another example for the practice to use 'existing' data without fully understanding the context of the 'measurement' and the implications of imbalanced data.
- This example emphasises the need to understand the causality of a phenomenon such as the basic biological processes of a medical syndrome. Yet, it could be undesirable to discuss the ethnic disparity of a disease.

Interestingly, a main result of the study that the four different machine learning models (Logistic Regression, Random Forest, Support Vector Machine, and Multi-layer Perceptron [unclear, whether the authors intend to say 'ANN'?]) show comparable general performance, is discussed only marginally.

Box 16.2 A brief example for the tension between the quality measures of statistical classifiers and the social perception of 'fairness'

17. Deep Fakes, Manipulation and Disinformation

The idea to apply technology to the manipulation of pictures and especially of filmmaking '*concentrating the audience's emotions in any direction dictated by the production's purpose*' is hundred years old and was articulated by Sergei Mikhailovich Eisenstein (1898 – 1948). As a Russian/USSR film director and film theorist he pioneered '*The Montage of Attractions*' in 1923/24 (see Eisenstein, 1998). His concept of *montage* was based on the 'generation' of new context to 'direct' the perception of spectators, and this *montage* re-used real elements (usually short film scenes, often scenes from different perspectives, but also music) to 'engineer' a new and manipulative message. Although *montage* applied a very different technology, the concept is similar to 'prompt engineering' to 'generate' images or video sequences based on a corpus of training data. In other words, technical 'fakes' are linked to mass media (or 'social media') from the beginning.

While the original proposal of the AIA did not address so-called 'deep fakes', a more principle-based 'pro-innovation approach to regulating AI' in UK presented in July 2022 (Dorries, 2022) explicitly mentioned [quote]: '*There is also concern that AI will amplify wider systemic and societal risks, for instance AI's impact on public debate and democracy, with its ability to create synthetic media such as deepfakes*. …'.

Recent changes to the AIA in the nick of time included GenAI – described as 'general-purpose AI systems' – and addressed the danger of deep fakes [quotes]:

> *Art. 3 (60) 'deep fake' means AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful;*
>
> *Art. 50 / 2. Providers of AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. ... 4. Deployers of an AI system that generates or manipulates image, audio or video content constituting a deep fake, shall disclose that the content has been artificially generated or manipulated.*

These clarifications and disclosure obligations are highly appreciated.

Unfortunately, any technology for watermarks, embedded signatures and/or identifications in a central repository can be manipulated by other technologies or even simple approaches, which would be sufficient to post a ‘deep fake’ in social media (potentially with lower quality, but still as a realistically looking image). One simplified way could look like this: original photo[41] taken in a location → ad-hoc manipulation with GenAI-tools (open source or from a non-regulated provider, without watermark or watermark removed) → ‘generated’ image with manipulations → printout or display → new photo of this manipulated picture with an external camera with the same geolocation data and hardware generated meta-information as ‘original image’ → post on social media.

The idea of *montage* highlights that the best manipulation is the ‘*audience's emotions … dictated by the production's purpose*’: from a special viewpoint of an image via a changed context (sometime simply the date) to a combination of sequences for a short video. There is no truth in LLMs, there is no truth in technical devices or tools, and there is no truth in media at all. The best measures against this problem are neither regulation nor technical solutions but socio-political approaches like independent journalism, double-check with independent sources, verification by a ‘second factor’ e.g. a photograph and a news feed, and last but not least common sense.

Skipping the issue of fraud in online chats of banks (with the possibility to generate ‘realistic’ video plus audio sequences), the most urgent case for ‘GenAI financial scammers’ will be indirect channels like ‘GenAI automated’ phishing (with generated e-mails to get access to ban credentials), spear phishing (addressed employees in specific organization with ‘individualized’ messages) or so-called CEO fraud (with generated messages, phone calls or even video calls, in which faked executives try to persuade employees to make ‘urgent’ payment transaction). Although the basic concepts are well-known and educated in employees’ trainings, the very realistic appearance of ‘generated’ content enters into a new dimension of something what could be called ‘Eisenstein-type fraud’.

[41] It should be noted that the manipulation of image is a problem for online claims management in insurance with possible fraud if customers up-load manipulated images e.g. for car insurance claims.

A final illustration of the challenges of GenAI-based fraud – and example for the never-ending race between The Hare and the Hedgehog – is 'synthetic identity fraud' as an emerging and insidious adversary. Currently, identity fraud is a problem for online merchants, but also online consumer finance providers, when stolen customer data are used to fake 'new' customers – i.e. without recorded shopping patterns - to make fraudulent order with deferred or buy-now-pay-later payments. Advanced fraud detection systems will (i) analyse the context (type of order, product, amount, time of day et cetera) but also (ii) use key-stroke patterns[42] with the difference in rhythm between a human entering his name, address and date of birth and a system, which will 'copy' a sequence as input.

Sometimes, there is biases perception about the risks of AI but ignoring the opportunities, as Yann LeCun pointed out in an interview end of 2023 (Levy, 2023):

> Steven Levy (Editor at Large for Wired)*: Why are so many prominent people in tech sounding the alarm on AI?*
>
> Yann LeCun (22.12.2023)*: Some people are seeking attention, other people are naive about what's really going on today. They don't realize that AI actually mitigates dangers like hate speech, misinformation, propagandist attempts to corrupt the electoral system. At Meta we've had enormous progress using AI for things like that. Five years ago, of all the hate speech that Facebook removed from the platform, about 20 to 25 percent was taken down preemptively by AI systems before anybody saw it. Last year, it was 95 percent.*

Finally, one should never forget that all this anti-fraud measures are statistical estimations about fraud/no fraud. As already elaborated for the case of credit card fraud, any provider has to calculate the commercial trade-off between False Positive / False Negative, i.e. accepting an amount of fraud to avoid churn of unsatisfied customer.

[42] It is an additional regulatory problem in Europe that this kind of 'individual patters recognition' could be regarded as either 'biometric' identification and/or processing of 'personal' data.

18. Risks, Fears and Misunderstandings

Although AI and especially GenAI are statistical classifiers and although systems with embedded GenAI from Retrieval Augmented Generation (RAG[43], see chapter 8) to GenAI-powered robots made by Figure have neither an own intentionality nor a free will, there is an increasing debate in the public. In general, three opinions are articulated:

1. The optimistic opinion – like Yann LeCun as mentioned in the previous chapter – that AI can be helpful and vice versa [quote]: '*AI doomism is quickly becoming indistinguishable from an apocalyptic religion*. *Complete with prophecies of imminent fire and brimstone caused by an omnipotent entity that doesn't actually exist*.' (LeCun, 2023)
2. A pragmatic perspective as pointed out by the U.S. Secretary of Commerce Gina Raimondo in a press release (U.S. Department of Commerce, 2024): *'AI is the defining technology of our generation. This partnership [*on Science of AI Safety*] is going to accelerate both of our Institutes' work across the full spectrum of risks, whether to our national security or to our broader society.*'
3. The perspective of 'doomerism' as summarized by Melissa Heikkilä (2023) saying [quote]: '*AI doomerism went mainstream ... Existential risk has become one of the biggest memes in AI. ... It's an ideology championed by many in Silicon Valley, ...*' Very often, this perspective conflates the contemporary technologies of AI and GenAI as statistical classifiers, the vision of some Artificial General Intelligence (AGI, in the sense of the mentioned McCarthy et al., 1955), and fear of some 'superintelligence'.

