THE CAPCO INSTITUTE

JOURNAL

OF FINANCIAL TRANSFORMATION

ORGANIZATION

The risks of artificial intelligence used for decision making in financial services

UDO MILKAU

NEW WORKING PARADIGMS

#52 JANUARY 2021

THE CAPCO INSTITUTE

JOURNAL OF FINANCIAL TRANSFORMATION

RECIPIENT OF THE APEX AWARD FOR PUBLICATION EXCELLENCE

CONTENTS

LEADERSHIP

WORKFORCE

ORGANIZATION

DEAR READER,

Welcome to edition 52 of the Capco Institute Journal of Financial Transformation.

Transformation has been a constant theme in our industry for several decades, but the events of 2020 have accelerated change in employee working patterns, and in the very nature of the workplace itself. This Journal examines three key elements of these new working paradigms – leadership, workforce, and organization.

As we explore in this edition, a key part of any firm's transformation agenda centers around digital leadership and how to tackle the novel challenges created by changes within organizations and society. Leaders need advanced organizational skills to build teams that use digital technologies, as well as to inspire millennial workers who have grown up in a digitally transformed world. They also need deeper technology skills to lead, and a broader understanding of the ethical paradigms introduced by the challenges created through new technologies such as AI. These enhanced skillsets will help today's leaders and their teams fully realize the benefits of new working models.

The topics reviewed in this Journal offer flexibility for employees, increased agility for teams, and a combination of both for organizations. When supported by the right technology, these can create collaborative, outcome-driven environments. Through the resulting remote or hybrid models, organizations can transform their workforce and operations to boost productivity, cost effectiveness and employee engagement, while enhancing resilience and customer experiences.

As always, our contributors to this Capco Journal are distinguished, world-class thinkers. I am confident that you will find the quality of thinking in this latest edition to be a valuable source of information and strategic insight.

Thank you to all our contributors and thank you for reading.

Lance Levy, **Capco CEO**

THE RISKS OF ARTIFICIAL INTELLIGENCE USED FOR DECISION MAKING IN FINANCIAL SERVICES

UDO MILKAU | Digital Counsellor, European Association of Co-operative Banks (EACB)[1]

ABSTRACT

The risks associated with the use of artificial intelligence (AI) have captured the attention of research, regulation, and industry practitioners in recent years. Given that this is a vast topic in its own right, we are using the experiences of the financial services industry, in specific credit scoring, as a proxy for some of the salient features of AI from a sociotechnical perspective. Although it shares some of the operational risk challenges associated with other technologies, a model for decision making reveals how the interfaces with the social context create two new types of risk: naiveté in the use of data for training AI as a statistical classifier and perceptions of the stakeholders regarding its societal implications. While the first can – and has – to be mitigated by increased literacy within an active internal risk management, the latter requires building trust.

1. INTRODUCTION: ARTIFICIAL INTELLIGENCE AS SOCIOTECHNICAL SYSTEM

The European Commission's white paper on AI [European Commission (2020a)] begins with a remarkable introduction: "Artificial Intelligence is developing fast. It will change our lives by improving healthcare […] At the same time, Artificial Intelligence (AI) entails a number of potential risks, such as opaque decision making, gender-based or other kinds of discrimination, intrusion in our private lives or being used for criminal purposes."

It is quite unique that benefits and risks are mentioned side by side from the onset, as compared to other political initiatives such as quantum technology or blockchain, which means that AI is put in the same category as other "high-risk technologies", such as genetically modified organisms, predominantly because of its sociotechnical implications.

The sociotechnical implications of AI demand a wider interpretation of its risks – especially when AI is used in decision making – beyond just operational (including criminal actions) and model risks (e.g., of credit risk models). Aligned with the discussions on sociotechnical safety by Aven and Ylönen (2018), we suggest that because of the risks associated with using AI, a holistic perspective, including the social implications of discrimination, is needed, as well as a recognition that complex (sociotechnical) systems can never be fully predicted and controlled.

One issue needs to be highlighted from the onset, as it has relevance in this context, and that is that it is quite shameful that we still have systemic racism, discrimination[2], antisemitism, and other 'isms' in the 21st century! Nevertheless, we have to analyze the process of decision making and the role of technology to understand how this process can exacerbate the situation, and to distinguish between freedom in an open diverse society and unequal treatment due to systemic discrimination, which violates equal rights and human dignity.

2. TWO ANTAGONISTIC EXAMPLES

The first example is "COVID-Net": an artificial neural network (ANN) designed for the detection of COVID-19 cases from chest radiography images [Wang and Wong (2020)]. The application

[1] I would like to thank Katja Langenbucher and Hans-Christian Boos for their very helpful comments and advice. The views expressed are those of the author and do not in any way represent those of the organizations he is associated with.

illustrates the possibilities of using AI for processing medical images, but also points out certain essential conditions:

- Typically, radiography images are taken under standardized conditions and with equivalent technical devices, which provides comparable images for a given scope.
- The medical images are labeled by experts (COVID-19 – other infections – no observations) according to the best existing human knowledge, and the trained ANN provides a statistical classification for each new case (COVID-19 – other diseases – no diagnosis) within the limits of statistical predictability.
- The tool neither "learns" nor "decides", but it makes a classification of a new image within the existing scope to support human decision making. Importantly, it is not "portable" to other scopes.

Pattern recognition is an archetype for ANN. A recent meta study [Xiaoxuan et al. (2019)] analyzed the diagnostic performance of AI tools versus the performance of healthcare professionals. The analysis showed that the pooled sensitivity (i.e., to correctly diagnose the disease) was 87.0% for AI and 86.4% for healthcare professionals, while the pooled specificity (i.e., the ability to accurately exclude patients who do not have a disease) was 92.5% and 90.5%, respectively. The results illustrate that AI can "emulate" the ability of human professionals for classification of medical images with a similar degree of accuracy.

As AI for pattern recognition has to be trained with labeled data derived from human experience, AI can automate examination and substitute human experts in places where healthcare professionals are not available. In many cases – from the COVID-19 pandemic to places where there are no medical staff for hundreds of miles – technical automation is more than welcome. However, AI can merely "copy" human experience in well-controlled circumstances [for an up-to-date overview of AI in general refer to Chowdhary (2020)].