[43] It is worth to note that sophisticated concepts like RAG, in which GenAI is embedded in an overall application with data retrieved from 'outside', introduce new technical risks due to new attach vectors: Indirect Prompt Injections are one example for this problem that the 'outside' data sources can be manipulated and compromised by an attacker. If these data are retrieved and included in a prompt to a LLM, the attacker could have injected some text in these data, which are used in the prompt but will change the original prompt to follow the 'instructions' of the attacker (see e.g. BIS, 2023 and BIS, 2024).

One year ago, three open letters have been published that an existential risk of AI ('x-risk') is near [quotes]:

- *'Should we risk loss of control of our civilization?'* (Bengio et al., 2023)
- *'…, superintelligence will be more powerful than other technologies humanity has had to contend with in the past. … Given the possibility of existential risk, we can't just be reactive. Nuclear energy is a commonly used historical example ...'* (Altman, Brockman, and Sutskever, 2023)
- *'Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.'* (Hinton, Bengio, Hassabis, Altman et al., 2023)

Similar arguments have been reiterated – for example in a recent Policy Forum contribution in the journal Science in April 2024 with overlapping authors [quote]:

- *'Highly capable and far-sighted RL agents* [Reinforcement Learning; also described as long-term planning agents 'LTPA' in this contribution] *are likely to accrue reward very successfully. … One path to maximizing long-term reward involves an RL agent acquiring extensive resources and taking control over all human infrastructure, which would allow it to manipulate its own reward free from human interference. Additionally, because being shut down by humans would reduce the expected reward, sufficiently capable and far-sighted agents are likely to take steps to preclude that possibility ... Progress in AI could enable such advanced behavior. (Cohen, Kolt, Bengio, Hadfield, and Russell, 2024)*

Yet, warnings that AI could be self-learning beyond any human control and achieve their individual non-human goals are not new. Such apocalyptic visions have been common in the last 60 years, and the plot of D.F. Jones novel 'Colossus' – written in 1966 – is a very, very similar story of computers taking control (sic!).

Of course, there is a risk – but the risk of misuse of AI by humans. One can rephrase the quote by Deborah G. Johnson (2006) to: *'AI-based systems have risks, the risks due to the intentionality put into them by the intentional acts of their designers.'* A possible reason for people to be frightened by assumed capabilities - beyond the intentionality of the designers, users and AI-supported manipulators – is the terminology ('intelligence', 'learning', 'autonomy', 'emergence') and especially the term 'self-learning'.

It is important to understand that both – supervised and un-supervised learning are both '*able to fit a function to a collection of historical data points*' as coined by Pearl and Mackenzie (2018). In the first case, the data consists of fix-length data (images) plus a label; in the second case, the data are sequences with a 'next token in sequence' to be 'learned'.

One example is a recently publish essay by Holger Lyre (2024) '*ChatGPT versteht es'*, in which he asked whether LLM do 'understand' the meaning of the texts they generate and whether they possess a so-called semantic grounding. To develop an answer, he started from the idea of 'self-learning' and continues with the questions, whether LLMs obtain a 'semantic grounding' to understand the meaning of the texts in three steps: functional grounding, social grounding, and causal grounding.

Deviant from Holger Lyre's essay, one can directly test these three requirements with the concept of LLMs as statistical classifiers, which are 'trained' on a text corpus with words mapped to tokens and parameters to be fitted to an internal representation of the probabilities of sequences of tokens and the estimation of a 'next-best-token'.

Taking the example in chapter 7 ('*During the day, all my cats were [?]*'), functional grounding can be simplified to the questions, whether the LLM 'understands' what the 'meaning' of a cat is? Does a LLM 'understands' the functional role of a cat (as an animal, as a pet, as a curious mammal, but not as a 'gambler')? The answer is simply: no, as a LLM is a representation of statistical relationship between token in sequences.

Concerning social grounding, it is trivial that LLMs are limited to language-based behaviour and cannot interact with social agents besides text messages (or technical text-to-speech generation), which is weak form of social grounding. If one looks only to a short-term interaction – such as a conversation - all the known problems with the ELIZA chatbot of Joseph Weizenbaum (1966) surface again: People can regard simple chatbots, which mirror the human input, as 'social' agents. And if the perspective is extended to long-term 'language games', which develop linguistic practices, LLMs can merely reproduce probabilistically 'next-best-tokens', which represent the statistical properties of the text corpus used for training and which consists of text produced by human beings before.

There is no dynamical and interacting 'language game', as LLMs are trained once with a fixed text corpus (including post-processing with reinforcement learning from human feedback and/or rule-based fine-tuning). They are deployed and used in an 'executable' implementation as any other computer software. Vice versa, any simple (and analogue) control system with a feedback loop can 'adapt' to the development of the system to be controlled. But neither analogue control systems nor LLMs have a social grounding.

Finally, Holger Lyre (2024) links the question of a potential causal grounding to the question whether LLMs have a 'world model', which is a tricky term, as a 'world model' in the context of control theory and also AI is a terminus technicus about the ability of a control system to '*map the environment*' in space and time like a robotic vacuum cleaner 'learns' a model representation of the geometry of a room with walls and objects but also cats running around, when they are not sleeping.

Adaption of control theory to AI started in the 1990s, and important papers about 'world models' with RNNs were authored *inter alia* by Jürgen Schmidhuber (1990), David Ha & Jurgen Schmidhuber (2018), and Yann LeCun (2022). And there are car navigation systems with sophisticated maps, which use online feeds with traffic information and even online recognition of traffic signs to calculate the route.

Other discussions about 'world models' should be skipped to return to the original question whether LLMs have a causal grounding. A brief look to '*The Book of Why*' by Judea Pearl and Dana Mackenzie (2018) helps to find a simple answer: They don't! As LLMs are statistical classifiers[44] with an internal representation of correlations between numbered tokens in sequences, they are a highly advanced version of a probability table how to find the 'next-best-token' (see Fig. 7.1). But correlations aren't causality, and causality requires some appropriate representation like directed graphs (see Fig. 15.1) or rules of physics (such as *actio* = *reactio*).

[44] One could argue that recently launched or expected versions of LLMs (like Antrophic's Claude 3 Opus (4.3.2024) or OpenAI's GPT-5) could have some 'emergent' capabilities. However, there is a clear '*end of LLMs*', when all available text corpora including programming language will be included in the training data. There will be more time left for multi-modal GenAI, as it will take some time to scrap all available videos, music, or even sequences of moves of cars. But in the end, all these attempts will lead to the 'perfect' model described by Léon Bottou and Bernhard Schölkopf (2023), which will contain all output produced by human beings.