While the first example presented the "technical prerequisites" for the correct use of AI (and the risks, if ignored), the second example highlights the "social implications" of decision making and the perceptions of the stakeholders regarding outcome.

In a *gedankenexperiment*, a stylized case is assumed with a simple algorithm, which can be executed by human (according to a manual) or technical agents (programmed):

- A European bank decides about new consumer loans solely based on the parameter "free average income" in relation to the required monthly repayment (other data such as the credit history of the borrower are not used for simplification).
- If (free monthly income) > (required monthly repayment + defined threshold) then loan is approved; else not.
- Explicitly, the bank neither processes nor stores sensitive date like "gender" in compliance with the European General Data Protection Regulation (EU-GDPR Art. 9), which prohibits data processing based on a natural person's sexual orientation[3] [European Parliament, 2016].
- 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), while on the other hand, it is obligated to assess the financial capabilities of the borrowers. The European Banking Authority emphasized that: "Creditworthiness assessment is important to avoid building up excessive risk and to embed responsible lending and borrowing practices, for both consumers and institutions" [EBA (2019)].
- In Germany (as in most countries), women have a lower average income; yet, the probability of approval will only differ between "women" and "men" if an external observer uses the protected sensitive data item "gender" to classify a certain sub-group.

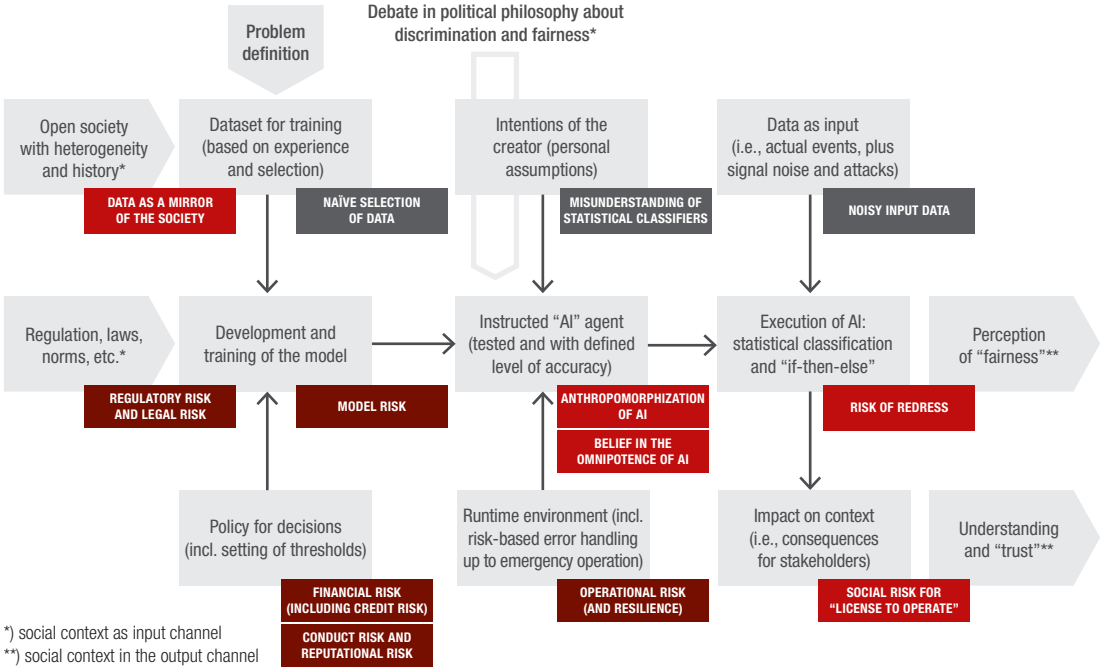
Is this algorithm discriminative? In other words: can an outcome evaluated ex-post on the basis of statistical averages of the entire population and with the use of sensitive data suggest discrimination by an individual economic agent deciding ex-ante, exclusively on the basis of objective financial data?

Unfortunately, there are no straightforward answers to this question. Langenbucher (2020), looking at the doctrines of "disparate impact" according to U.S. law or "indirect discrimination" according to E.U. law, suggested that "Under these doctrines, **intention to discriminate is not a necessary element**. Instead, a facially neutral rule or practice is under scrutiny **because of the real-world effects it**

[2] Systemic discrimination can also be documented with regards to financial inclusion. In the Annual Economic Report 2020 of the Bank of International Settlement [BIS (2020)], it was pointed out that "nearly half of Black and Hispanic US households are unbanked or underbanked" (approximately 15% unbanked and an additional 30% underbanked).

[3] EU-GDPR Art. 9/1: "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."

Figure 1: Lifecycle of an AI system embedded in its sociotechnical context with the three main steps of preparation, implementation, and execution



Adopted from Milkau and Bott (2019)

triggers when applied to a mixed group, constituted of members of a protected class and members of a not protected one" (emphasis added by author).

Scholars have proposed different legal doctrines backed by dissimilar political philosophies [Langenbucher (2020)]. One school accepts the "indirect" impact of decision making as long as there is no active evasion of law and masking prejudice. This perspective is linked to the so-called anti-classificatory theories of equality to exclude "artificial, arbitrary, and unnecessary barriers". The other school proposes corrective and redistributive methods and an obligation for lenders to accept a loss of profits to compensate for "indirect discrimination" irrespective of any causality (but based on correlations or the possibilities that "proxies" could be linked statistically to sensitive data).

It is beyond the scope of this paper to discuss these philosophies[4] but a brief comparison reveals that decision making with an impact on the social context is part of the political debate, which leaves financial institution with uncertainty about the accepted norms. This causes a new risk in balancing demographic blindness versus corrective redistribution, although the stylized decision making process is fully compliant to anti-discrimination regulations.