The discussion about ‘x-risk’ has a number of components, which are far from any assessment of actual risks of AI such as disinformation:

- individual ‘hidden’ agendas and a strange ‘*ideology of doomerism’,*
- many misunderstandings of a specific terminology in AI with termini technici, which have to be understood in the right way,
- a very high level of ambition since 1956 with visionary concepts of an *Artificial Intelligence*, which went belly-up very often,
- and a general fear of robots and computers since Karel Capek’s *‘R.U.R.: Rossum's Universal Robots’*

Today, all computer code written by humans is executing pre-defined ‘*if - than - else*’ statements, calculating parameters like statistical probabilities, and looking up internal tables how to find a ‘next-best-token’. Still, Stanley McChrystal und Anna Butrico (2021) are very right [quote]: „*The Greatest Risk Is Us!*“

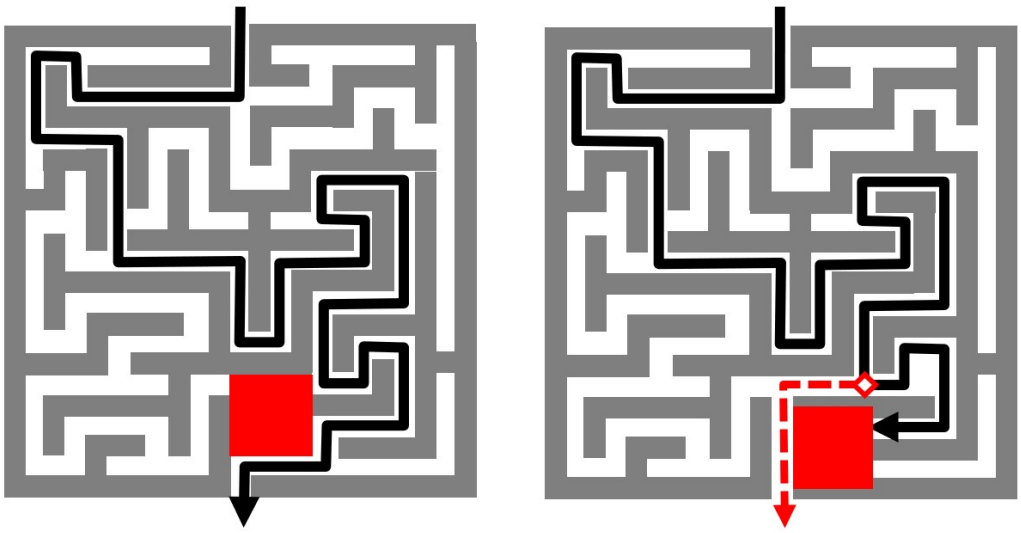


Figure 19.1: A labyrinth for an agent interactively 'learning' its way by recording the way and adopting existing 'knowledge items' (i.e. if-the-else rules like 'if there is a junction: then turn left' or 'if the way is blocked: then go back' etc.). Between two test runs, the red block is moved so that the 'old' way is blocked, and the agent has to find a 'new' way by reasoning from the knowledge items (adapted from Milkau, 2020).

19. From Machine Learning to Machine Reasoning

The title of this chapter is quoted from an essay written by Léon Bottou (2014), and he started with a definition that 'machine reasoning'[45] could be described as [quote]: '*algebraically manipulating previously acquired knowledge in order to answer a new question'*.

In principle, the concept of 'machine reasoning' is the old symbolic-logic approach to AI (see chapter 2) with an extension that the basic knowledge can be 'acquired' (i.e. recorded, but not pre-defined) by an agent at run-time. However, the 'acquired' knowledge has to be distinguished from the 'provided' knowledge items: While the basic '*if-then-else*' rules for an agent navigating in the labyrinth shown in Fig. 19.1 have to be provided, the agent is assumed to be capable to map/document its way and acquire an internal representation of the geometry of the labyrinth.

[45] This definition of ‚reasoning' has to be distinguished from another definition used in the context of so-called 'emerging' capabilities of LLMs. For example, a benchmark text for LLMs '*MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI*' (Yue et al., 2023) proposed a test for 'reasoning' but is only a collection of multiple-choice question from college exams et cetera, which reveals that LLMs were trained on text corpora including these exam texts.

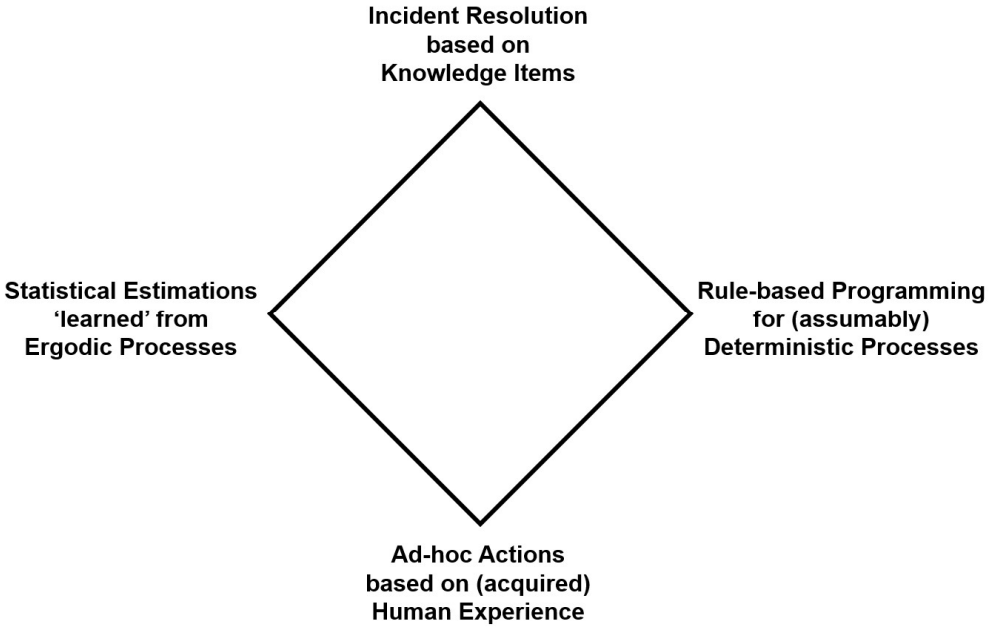


Figure 19.2: A schematic taxonomy of different types of processes and related technical solutions.

Similar concepts were discussed especially by Jurgen Schmidhuber (1990) or David Ha and Jurgen Schmidhuber (2018), who described a system ofVision (V, for detecting the environment), Memory (M, to store and retrieve parameters), and Controller (C, to compare an objective with the result of an action).

Today, similar concepts have been implemented for automated incident management in datacentres and for IT infrastructures. In this case the 'knowledge item' are documented procedures how things can be done (e.g. re-starting a server, re-setting a queue, re-connecting a device et cetera). Such systems can use 'knowledge item' at run-time to solve emerging problems based on old (= recorded) solutions but also with adapted ways, if the old solutions do not work. The basic 'knowledge item' themselves can be simple rules or more complicated building blocks like statistical estimators or even ANNs. The 'knowledge' can be stored in knowledge bases as simple list of rules or as knowledge-graph databases with a representation of the relationship between the items, from which the items can be retrieved later.

It would be far beyond the ambition level of this summary to dive deeper into the development of 'machine reasoning'. The brief introduction to 'machine reasoning' shall illustrate that there is much more AI than only simple GenAI-based statistical classifiers to estimate a 'next-best-token' on a probabilistic basis.

As illustrated with the schematic taxonomy in Fig. 19.2, there are different types of processes – and depending on these types of processes and the related problems there are different technical solutions:

- Statistical estimations 'learned' from ergodic processes,
- Rule-based programming[46] for (assumably) deterministic processes,
- Incident resolution based on knowledge items,
- Ad-hoc actions based on (acquired) human experience.