This issue is even more crucial if AI is supporting the decision making process. For example, with the emergence of "fairness in machine learning" [Binns (2018), MacCarthy (2019), Hu and Chen (2020)], there are expectations that machine learning comes with an obligation for equal (re-)distribution of social welfare across various social sub-groups, guided by a social planner in search of an optimum of social welfare. Such demands create dilemmas, because it would require the use of protected sensitive personal data to distinguish between sub-groups to redress historical discrimination in the society, which is prohibited under EU-GDPR and anti-discrimination legislation.

[4] The current discussion echoes an old debate about "social justice" or "distributive justice" between John Rawls and Robert Nozick in the 1970s. F.A. von Hayek (1976) pointed out that the concept of social justice (or fairness) belongs to families, warlords with retinue, or tribal societies in general, but has no meaning in a free open society. The moral idea of "dividing justly" is suitable for a birthday cake or plunder of pirates only. Concerning "economic equality" see also Harry Frankfurt's seminal essay (1987).

3. DECISION MAKING WITH AI AS A SOCIOTECHNICAL SYSTEM

A simplified model for the process of "decision making with AI" is presented in Figure 1. Independent of the implementation and whether the decision making will be undertaken by a human agent (e.g., a loan process defined in a manual including control by the 4-eyes-principle) or a technical agent (with the same algorithm programmed in software), this sociotechnical system reveals a number of risks along the *preparation* → *implementation* → *execution* sequence, with the following 3x3 steps:

1. Preparation (with definition, development, and training)
   a) Dataset for training or as benchmark (as a statistical sample)
   b) Development of the model (for classification)
   c) Policy for decisions (including setting of thresholds)
2. Implementation (with rollout to a runtime environment)
   a) Intention of the creator
   b) Instructed agent
   c) Runtime environment (including testing, risk assessment, monitoring, and resilience)
3. Execution (within the sociotechnical context)
   a) Input data
   b) Execution of "if-then-else"
   c) Impact on social context

1.a) A dataset for training of AI is needed, which has to be representative of the defined problem, but inevitably mirrors the diversity of an open society. As "data" must be generated, data – inevitably – reflect the context of their production and are never "neutral". In general, processing of personal data includes a potential legal risk concerning GDPR, as the GDPR (i) requires minimization of processing of personal data as a first principle (even if consent of the data subject was given), and (ii) is interpreted differently by national data protection authorities. Far more important, there is a new risk of a naïve use of data due to a trend to use "available" data instead of preparing "suitable" data for the specific problem (e.g., image recognition with training data taken from "public" picture databases).

1.b) A model can be based on rules with parameters, on traditional machine learning such as support vector machines (SVM), or on training of an ANN. Every model is – by definition – a hypothesis with parameters to be fitted to measured data and includes a model risk (assumptions, choice of a specific model, parametrization of the model, etc.). Ali Rahimi, a researcher in AI, argued that machine learning has become a form of "alchemy" [Hutson, 2018]. However, this is a generic problem with all sophisticated models – especially non-linear ones. Additional to technology, there is a "regulatory risk", as regulation might be non-proportional, fragmented, or inconsistent.

1.c) Every economic agent defines its individual policy for decision making, based on the freedom of contract and compliance to legislation. One example is loan origination of a bank, based on the individual financial risk management of the institution. Every lender applies its own statistical predictions of the future, including expected losses as an estimation of mean values, unexpected loss contribution due to the standard deviation from the mean, and cost of capital (based on the bank's individual balance sheet structure and rating). Consequently, credit scoring is a statistical concept of the risk-taker, and does not necessarily represent an assessment of the borrower's "worthiness" [as indicated in Hao (2019)].

2.a) The first step in implementation is an articulated human intention, which comes with subjective beliefs and bounded rationality [Simon (1991)]. While decision making has an economic rational (e.g., in credit risk management, the balance between margin and (un)expected losses), there is the danger of misunderstanding statistical classifiers. A statistical classifier can neither provide better results than the input distribution, nor be generalized beyond the defined scope.

2.b) At its core, decision making is an instructed agent, be it a human with a manual, a rule-based program, or a trained AI tool. The original description of AI, as it was presented in the Dartmouth conference of 1956, that AI "is to proceed on the conjecture that that every aspect of learning or any other feature of intelligence ... a machine can be made to simulate it" [McCarthy et al. (1955)], has unfortunately resulted in confusion, since contemporary AI is only "able to fit a function to a collection of historical data points" [Pearl and Mackenzie (2018)]. This confusion culminates in the term "self-learning", as AI systems neither act by themselves (but follow the human intention), nor learn in a human way (but are trained). Johnson (2006) suggests that computer systems do not have any intentions to act, compared to the free will of human beings. However, computer systems – and instructed agents – have intentionality, but this is the "programmed" intentionality of their designers.

2.c) An underrated element in the implementation of AI is the runtime environment. Of course, every computer program has to be tested (for executing as designed), reviewed (for correct design and use of proxies), and assessed (for potential new operational risks), as well as monitored in operation (for

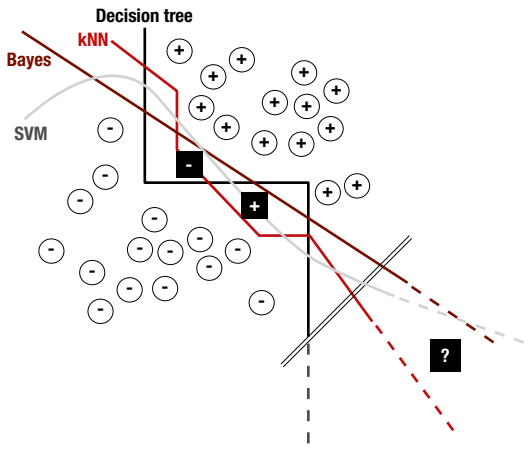
actual incidents or derivation from design parameters). Every software inevitably includes errors, suffers from the so-called "software aging" due to interdependences in the software [Parnas (1994)], could be a target of cyberattacks, or suffer from problems when AI is embedded in extended software systems [see, for example, the so-called "Uber accident" in 2018; NTSB (2019), and the debate about autonomous spacecrafts; Patel (2020)]. Depending on the degree of operational risk, "error handling" could range from controlled exit via emergency operation features (run-flat tyres) to resilience (such as redundant triple systems in airplane auto pilots).