Those types of processes are abstract definitions, which are not likely to be found in reality completely. For example, the requirement for an ergodic process - that 'expected' statistical properties can be deduced from a sufficiently long sample of the process - is not given in case of the 'Turkeys on Thanksgiving', which live a 'regularly' life until an 'unexpected' end[47]. And internal processes in a company are expected to match the 'official' process description in manuals, while in reality the processes are 'living' and change often unnoticeably under the radar[48].

The second dimension in Fig. 19.2 shows step-wise approach (compared to regular processes) such as the mentioned incident management (combining single 'knowledge items') or ad-hoc actions by expert staff. Examples for the latter type could be middle-office or back-office operations to correct a SWIFT message with inconsistent data elements, to discuss differences in fee calculations for a brokerage order, or the evaluate deviations in a collateral margin call, which usually require a communication with a counterparty outside the bank's perimeter.

These types of processes are limits for a reasonable usage of AI in the sense of a 'statistical classifier' or of 'machine reasoning' as algebraical manipulation of knowledge items. Once again, we have to understand first, what the problem is, before we select the tool, which matches the challenge.

[46] It does not depend on the formal structure, in which a program was written: whether in a computer language like Java, Python, FORTRAN or COBOL, for a proprietary Business Process Management (BPM) platform, or in a graphical programming environment.
[47] This is a non-trivial issue in risk management, as time series of recorded data have to be 'long enough' to include rare, but severe events in the tails of the probability distribution.
[48] Vice versa, the method of 'Process Mining' used event log data from software systems to create a picture of the actual processes (see especially: Wil M.P. van der Aalst, 2011).

20. Conclusion

First, AI is about understanding data, the unavoidable 'noise' in data, and statistics. Second, AI depends on defined objectives and statistical quality criteria, what to achieve. Third, contemporary AI tools are – *cum grano salis* - statistical classifiers to give estimations about future instances based on historical data, but no engines for rule-based[49] transaction processing. Fourth, there has been much hype and ambitions since 1956: Yet so-called expert systems faced two 'AI winters', and even autonomous driving is struggling with 'corner cases' at the time being. And finally, the current hysteria about 'Generative AI' (GenAI) focused much on things like students' homework - and, of course, they will use such tools, but with GenAI they can only achieve a mediocre result in the sense of a statistical average of all historical text written for the same question – and the generation of fake images.

It was beyond the scope of this paper, to dive into technical details or speculate about imaginary use cases. At the end of the day, banks and financial institutions have to decide for themselves, what they want, what they are allowed to do, what resources they have, and what they will pay for. However, it is not reasonable to hinge on the current hype, doomerism, or promises for an incredible wave of productivity. The toolbox of AI offers a lot of different building blocks, which can be selected and combined depending on the problem to be solved, the resources available and, especially, an understanding of data and statistics.

I would like to quote Wolfgang Wahlster, a pioneer of AI and founding director of the German Research Center for Artificial Intelligence (DFKI), what he said in an interview worth reading [quote:]

> *Künstliche Intelligenz ist besser als natürliche Dummheit*. (Wahlster, 2015)

And I would like to add:

> *Aber unsere natürliche Dummheit ist die größte Gefahr im Umgang mit Künstlicher Intelligenz.*

Therefore, we should avoid science fiction, hype, and marketing messages, but continuously attempt to learn about the benefits and risks of AI without prejudices.

[49] As already mentioned, a method like Robotic Process Automation (RPA) is sometimes included in a definition, what AI could be, RPA is at its core a rule-based approach to automate the flow of transaction between different applications.

References

Bottou, Léon und Bernhard Schölkopf (2023) “Borges und die Künstliche Intelligenz”, Frankfurter Allgemeine Zeitung, 18.12.2023, p. 18; and ‘Borges and AI’, arXiv, 4.10.2023 (available at https://arxiv.org/pdf/2310.01425.pdf, accessed 13.2.2024).

Schnabel, Deborah and Eva Berendsen (2024) “Die TikTok-Intifada - Der 7. Oktober & die Folgen im Netz”, Bildungsstätte Anne Frank, 31.1.2024 (available at: https://www.bs-anne-frank.de/fileadmin/content/Publikationen/Weiteres_P%C3%A4dagogisches_Material/TikTok_Studie-Bildungsst%C3%A4tte_2024-WEB.pdf, accessed 15.2.2024)

Ibrahim, Hazem, Fengyuan Liu, Yasir Zaki and Talal Rahwan (2024) "Google Scholar is manipulatable", arXiv.org, 7.2.2024 (available at https://arxiv.org/pdf/2402.04607.pdf, accessed 21.2.2024)

Warren, Tom (2024) "Google pauses Gemini’s ability to generate AI images of people after diversity errors", The Verge, 22.2.2024 (available at https://www.theverge.com/2024/2/22/24079876/google-gemini-ai-photos-people-pause, accessed 24.2.2024)

Schmidhuber, Jürgen (2022) "Annotated History of Modern AI and Deep Learning", 29.12.2022 (available at: https://arxiv.org/ftp/arxiv/papers/2212/2212.11279.pdf, accessed 4.3.2024)

McCarthy, John, Marvin Minsky, Nathan Rochester, and Claude E. Shannon (1955) “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence”, Aug. 1955 (available at: http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf; accessed 11.4.2020)

Schmidhuber, Jürgen (2015) "Deep learning in neural networks: An overview", Neural Networks, Vol. 61, Jan. 2015, pp. 85-117 (available at https://www.sciencedirect.com/science/article/abs/pii/S0893608014002135, accessed 28.2.2024)

Kurzweil, Ray (2012) “How to Create a Mind”, Viking Penguin, New York, USA

Melanie Mitchell (2023) ‘How do we know how smart AI systems are?’, Science, 13.7.2023, Vol 381/6654 (available at https://www.science.org/doi/10.1126/science.adj5957, accessed 21.2.2024)

Berg, Aksel (also Axel I.; 1961) “Kibernētiku - na službu kommunizmu" [Cybernetics at the service of communism]”, Gosenergoizdat, Leningrad, quoted after: Semenov, Alexei at al. (2020) "Axel Berg’s Legacy in Cybernetics and Education. From the Council on Cybernetics to Axel Berg Institute", 2020 Fifth International Conference “History of Computing in the Russia, former Soviet Union and Council for Mutual Economic Assistance countries” (SORUCOM), Moscow, Russia, 2020, pp. 152-157, IEEE Xplore: 30 June 2021, https://ieeexplore.ieee.org/document/9464970

Winograd, Terry A. and Fernando Flores (1986) "Understanding Computers and Cognition", Ablex Publishing Corporation, Norwood, NJ, USA

Pearl, Judea and Dana Mackenzie (2018) “The Book of Why”, Basic Books/Hachette Book Group, N.Y. (15.5.2018)

Ben-David, Shai and Shai Shalev-Shwartz (2014) "Understanding Machine Learning: From Theory to Algorithms", 17.7.2014, Cambridge University Press, New York, USA

Yann LeCun et al. (1988) "Gradient-based learning applied to document recognition", Proceedings of the IEEE, Vol. 86/11, pp. 2278 - 2324

Pearl, Judea (1988) "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference", Morgan Kaufmann Series in Representation and Reasoning, 1.9.1988, Morgan Kaufmann Publishers/Elsevier, San Francisco, USA

Schmidhuber, Jürgen (2015) "Deep learning in neural networks: An overview", Neural Networks, Vol. 61, Jan. 2015, pp. 85-117 (available at https://www.sciencedirect.com/science/article/abs/pii/S0893608014002135, accessed 28.2.2024)

Lam, Remi et al. (2023) "Learning skillful medium-range global weather forecasting", Science, 14.11.2023, Vol. 382/6677, pp. 1416-1421

Heindel, Walter (2024) “Zu riskant für junge Frauen”, Interview in Frankfurter Allgemeine Zeitung, 6.3.2024, p. N1

Degtiar, Irina and Sherri Rose (2023) "A Review of Generalizability and Transportability", Annual Review of Statistics and Its Application, Vol. 10 (Volume publication date March 2023), p. 501-524 (available at: https://www.annualreviews.org/doi/full/10.1146/annurev-statistics-042522-103837, accessed 3.3.2024)

Pearl, Judea (2000) “Causality”, Cambridge University Press, New York, USA; reprint 2017 of the second edition of 2009 with updates

Pearl, Judea (2010) "Causal Inference", NIPS 2008 workshop on causality, JMLR Workshop and Conference Proceedings, Vol. 6, pp. 39–58 (available at: https://proceedings.mlr.press/v6/pearl10a/pearl10a.pdf, accessed 3.3.2024)

Lapuschkin, Sebastian et al. (2019) "Unmasking Clever Hans predictors and assessing what machines really learn", Nature Communications, Vol. 10, Art: 1096 (available at: https://www.nature.com/articles/s41467-019-08987-4.pdf, accessed 3.3.2024)

Anders, Christopher J. (2022) "Finding and removing Clever Hans: Using explanation methods to debug and improve deep models", Information Fusion, Vol. 77, pp. 261-295 (available at: https://pdf.sciencedirectassets.com/272144/1-s2.0-S1566253521X00085/1-s2.0-S1566253521001573/main.pdf, acessed 3.3.2024)

HHI (2019) ‘Der Blick in Neuronale Netze’, Forschung Kompakt (in German), Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut (HHI), 1.7.2019 (available at: https://www.fraunhofer.de/de/presse/presseinformationen/2019/juli/der-blick-in-neuronale-netze.html; accessed 20.6.2023).

Goodfellow, Ian, Yoshua Bengio and Aaron C. Courville (2015) "Deep Learning", MIT Press, Cambridge, MA, USA (available at: http://alvares-tech.com/temp/deep/Deep%20Learning%20by%20Ian%20Goodfellow,%20Yoshua%20Bengio,%20Aaron%20Courville%20(z-lib.org).pdf, accessed 5.3.2024)

Jurafsky, Daniel and James H. Martin (2024) "Speech and Language Processing", Draft 3.2.2024, (available at: https://web.stanford.edu/~jurafsky/slp3, accessed 19.3.2024)

Gu, Albert et al. (2022) "Structured State Spaces: Combining Continuous-Time, Recurrent, and Convolutional Models", Hazy Research, Stanford University, 14.1.2022 (available at: https://hazyresearch.stanford.edu/blog/2022-01-14-s4-3, accessed 31.3.2024)

Vaswani, Ashish et al. (2017) "Attention Is All You Need", Advances in Neural Information Processing Systems 30 (NIPS 2017), (available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf, accessed 5.3.2024)

OpenAI (2023) "GPT-4 Technical Report", 27.3.2023 (available at: https://cdn.openai.com/papers/gpt-4.pdf, accessed 12.3.2024)

Shin, Minkyu et al. (2023) "Superhuman Artificial Intelligence Can Improve Human Decision Making by Increasing Novelty", PNAS 2023, Vol. 120/12, Art.: e2214840120 (available at: https://www.pnas.org/doi/epdf/10.1073/pnas.2214840120, accessed 8.3.2024)

Warren, Tom (2024) "Google pauses Gemini's ability to generate AI images of people after diversity errors", The Verge, 22.2.2024 (available at: https://www.theverge.com/2024/2/22/24079876/google-gemini-ai-photos-people-pause, accessed 24.2.2024)

Feng, Shangbin, Chan Young Park, Yuhan Liu and Yulia Tsvetkov (2023) "From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models", Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 9.-14.7.2023, Volume 1, pp. 11737–11762, Association for Computational Linguistics, Toronto, Canada (available at: https://aclanthology.org/2023.acl-long.656.pdf, accessed 11.3.2024)

Rozado, David (2024) "The Political Preferences of LLMs", arXiv, 2.2.2024 (available at: https://arxiv.org/ftp/arxiv/papers/2402/2402.01789.pdf, accessed 10.3.2024)

Juvenal (58-138?) Decimvs Ivnivs Ivvenalis, "Satura VI", lines 347–348 (available at: https://www.thelatinlibrary.com/juvenal/6.shtml, accessed 12.3.2024)

Liu, Xiaoxuan et al. (2019) "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis", The Lancet - Digital Health, Vol. 1/6, E271-E297, online: 25.9.2019 (available at: https://www.thelancet.com/action/showPdf?pii=S2589-7500%2819%2930123-2, accessed 14.3.2024)

Plesner, Louis Lind et al. (2023) "Commercially Available Chest Radiograph AI Tools for Detecting Airspace Disease, Pneumothorax, and Pleural Effusion", Radiology, Vol. 308/3, online: 26.9.2023 (Link: https://pubs.rsna.org/doi/full/10.1148/radiol.231236, accessed 14.3.2024

Fama, Eugene F. (2013) "Two Pillars of Asset Pricing", Prize Lecture, 8.12.2013 (https://www.nobelprize.org/uploads/2018/06/fama-lecture.pdf, accessed 13.6.2023)

Edwards, Tim et al. (2024) "SPIVA® Europe Scorecard", S&P Dow Jones Indices, 16.4.2024 (available at: https://www.spglobal.com/spdji/en/documents/spiva/spiva-europe-year-end-2023.pdf, accessed 19.4.2024)

Waabi (2024) "Introducing Copilot4D: A Foundation Model for Self-Driving", 15.3.2024 (available at https://waabi.ai/introducing-copilot4d/, accessed 17.4.2024)

Zhang, Lunjun et al. (2024) "Learning Unsupervised World Models for Autonomous Driving via Discrete Diffusion", arXiv, 16.1.2024, published as a conference paper at ICLR 2024 (available at: https://arxiv.org/pdf/2311.01017.pdf, accessed 17.4.2024)

Coldewey, Devin (2015) "Google's Self-Driving Cars Use Halloween to Learn to Recognize Costumed Kids", NBC News, 2.11.2015 (available at: https://www.nbcnews.com/tech/innovation/googles-self-driving-cars-use-halloween-learn-recognize-costumed-kids-n455901, accessed 16.3.2024)

Armbruster, Alexander and Johannes Winterhagen (2023) "Wo autonomes Fahren Sinn macht", Interview mit Steven Peters, Frankfurter Allgemeine Zeitung, 11.12.2023, S. 18

David Autor (2024) "Applying AI to Rebuild Middle Class Jobs", NBER, Working Paper 32140, Feb. 2024 (available at: https://www.nber.org/system/files/working_papers/w32140/w32140.pdf, accessed 16.3.2024)

Kolodny, Lora (2023) “Cruise confirms robotaxis rely on human assistance every four to five miles”, CNBC, 6.11.2023 (available at: https://www.cnbc.com/2023/11/06/cruise-confirms-robotaxis-rely-on-human-assistance-every-4-to-5-miles.html, accessed 22.11.2023)

Coeckelbergh, Mark (2022) "Robot Ethics", The MIT Press, 6.9.2022, Cambridge

Engisch, Karl (1930) "Untersuchungen über Vorsatz und Fahrlässigkeit im Strafrecht", O. Liebermann, Berlin (Neudruck: Scientia, Aalen, 1964)

Welzel, Hans (1951) "Zum Notstandsproblem", Zeitschrift für die gesamte Strafrechtswissenschaft, Vol. 63/1, S. 47–56.