3.a) The execution of decision making starts with actual input data of various quality, which typically include "signal noise". For example, using AI for traffic sign recognition (which is quite simple compared to face recognition) could be susceptible to damage, dirt, snow, night, graffiti, manipulation, or gaming the system. Additionally, AI systems can be vulnerable to adversarial attacks [Eykholt et al. (2018)].

3.b) Execution of an AI-based decision making is – put simply – a statistical classification plus an "if-then-else" rule. While this execution of a pre-programmed decision is trivial from a technical perspective, the social awareness can be different. Decision making based on statistical classification might be regarded as unfair. This perception leads to the above-mentioned call for fairness and the risk of redress for users of AI beyond intent or casualty.

3.c) Finally, the societal impact assessment of decision making (i.e., the consequences for all stakeholders) requires trust in the decision making process based on understanding. As Ebers (2020) stated: "Algorithms also play an increasing role in making substantive decisions. Many important decisions which were historically made by people are now either made by computers or at least prepared by them. [...] Some algorithmic scores have existential consequences for people: They decide to an increasing extent whether someone is invited for a job interview, approved for a credit card or loan, or qualified to take out an insurance policy." In an extreme case, disappointed stakeholders could cause an "outside-in social risk", resulting in the firm losing its "license to operate".

Figure 2: Illustration of machine leaning as "statistical classifier" with the examples of SVM, kNN, Naive Bayes, and Decision Trees



This Figure is adopted from Domingos (2012)

The training data are shown as circles with + and - and new events as squares. Event at the frontiers (or hyperplanes in general) can cause estimated classifications with uncertainty depending on the selected method. New events outside the training data "?" exceed the scope of the classifiers.

The various risks indicated in Figure 1 belong to three different groups. At the bottom left, one can find (traditional) financial and non-financial risks. At the top, there are (internal) risks of AI implementation in a firm. And in the diagonal, there are the (external) social and political issues, when AI is applied in a sociotechnical context. It is beyond the scope of this paper to provide a comprehensive discussion of all aspects, but the following sections will address the main issues of AI as a sociotechnical system: naiveté concerning data, the conflict of statistics versus fairness, public perception and trust-building, and finally the issue of suitable explainability.

4. TECHNICAL RISKS OF AI AND NAIVETÉ

For the purposes of this article, I will use "machine learning" as an example of AI. Machine learning can be conventional, such as "support vector machines" (SVM), or advanced, "artificial neural networks" (ANN). To start with, I will focus on "weak AI", with the intention of solving one specific problem at a time (more advanced concepts are discussed later).

It should be noted that it is not clear how an "artificial general intelligence" (AGI) in the sense of a "general problem solver" could look like. Pearl et al. (2018) state that current AI is "able to fit a function to a collection of historical data points."

The schematic example of machine learning in Figure 2 helps to illustrate its capabilities and the limitations. Different machine learning methods with a distinct "fit function" can provide similar classifications within the scope of the training data, but could result in (i) model-dependent predictions for new events "on the edge", and (ii) doubtful estimations for events "outside the original scope".

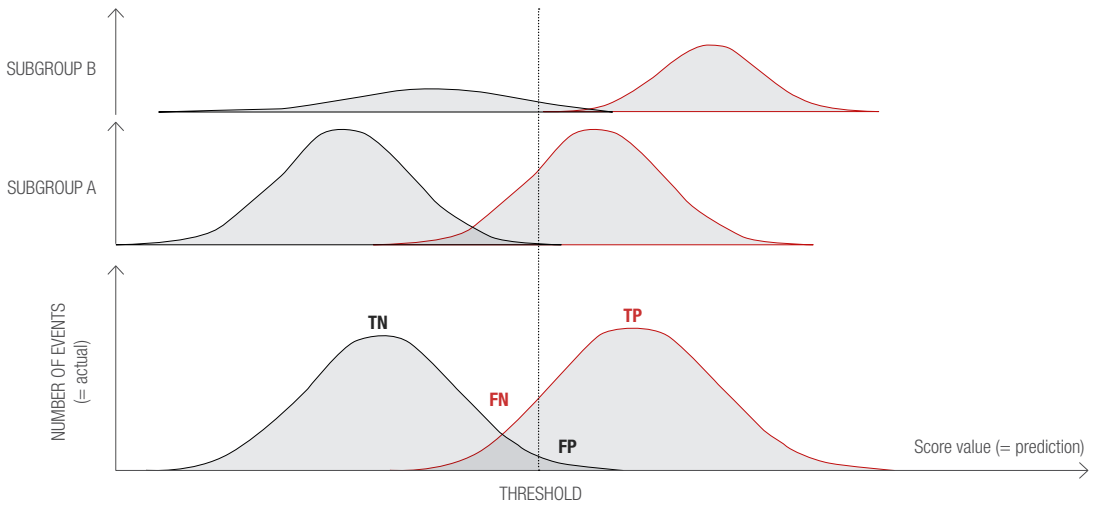
A word of warning regarding ANN, since such "neural networks" do not resemble a human brain, or even a mouse's brain, but single nerve cells, "perceptron" [Rosenblatt (1957)]. An ANN is a transformation of an input vector (e.g., enumerated pixels of an image) via a network of nodes to an output classification and has a straightforward mathematical representation [Erdmann (2020)]. Equivalent to the fitted frontiers in Figure 2, the parameters of the network are fitted to achieve an optimized classification of training data to a given label. After the training, the ANN computes straightforward classification values for a new event. If, for example, an image recognition is trained with pictures of "cats" (as 1), "dogs" (as 2), and "others" (as 0), it will classify new images as 1, 2, or 0, but does not "understand" the high-level concept, i.e., what a cat or a dog is (which cat and dog owners know well!). While ANNs with few layers and few nodes have been around for decades [Schmidhuber (2015)], "deep learning" with a nested structure of numerous layers was developed in recent years.

Unfortunately, billions of parameters exceed our ability as humans and appear to us as "black boxes". However, such a black box cannot achieve more than statistical classifications based on the original training data: classify images as 1, 2, or 0 (for cats, dogs, and other animals).