Foot, Philippa (1967) "The Problem of Abortion and the Doctrine of the Double Effect", Oxford Review, No. 5; ebenso in: Philippa Foot "Virtues and Vices and Other Essays in Moral Philosophy", 1977 (verfügbar unter https://philpapers.org/archive/FOOTPO-2.pdf, abgerufen am 9.11.2023)

Thomson, Judith Jarvis, Killing (1976) Letting Die, and the Trolley Problem, 59 The Monist 204-17

Misselhorn, Catrin (2022) "Artificial Moral Agents - Conceptual Issues and Ethical Controversy", in: S. Voeneky et al. (Hrsg,) "The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives", S. 31-49, Cambridge University Press, Cambridge

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon and Iyad Rahwan (2018) "The Moral Machine experiment", Nature, 24.10.2018, Vol. 563, S. 59-64

Johnson, Deborah G. (2006) "Computer systems: Moral entities but not moral agents", Ethics and Information Technology, Vol. 8/4, S. 195–204

Anja Utler (2024) "Die Maschine agiert wie eine ertappte Schülerin", Frankfurter Allgemeine Zeitung, 11.3.2024, S. 12

Cao, Hanqun et al. (2023) "A Survey on Generative Diffusion Models", arXiv, 23.12.2023 (available at: https://arxiv.org/html/2209.02646v10, accessed 9.4.2024)

NVIDIA (2024) NVIDIA Blackwell Architecture, 18.3.2024 (available at: https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/, accessed 20.3.2024)

Choi, Yejin (2023) “An AI Odyssey: the Dark Matter of Intelligence” Keynote at Computer Vision and Pattern Recognition, Vancouver, 21.7.2023 (slides available at: http://lxmls.it.pt/2023/slides/yejin.pdf, accessed 22.7.2023)

Schaeffer, Rylan, Brando Miranda and Sanmi Koyejo (2023) "Are Emergent Abilities of Large Language Models a Mirage?", 37th Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, 10.-16.12.2023 (available at: https://openreview.net/pdf, accessed 23.12.2023)

Figure (2024) "Figure Status Update - OpenAI Speech-to-Speech Reasoning", www.figure.ai, on YouTube, 9.3.2024 (available at: https://www.youtube.com/watch?reload=9&v=Sq1QZB5baNw, accessed 25.3.2024)

Bloomberg (2024) "Humanoid Robots at Amazon Provide Glimpse of an Automated Workplace", 4.3.2024 (available at: https://www.bloomberg.com/news/articles/2024-03-04/amazon-warehouses-provide-glimpse-of-workplace-humanoid-robots, accessed 25.3.2024)

Tesla (2023) "Optimus - Gen 2", 14.12.2023 (https://www.youtube.com/watch?v=D2vj0WcvH5c, accessed 25.3.2024)

Sean Trott (2023) "In cautious defense of LLM-ology", Substack, 2.3.2023 (available at https://seantrott.substack.com/p/in-cautious-defense-of-llm-ology, ac-cessed 23.2.2024)

Prompt Engineering Guide (2024) "Prompt Engineering Guide", a project by DAIR.AI, (available at: https://www.promptingguide.ai/, accessed 21.3.2024)

Zhou, Pei et al. (2024) "Self-Discover: Large Language Models Self-Compose Rea-soning Structures", arXiv, 6.2.2024 (available at: https://arxiv.org/pdf/2402.03620.pdf, accessed 19.4.2024)

BIS (2024) "Project GAIA", BIS Innovation Hub, 18.3.2024 (available at: https://www.bis.org/publ/othp84.pdf, accessed 21.3.2024)

Milkau, Udo (2023) "The Challenges of Generative Artificial Intelligence in Asset Management", Capco Journal of Financial Transformation, Vol. 58, Nov. 2023, 4.12.2023, pp. 138-149

Milkau, Udo (2024) "Large Language Models im Banking - Nutzen und Nutzbarkeit", BIT (Banking and Information Technology), Issue 1/2024, 19.3.2024

Capgemini Research Institute (2023) 'Why consumers love generative AI', Capgemini Research Institute, 19.6.2023 (available at: https://prod.ucwe.capgemini.com/wp-con-tent/uploads/2023/05/Final-Web-Version-Report-Creative-Gen-AI.pdf, accessed 21.6.2023)

Bauer, Dominik (2024) “Diese Kombination ist ungewöhnlich, aber warum nicht?”, Frankfurter Allgemeine Sonntagszeitung, 24.3.2024, p. 16

Perrault, Ray and Jack Clark (2024) "Artificial Intelligence Index Report 2024", 15.4.2024, HAI, Stanford Institute for Human-Centered Artificial Intelligence, Stanford University (available at: https://aiindex.stanford.edu/report/, accessed 18.4.2024)

Waber, Ben and Nathanael J. Fast (2024) "Is GenAI's Impact on Productivity Over-blown?", Harvard Business Review, 8.1.2024

Shumailov, Ilia et al. (2024) "The Curse of Recursion: Training on Generated Data Makes Models Forget", arXiv, 14.4.2024 (available at: https://arxiv.org/html/2305.17493v3, accessed 19.4.2024)

Alemohammad, Sina et al. (2024) "Self-Consuming Generative Models Go MAD", Published as a conference paper at ICLR 2024 (available at: https://open-review.net/pdf?id=ShjMHfmPs0, accessed 19.4.2024)

Weizenbaum, Joseph (1966) "ELIZA – A Computer Program For the Study of Natural Language Communication Between Man And Machine", Communications of the ACM, Vol. 9/1, pp. 36–45 (available https://dl.acm.org/doi/pdf/10.1145/365153.365168, accessed 24.3.2024)

Maersk (2022) "A.P. Moller - Maersk and IBM to discontinue TradeLens, a block-chain-enabled global trade platform", press releases, 29.11.2022 (available at: https://www.maersk.com/news/articles/2022/11/29/maersk-and-ibm-to-discontinue-tradelens, accessed 24.3.2024)

Graves, Axel and Jürgen Schmidhuber (2008) “Offline handwriting recognition with multidimensional recurrent neural networks", Advances in Neural Information Pro-cessing Systems 21, 2008, pp. 545–552

BIS (2023) "Project Aurora: the power of data, technology and collaboration to com-bat money laundering across institutions and borders", BIS Innovation Hub, 31.5.2023 (available at: https://www.bis.org/publ/othp66.pdf, accessed 25.3.2024)

db (2014) "How AI is changing banking", undated (available at: https://www.db.com/what-next/digital-disruption/better-than-humans/how-artificial-intelligence-is-changing-banking/index, accessed 4.4.2024)

Baur, Sebastien et al. (2014) "HeAR - Health Acoustic Representations", arXiv, 4.3.2024 (available at: https://arxiv.org/abs/2403.02522, accessed 5.4.2024)

Bellefonds, Nicolas de et al. (2023) “Turning GenAI Magic into Business Impact”, BCG, 11.12.2023 (available at: https://www.bcg.com/publications/2023/maximizing-the-potential-of-generative-ai, accessed 16.2.2024)

Chui, Michael et al. (2023) "The economic potential of generative AI", McKinsey & Company, 14.6.2023 (available at: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier, accessed 6.4.2024)

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond (2023) “Generative AI at Work”, NBER, Working Paper 31161,21.4.2023 (available at: https://danielle-li.github.io/assets/docs/GenerativeAIatWork.pdf, accessed 28.3.2024)

Kyogo Kanazawa et al. (2022) “AI, skill, and productivity: The case of taxi drivers”, IZA - Institute of Labor Economics, DP No. 15677, Oct. 2022 (available at: https://docs.iza.org/dp15677.pdf, accessed 28.3.2024)

Noy, Shakked and Whitney Zhang (2023) “Experimental evidence on the productivity effects of generative artificial intelligence”, Science, Vol. 381, 14.7.2023, pp. 187–192

Sayan Chatterjee et al. (2024) "The Impact of AI Tool on Engineering at ANZ Bank", 8.2.2024 (available at: https://arxiv.org/pdf/2402.05636.pdf, accessed 13.2.2024)

Economist (2024) “The AI doctor will see you…eventually”, 27.3.2024 (available at: https://www.economist.com/leaders/2024/03/27/the-ai-doctor-will-see-youeventually, accessed 29.3.2024)

Obermeyer et al. (2019) "Dissecting racial bias in an algorithm used to manage the health of populations", Science, Vol. 366/6464, 25.10.2019, pp. 447-453 (available at: https://www.science.org/doi/epdf/10.1126/science.aax2342, accessed 29.3.2024)

Milkau, Udo (2021) “Algorithmic Credit Scoring in USA, Europe, and China - a Comparison of Developments”, Virtual Workshop on AI & FINANCE, 29. Oct./12. Nov. 2021, hosted by Prof. Dr. Katja Langenbucher and the associated ZEVEDI investigators, in: WatchIT, Nr. 3, 2021

StM.BW (2023) “Künstliche Intelligenz in der Verwaltung”, Staatsministerium Baden-Württemberg, 10.5.2023 (in German, available at: https://stm.baden-wuerttemberg.de/de/service/presse/meldung/pid/kuenstliche-intelligenz-in-der-verwaltung, accessed 26.7.2023)

Carsten Maple et al. (2024) "The Impact of Large Language Models in Finance: Towards Trustworthy Adoption", The Alan Turing Institute, 2024 (https://www.turing.ac.uk/sites/default/files/2024-04/alan_turing_the_impact_of_large_language_models_in_finance_towar.pdf, accessed 9.4.2024)

Shijie Wu et al. (2023) "BloombergGPT: A Large Language Model for Finance", arXiv, 21.12.2023 (available at: https://arxiv.org/pdf/2303.17564.pdf, accessed 29.3.2024)

Xianzhi Li et al. (2023) "Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks", Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, Association for Computational Linguistics, 6-10.12.2023, pp. 408–422 (available at: https://aclanthology.org/2023.emnlp-industry.39.pdf, accessed 29.3.2024)

EU (2019) "EU Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC" (available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%3A32019L0790, accessed 29.3.2024)

Artist Rights Alliance (2024) "Artists' Bill of Rights", April 2024 (available at: https://artistrightsalliance.org/, accessed 4.4.2024)

Fenwick & West (2013) "Google Wins Summary Judgment in Books Case", Fenwick & West LLP, 15.11.2013 (available at: https://www.jdsupra.com/legalnews/google-wins-summary-judgment-in-books-ca-23867/, accessed 4.4.2024)

DFS (2021) "DFS Issues Findings on the Apple Card and its Underwriter Goldman Sachs Bank", New York State Department of Financial Services, March 23, 2021 (available at: https://www.dfs.ny.gov/reports_and_publications/press_releases/pr202103231, accessed 22.3.2021)

Ivana Bartoletti and Raphaële Xenidis (2023) "Study on the impact of artificial intelligence systems, their potential for promoting equality, including gender equality, and the risks they may cause in relation to non-discrimination", Council of Europe, GENDER EQUALITY COMMISSION (GEC), Aug. 2023 (available at: https://rm.coe.int/study-on-the-impact-of-artificial-intelligence-systems-their-potential/1680ac99e3, accessed 30.3.2024)

EACB (2024) "3 Questions to Mr Axel Voss", Interview, Newsletter - EACB Monthly Interviews - n°69 - March 2024, 28.3.2024

European Parliament (2024) "Artificial Intelligence Act", P9_TA(2024)0138, 13.3.2024 (available at: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf, accessed 30.3.2024)

EU (2023) "Directive (EU) of the European Parliament and of the Council of 18 October 2023 on credit agreements for consumers and repealing Directive 2008/48/EC", 18.10.2023 (Consumer Credit Directive, CCD; available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202302225&qid=1699861249729, accessed 25.11.2023)

Destatis 82024) "Gender Pay Gap 2023: Frauen verdienten pro Stunde 18% weniger als Männer", Statistisches Bundesamt, Pressemitteilung Nr. 027, 18.1.2024 (available at: https://www.destatis.de/DE/Presse/Pressemitteilungen/2024/01/PD24_027_621.html, accessed 30.3.2024)

Milkau, Udo (2023) "Algorithmic Credit Scoring as a High-Risk Issue?, Journal of Digital Banking", Vol. 7/3, 3.4.2023, pp. 249–265

Milkau, Udo (2022) "Artificial Intelligence im Credit Scoring: Lernfähigkeit oder statistische Fit-Funktion?", BIT (Banking and Information Technology), Issue 2/2022, 2.11.2022, pp. 23-37

CFPB (2019) "Interagency statement on the use of alternative data in credit underwriting", Consumer Financial Protection Bureau, Dec. 3, 2019 (available at: https://files.consumerfinance.gov/f/documents/cfpb_interagency-statement_alternative-data.pdf, accessed 22.3.2021)

Gambacorta, Leonardo et al. (2019) "How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm"', BIS Working Papers, No. 834, Dec. 19, 2019 (available at: https://www.bis.org/publ/work834.htm, accessed 5.4.2021)

Kleinberg, Jon, Sendhil Mullainathan and Manish Raghavan (2017) "Inherent Trade-Offs in the Fair Determination of Risk Scores", 8th Innovations in Theoretical

Computer Science Conference, ITCS 2017 (available at: https://drops.dagstuhl.de/opus/volltexte/2017/8156/pdf/LIPIcs-ITCS-2017-43.pdf, accessed 30.3.2024)

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan (2018) „Algorithmic Fairness“, AEA Papers and Proceedings 2018, Vol. 108, pp. 22–27