Developers seem to be keen to use "available" data for training of ANNs (e.g., image collections from social media) without checking for the context. This is different from scientific experiments, which start with a well-defined research questions, followed by the design of a "detector" for data collection, and investing much effort into the analysis of the "detector sensitivity". As all detectors have "blind" areas, researchers need to understand the detector sensitivity before data analysis can be performed. Otherwise, the results would be biased by artefacts due to active versus non-active areas of detection or by some random selection instead of a full "360 degree" perspective.

Given that AI systems must be trained with real-world data, they perpetuate a past situation to future classifications. Obermeyer et al. (2019) looked at "Dissecting racial bias in an algorithm used to manage the health of populations," which analyzed commercial prediction algorithms in the U.S. to identify and help patients with complex health needs. As this application used data about past healthcare costs of a patient (rather than a real illness) as a proxy for the needs of the patient, the data – and not the algorithm – were biased by unequal access to the U.S. healthcare system. Patients with numerous/more expensive medical treatments in the past qualified for more (predicted) preventive treatments, which reflected access to

Figure 3: Schematic distribution of positive and negative events in a population in dependence of a statistical score value with the four classes TP, TN, FP, and FN



A schematic split of the original population in two sub-groups (A and B) is shown, together with a single threshold (for the whole population). Without constraints, the distributions for the sub-groups cannot be assumed to be identical or even similar. Additionally, the variance of distributions can be broader, compared to the difference of the mean values. Any representative sample for such sub-groups has to reflect the statistical characteristics of the distribution, including the TP, TN, FP, and FN values, which do not have to match between sub-classes without additional constrictions.

healthcare and correlated to social stratigraphy, but not to actual illnesses. Additionally, the "healthcare costs" proxy strongly depends on commercial agreements and incentives (e.g., which treatment to be prescribed) [Balzter, 2018]. To avoid such bias, all "sensitivities" have to be known and taken into account: whether it is the "blind spots" of detector systems, selective recording of data due to assumptions of programmers, or our idealizing perception of the structures of society.

Cassie Kozyrkov (2019) pointed out that: "bias doesn't come from AI algorithms, it comes from people" – and from people who do not understand the context of data taking, data selection, or bias in datasets in general. This misunderstanding, ignorance, or naiveté causes an essential risk for the application of AI in all industries, including financial services. Even experts who are very interested in development of sophisticated AI tools are unenthusiastic about the tiresome work of data quality management. This "naiveté" about the training data has to be regarded as a new category of risk for the implementation of AI in every industry, but especially in financial services.

5. STATISTICS AND FAIRNESS

While the previous section elaborated on human carelessness and naiveté, which can be monitored and managed with improved technical and legal literacy within an active internal risk management, this section will focus on the general problem of "statistics versus public expectations".

The example presented in Figure 3 will be used for the following discussion. At the bottom of Figure 3, we have a distribution of "positive" (right side) and "negative" (left side) events in a population based on a statistical score value derived as an ex-ante prediction. The actual positive (e.g., decease) and negative (e.g., no decease) events have some overlap and define four classes: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). This reflects the conditional probabilities P(ex-post actual=x | if ex-ante prediction=y), which always includes "false" predictions (from an ex-post perspective).

For an economic credit decision, a trade-off for a threshold has to be made by the lender, i.e., how many FPs a bank is willing to accept (i.e., accepted loans, but with a negative margin) versus rejecting too many FNs (i.e., lost margin). "If, and only if, additional (hidden) parameters" are used ex-post to separate the population into two sub-groups A and B, these subgroups will have different distributions, and no choice of a single threshold will provide identical metrics. The same holds true, if one uses normalized metrics in the form $m = a/(a + b)$ with $a, b \in \{TP, TN, FP, FP\}$.

One could argue that the choice of one threshold would be "unfair", as the metrics would not provide an equal "fairness measure" to all sub-groups. However, every possible "fairness measure" comes with significant shortcomings:

- Kleinberg et al. (2017) made clear that "except in highly constrained special cases, there is no method that can satisfy [… all fairness] conditions simultaneously."
- As there are various philosophical, sociological, psychological, or cultural conditions of fairness, who should be in charge of selecting the "right" one?
- Finally, the dilemma remains that pre-planned "fairness conditions" for sub-groups, which mirror the structure of the society, requires processing of sensitive personal data, which is non-compliant with the intentions of legislations against discrimination.

While no fairness measure is coherent, there is a public perception that AI should be fair. Yona (2017) states that "One immediate observation that appeared when machine learning algorithms were applied to human beings […], was that the algorithms were not always behaving 'fairly' […] sometimes resulted in algorithms that behaved in a way in which a human observer will deem unfair, often especially towards a certain minority." This perception presents a new type of risk in the sense of an external requirement of fairness, which is independent of any evidence for non-compliant behavior of the individual economic agent.

6. MARKET, MORAL, AND TRUST

It is worth repeating that as Johnson (2006) pointed out, "Computer systems [are] moral entities but not moral agents." This is a crucial synopsis of two important aspects: a warning against anthropomorphization of AI, but in parallel an emphasis on the embeddedness of sociotechnical systems. The following examples – without aiming at completeness – can demonstrate this embeddedness concerning AI and credit scoring.

Discrimination in lending is a long-known issue in the U.S., [Black et al., (1978), Ladd (1998)]. A recent meta study [Quillian et al. (2020)] suggested that "racial gaps in loan denial have declined only slightly, and racial gaps in mortgage cost have not declined at all," in the U.S mortgage market. Unfortunately, this study did not disentangle intended discrimination and

after-effects of historical inequalities. Justifiably, credit scoring demands scrutiny, and some recent studies have described the development of credit scoring in the U.S. [Kiviat (2019a), Fourcade and Healy (2017)].

Kiviat (2019b) suggested that (emphasis added by author): "For policymakers, predictive validity was necessary, but not sufficient, to establish credit scoring as fair. To fill out the picture, policymakers drew on a competing **moral framework, one in which moral deservingness indicated how the market ought to treat people.**"

Consequently, it is of paramount importance to achieve "trust as a reduction of complexity" [Luhmann (1968)]. Coyle and Weller (2020) pointed out that "If an organization is not trusted, its automated decision procedures will likely also be distrusted."