Kleinberg, Jon, Jens Ludwig Sendhil Mullainathan, and Cass R. Sunstein (2019) „Discrimination In The Age Of Algorithms“, NBER Working Paper No. 25548 (available available at:https://www.nber.org/papers/w25548.pdf, accessed 1.7.2019)

Celeste, Cameron et al. (2023) "Ethnic disparity in diagnosing asymptomatic bacterial vaginosis using machine learning", npj Digital Medicine, Vol. 6, Art.: 211, 17.11.2023 (available at: https://www.nature.com/articles/s41746-023-00953-1.pdf, accessed 1.4.2024)

Eisenstein, Sergei (1998) "The Eisenstein Reader", edited by Richard Taylor, translated by Richard Taylor and William Powell, first published in 1998 by the British Film Institute, Bloomsbury Publishing, 25.07.2019

Dorries, Nadine (2022) “Establishing a pro-innovation approach to regulating AI”; Presented to Parliament by the Secretary of State for Digital, Culture, Media and Sport by Command of Her Majesty, 18.7.2022 (available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1092630/_CP_728__-_Establishing_a_pro-innovation_approach_to_regulating_AI.pdf, accessed 1.4.2024)

Levy, Steven (2023) "How Not to Be Stupid About AI", Interview with Yann LeCun, Wired - Backchannel, 22.12.2023 (available at: https://www.wired.com/story/artificial-intelligence-meta-yann-lecun-interview/, accessed 3.4.2024)

BSI (2023) "Indirect Prompt Injections - Intrinsische Schwachstelle in anwendungsintegrierten KI-Sprachmodellen", CSW-Nr. 2023-249034-1032, 18.07.2023 (available at: https://www.bsi.bund.de/SharedDocs/Cybersicherheitswarnungen/DE/2023/2023-249034-1032.pdf, accessed 9.4.2024)

BSI (2024) "Generative KI-Modelle - Chancen und Risiken für Industrie und Behörden", 27.3.2024 (available at: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Generative_KI-Modelle.pdf, accessed 9.4.2024)

LeCun, Yann (2023), Post on X, 1.4.2023 (available at: https://twitter.com/ylecun/status/1642205736678637572?lang=de, accessed 3.4.2024)

Heikkilä, Melissa (2023) "Four trends that changed AI in 2023", MIT Technology Review - The Algorithm, 19.12.2023 (available at: https://www.technologyreview.com/2023/12/19/1085696/four-trends-that-changed-ai-in-2023/, accessed 3.4.2024)

Bengio, Yoshua et al. (2023) "Pause Giant AI Experiments: An Open Letter", Future of Life Institute, 22.3.2023 (available at: https://futureoflife.org/open-letter/pause-giant-ai-experiments/, accessed 1.4.2024)

Altman, Sam, Greg Brockman, and Ilya Sutskever (2023) "Governance of superintelligence", OpenAI, 22.5.2023 (available at: https://openai.com/blog/governance-of-superintelligence, accessed 1.4.2024)

Hinton, Geoffrey, Yoshua Bengio, Demis Hassabis, Sam Altman, et al. (2023) "Statement on AI Risk", Center for AI Safety, 31.5.2023 (available at: https://www.safe.ai/statement-on-ai-risk, accessed 1.4.2024)

Cohen, Michael K., Noam Kolt, Yoshua Bengio, Gillian K. Hadfield, Stuart Russell (2024) "Governance frameworks should address the prospect of AI systems that cannot be safely tested", Science, Vol. 384/6691, 5.4.2024, pp 36-38

Lyre, Holger (2024) “ChatGPT versteht es”, Frankfurter Allgemeine Zeitung, 3.4.2024, p. N2; based on “’Understanding AI’: Semantic Grounding in Large Language Models”, arXis, 16.2.2024 (available at: https://arxiv.org/abs/2402.10992, accessed 3.4.2024)

Schmidhuber, Jürgen (1990) "Making the world differentiable: On using self-supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments", 1990 (available at: https://people.idsia.ch/~juergen/FKI-126-90_(revised)bw_ocr.pdf, accessed 4.4.2024)

Ha, David and Jurgen Schmidhuber (2018) "World Models", arXiv, 9.5.2018 (available at: https://arxiv.org/pdf/1803.10122.pdf, accessed 4.4.2024)

LeCun, Yann (2022) "A Path Towards Autonomous Machine Intelligence", Version 0.9.2, 27.6.2022 (available at: https://openreview.net/pdf?id=BZ5a1r-kVsf, accessed 4.4.2024)

McChrystal, Stanley and Anna Butrico (2021) "Risk: A User's Guide", Portfolio/Penguin Random House, 5.10.2021, New York, USA

Milkau, Udo (2020) “Banken am digitalen Scheideweg - Verharren in der Vergangenheit oder Mut zur Zukunft?“, Fritz. Knapp Verlag, Dez. 2020

Léon Bottou (2014) "From machine learning to machine reasoning", Machine Learning, Vol. 94, pp. 133–149 (available at: https://link.springer.com/content/pdf/10.1007/s10994-013-5335-x.pdf, accessed 5.4.2024)

Yue, Xiang et al. (2023) "MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI", arXiv, version 3, 21.12.2023 (available at: https://arxiv.org/pdf/2311.16502.pdf, accessed 19.4.2024)

van der Aalst, Wil M.P. (2011) "Process Mining - Discovery, Conformance and Enhancement of Business Processes", Springer, Heidelberg

Wahlster, Wolfgang (2015) "Künstliche Intelligenz ist besser als natürliche Dummheit", Interview, Computerwoche, Nr. 23/2025 (available at: https://www.dfki.de/fileadmin/user_upload/DFKI/Medien/News_Media/Presse/Presse-Highlights/Computerwoche-Interview-Wahlster-2015-cw23-s-s014.pdf, accessed 4.4.2024)

About the author

Udo Milkau, Digital Counsellor, Frankfurt, Germany, udo.milkau@web.de

Udo Milkau is a 'digital dinosaur' with first experiences in digital technology in 1974, many innovation projects including the first European securities online-brokerage in 1995 and working as a Digital Counsellor now. For three decades he holt management positions with automotive industry, professional services firms, and transaction banking, served customers in Asia and Europe, the European banking industry, and was Chief Digital Officer, Transaction Banking until 2020. After his academic education in physics, he worked as a research scientist in large collaborations at different European research centres incl. CERN, CEA de Saclay, and GSI.

He was chairman of the European Association of Co-operative Banks (EACB) Digital and Data Working Group, member of the EACB Payment Services Working Group and member of the European Central Bank's Operation Managers Group (ECB OMG). And he is member of the Editorial Board of the Journal of Digital Banking.

Udo Milkau published more than 120 papers on digitalization of banking, risk management / risk culture, digital economies, and law & digitalization. He lectured at Goethe University Frankfurt am Main, Frankfurt School of Finance and Management, and is currently lecturing at Baden-Wuerttemberg Cooperative State University (DHBW in Mosbach).

He is also the author of the books ‘Banken am digitalen Scheideweg‘, ‘Operational Resilience in Finanzinstituten‘, ‘Decentralized Finance and Tokenization‘ und ‘Risiken jenseits wiederholter Spiele”.