Initiatives about "trustworthy" AI are steps in the right direction. The "G20 AI Principles" [G20 (2019)] proposes that "Principles for responsible stewardship of trustworthy AI," and point out that "AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle." These principles are based on the OECD's "Recommendation of the Council on Artificial Intelligence" [OECD (2019)]. The Principles feature a combination that "include[s] freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognized labor rights" going from the human rights via existing legislation (data protection, non-discrimination) to political philosophy (including fairness, social justice), which render the Principles more conceptual than actionable.

The European Commission (2020a) propose an approach of trust by regulation. The independent High-Level Expert Group on Artificial Intelligence, which was set up by the European Commission in June 2018, provided "Ethics guidelines for trustworthy AI" [AI HLEG (2019)], which reiterates (emphasis added by the author):

- "[...] respect for human dignity entails that **all people** are treated with respect due to them as moral subjects, **rather than merely as objects to be sifted, sorted, scored**, herded, conditioned or manipulated."
- "This goes **beyond non-discrimination**, which tolerates the drawing of distinctions between dissimilar situations based on objective justifications. In an AI context, equality entails that the system's operations cannot generate unfairly biased outputs [...]."

One can only appreciate these initiatives to support trust-building in sociotechnical systems. However, they contain a hidden risk of exaggeration. There is a danger for companies to forfeit their "license to operate" if unbalanced expectations of stakeholders would be failed. The European Commission (2020b) revealed that the main concerns raised by contributors to the consultation were (i) possible breach of fundamental rights, and (ii) possible discriminatory outcomes. Give that 90% and 87%, respectively, of respondents found these issues to be either important or very important, a fundamental mistrust of the use of AI among the public has to be recognized.

7. EXPLAINABILITY AND UNDERSTANDABILITY

The European "Ethics guidelines for trustworthy AI" identified four ethical principles that must be respected in the development, deployment, and use of AI systems: respect for human autonomy, prevention of harm, fairness, and explicability. The latter is described as "Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected."

The principle of "explicability" belongs to philosophical terminology, but the guidelines clarify that transparency is composed of (i) traceability, (ii) *explainability*, and (iii) communication with a key requirement: "Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned […]"

It is important that the requirement for suitable explanation focuses on the entire process (not on a single tool or method) and on communication adapted to the target audience. Likewise, there are different levels of explainability: a global explainability of a model (e.g., for an auditor or a supervisor) and a local one for an individual decision (typically for a consumer or a patient).

Suitable explainability requires "understandability" by stakeholders. In human communication, we do not interpret models by formulas, but explain our decisions. The way a doctor explains the result of a diagnosis (ex-post) and the reasons for a therapy (ex-ante) helps to build trust, because the explanation could be understood by the patient and covers

the whole process. Schematically, transparent communication for credit scoring may possibly be composed in the following way:

- Credit score value with **weighted elements from different sources** covering different time ranges, such as z% current free monthly income, y% aggregated credit history, x% payment pattern of last months,[5] <w% of other data (e.g., employee/freelancer/retired), and no use of sensitive/protected data.
- Threshold for score value represents, for example, an average 30% rejection rate (**also a protection for consumers** with high debt-to-incomes ratio not to run into excessive indebtedness).
- Additional **checks for statistical outliers** – not fitting into the typical distribution within a confidence interval – independent of whether the classification was made by a human or a computer software (so-called "yellow" cases with some ambiguity).
- In the case of a rejection, a suggestion for **possible social support** by governmental promotional banks, social benefit programs, etc.

Such an approach could be a starting point for discussion between financial institutions and regulators to explain that:

1) There is no significant criticality for the decision making process, as independent input data from multiple sources are used, while data from a credit agency would be only one element.
2) There is no discrimination, as no sensitive personal data are used. Only economic criteria for prediction of the individual financial situation of the borrower are applied.
3) Support is provided in the case of a rejection, which could provide help for people with financial problems by the society.

A combination of different types of data can improve the predictive power of a model, as analyzed in a recent Bank of International Settlement (BIS) working paper [Gambacorta et al. (2019)] using leading transaction-level data from a fintech company in China. In this case, traditional information (credit card information) and non-traditional information (usage of mobile apps and e-commerce) were combined.

However, the approach of credit scoring in China is a debated issue. While there is a lot of discussion about (governmental) social scoring systems in China [Matsakis (2019)], less information is available in English media about the financial credit scoring in China. Ant Financial started in 2015 with the "sesame score" [Ant Financial (2015)] based on (emphasis provided by the author):

- "Credit History reflects a **user's past payment history and indebtedness**, for example credit card repayment and utility bill payments.
- Behavior and Preference reveals a user's **online behavior on the websites they visit, the product categories they shop**, etc.
- Fulfillment Capacity shows a user's ability to fulfill his/her contract obligations. Indicators include use of financial products and services and **Alipay account balances**.
- Personal Characteristics examine the **extent and accuracy of personal information**, for example home address and length of time of residence, mobile phone numbers, etc.
- Interpersonal Relationships reflect the online characteristics of a **user's friends and the interactions between the user and his/her friends.**"

The first and the third element resemble financial scores discussed above, the second and the fourth are typical for online merchants (but unusual in the combination of financial data and shopping history), and the last element (behavior in social media) seems dubious from a conservative perspective. Nevertheless, the second major payment system in China, Tencent's WeChatPay, recently announced its own competing credit score system [Gill (2020)], which is to be based on consumers' personal and credit records, as well as "habits", such as their behavior as players of online games – one of the traditional business lines of Tencent. This development raises questions about the boundary between financial credit scoring and behavioral social scoring.

[5] Recently, U.S. agencies published an interagency statement on the use of alternative data in credit underwriting [CFPB (2019)] and pointed out that: "Improving the measurement of income and expenses through cash flow evaluation may be particularly beneficial for consumers who demonstrate reliable income patterns over time from a variety of sources rather than a single job. Cash flow data are specific to the borrower and generally derived from reliable sources, such as bank account records, which may help ensure the data's accuracy."

8. EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

Adequate explanations and clear responsibility of the decision-makers are the cornerstones for building trust among all stakeholders into a new technology like AI. This communication has to be "non-technical" and take people's potential fear of the technology into account. In the case of medical diagnosis, as mentioned earlier, there is a significant difference between whether the results need to be explained to computer experts at the level of "pixels", whether the classification of histologic patterns should be visualized and annotated to a pathologist [Wei et al. (2019)], or whether the diagnosis and therapy should be explained to the patient by a doctor.

The broader usage of AI has increased the demand for "explainable AI" [XAI; Gunning et al. (2019)]. Samek et al. (2019) provide an excellent overview of XAI and an introduction to different concepts. It is beyond the scope of this paper to discuss the technical aspects of XAI, but two remarks are important. Firstly, current XAI tends to be focused on image processing with deep learning. Secondly, the different XAI concepts, such as LIME [Ribeiro et al. (2016)], LRP [Bach et al. (2015)], GAM [Selvaraju et al. (2017)], and TSViz [Siddiqui et al. (2020)], require in-depth technical knowledge and are hardly suitable for communications with consumers or patients.

Advanced XAI concepts like "spectral relevance analysis" [SpRAy; Lapuschkin et al. (2019)] are able to provide meta-explanations. Such approaches can help to evaluate the reliability of the training data by back-tracing classifications to input patterns. For example, analysis has revealed [HHI (2019)] that AI tools might apply unreliable approaches. Although a majority of images could be classified correctly, a tool can lack reliability when context determines the outcome. For example, "ships" were classified due to surrounding water, "trains" due to railways, or "horses" due to copyright watermarks on the images (as training pictures with horses came from a source with such watermarks).

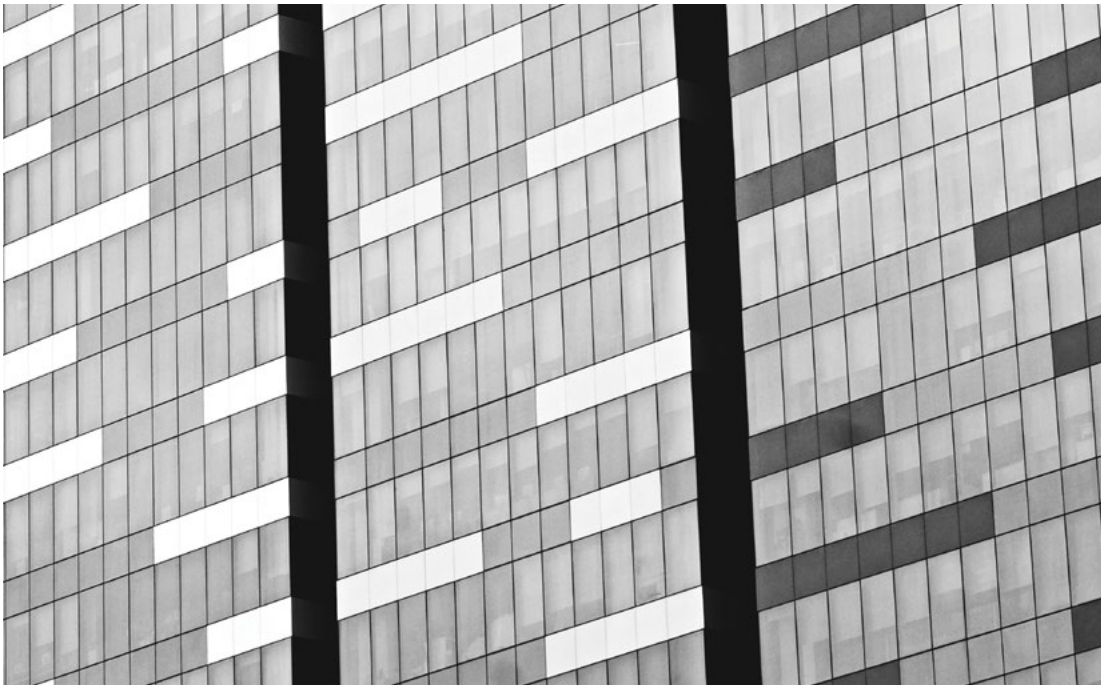
9. A REMARK ABOUT AI BEYOND MACHINE LEARNING

In this article, AI was limited to analysis of machine learning. We have to acknowledge that we neither have any idea what human (natural) intelligence really is, nor how we could emulate it as artificial intelligence. Nonetheless, the technology of AI is a huge toolbox with different methods from "expert systems" of the 1960s to computer vision, robotics, and autonomous vehicles today. Taking AI as a synonym for machine "learning" – as developed in a combination of training data plus chosen method – excludes advanced approaches of AI, which do not depend on tremendous amounts of data. One example is "machine reasoning", which was previously defined by Kaplan et al. (1988) as "computer systems that emulate reasoning tasks by using an 'inference engine' to interpret encoded knowledge of human experts stored in a 'knowledge base'."

Neither the first-generation expert systems (typically programmed in PROLOG or LISP), nor the second approach of Kaplan et al. (1988) (still based on programmed structures) were successful, but there has been a recent renaissance of this idea. With semantic graphs, an atomic piece of knowledge (a "knowledge item" consisting of factual knowledge about the environment, "situational knowledge" about the conditions under which it should be triggered, and "actionable knowledge" about what should be done) can be stored in a "knowledge base" with semantic relations. A graph-based inference engine can use such knowledge items to solve a certain problem, or to derive a new solution based on a new combination of given knowledge items.

> *"The critical success factors for sustainable implementation of AI are awareness about sociotechnical complexities and suitable communication to external stakeholders."*

Machine reasoning is an exception to the current machine learning and is applied for selected use-cases [Boos (2018)], such as automation of IT processes and/or incident handling. Nevertheless, machine reasoning depends on the knowledge items provided by human subject matter experts as an input. Based on this given "experience", the inference engine is capable of linking single knowledge items and of finding new combinations to solve novel problems. Other examples for advanced AI concepts are "causal inference" [Pearl (2016)] and "curiosity-driven learning". The latter was developed by Jürgen Schmidhuber and his co-authors [Kompella et al. (2012)] and Pierre-Yves Oudeyer and his co-authors [Colas et al. (2019)], and applies the concept of "embodied cognitive neuroscience", which states that cognitive processes depend on mind and body as a single entity and origin in an organism's sensory motor experience.



10. FROM STATISTICAL CLASSIFIERS TO A PROXY IN A POLITICAL DISCUSSION

Applications of AI are "statistical classifiers" – with few exceptions. Machines execute predefined programs, while humans make decisions about the uncertain future under conditions of bounded rationality and, consequently, have commercial, legal, and moral responsibility and accountability.

When a bank (as risk-taker) decides – usually with a threshold parameter for "if-then-else" – not to approve a consumer loan based on economic criteria, because the borrower cannot be expected to repay the loan, it is their responsibility to reject the loan. If the society decides that some sub-group in the society suffer from historical discrimination, the society can decide for (tax-paid) redistribution to this sub-group, e.g., with social benefit programs or with guarantees by a governmental promotional bank.

However, there is a subtle change of paradigm from AI being a tool for statistical classification towards AI as a proxy for a fundamental debate about responsibility and accountability. The changing perception was exemplified by Zuiderveen Borgesius (2018): "Most non-discrimination statutes apply only to discrimination on the basis of protected characteristics, such as skin colour. Such statutes do not apply if an AI system invents new classes, which do not correlate with protected characteristics, to differentiate between people. Such differentiation could still be unfair, however, for instance when it reinforces social inequality." "Suppose, for instance, that poorer people rarely live in the city centre and must travel further to their work than other employees. Therefore, poorer people are late for work more often than others because of traffic jams or problems with public transport. The company could choose "rarely being late often" as a class label to assess whether an employee is "good". But if people with an immigrant background are, on average, poorer and live further from their work, that choice of class label would put people with an immigrant background at a disadvantage [...]"

Zuiderveen Borgesius (2018) highlights a shift from autonomy and individual responsibility (of an employee to arrive in time according to the agreed employment contract) to a notion of unfairness based on correlations in the population (by an employer in a performance assessment of the agreed employment contract).

There is a danger for users of AI to get trapped into a political discussion between the traditional nexus of freedom of contract, and individual decision making, responsibility and accountability, and the demand for ex-ante planned outcome with an obligation for individual economic agents – such as bank lenders – to be made responsible to redress[6] historical developments of society.

11. CONCLUSION

As Rosa et al. (2014) stated, it is important to integrate "the lofty whiteness of risk society theory with the sooty details of risk decision making." Perceptions of external stakeholders should be taken into account, as it can result in an increase in the "risks of using AI." In this special sense, a perspective of constructivism helps, as stated by Beck (1986): "because risks are risks in knowledge, perceptions of risks and risk are not different things, but one and the same." People might be concerned that "autonomous machines" could degrade humans to pure objects (as in science fiction movies from Colossus to Terminator), or that "self-learning" AI could amplify existing discrimination in the society. These perceptions of risks by external stakeholders must be taken seriously. With this in mind, perceived risks can construct actual risks for users of technology to lose their "license to operate" in a society fragmenting into identitarian sub-groups.

These outside-in "social risks" arise from external actions of stakeholders and can be triggered by seemingly innocuous decisions – e.g., concerning the use of AI tools – if not communicated effectively.

The suggested model of decision making with AI illustrates how decision making is ingrained within the sociotechnical context and reveals the importance of the end-to-end process from assumptions to the social impact. While programming and usage of data can, and must, be educated (ex-ante), tested (during development and roll-out), and monitored (ex-post), it would be an illusion of control to believe that the external "perception of risk" could be contained. Open communication regarding the functioning of AI tools and transparency with explanation about the decision making processes are the building blocks for mitigating this new risk, while any "security by obscurity" would contradict trust-building at its core.

Although AI requires profound knowledge of sophisticated technical tools, the critical success factors for sustainable implementation of AI in financial services are awareness about sociotechnical complexities and suitable communication to external stakeholders. More research about the aspect of communication to stakeholders concerning risk management of complex sociotechnical systems would be needed to address such new risks of using AI.

[6] The UN (2011) 'Guiding Principles on Business and Human Rights' explicitly do not require remedy, if a firm did not cause or contribute to any adverse impact [quote]: 'Where adverse impacts have occurred that the business enterprise has not caused or contributed to, but which are directly linked to its operations, products or services by a business relationship, the responsibility to respect human rights does not require that the enterprise itself provide for remediation, though it may take a role in doing so.'

REFERENCES

AI HLEG, 2019, "Ethics guidelines for trustworthy AI," High-level expert group on artificial intelligence, European Commission, 8.4.2019

Ant Financial, 2015, "Ant Financial unveils China's first credit-scoring system using online data," https://bit.ly/2E3je5A

Aven, T., and M. Ylönen, 2018, "A risk interpretation of sociotechnical safety perspectives," Reliability Engineering & System Safety 175, 13-18

Bach, S., A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, 2015, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," PLoS ONE 10/7, e0130140

Balzter, S., 2018, "Im Krankenhaus fällt Watson durch," Frankfurter Allgemeine Sonntagszeitung, June 3,

Beck, U., 1986, "Risikogesellschaft - Auf dem Weg in eine andere Moderne," Suhrkamp

Binns, R., 2018, "Fairness in machine learning: lessons from political philosophy," Journal of Machine Learning Research 81, 1–11

BIS, 2020, "III. Central banks and payments in the digital era," https://bit.ly/310HCxV

Black, H., R. L. Schweitzer, and L. Mandell, 1978, "Discrimination in mortgage lending," American Economic Review 68/2, 186–191

Boos, H.-C., 2018, "AI and the future of business," https://bit.ly/3g5kv9w

CFPB, 2019, "Interagency statement on the use of alternative data in credit underwriting," Consumer Financial Protection Bureau, https://bit.ly/3asnZSi

Chowdhary, K. R., 2020, Fundamentals of artificial intelligence, Springer

Colas, C., P. Y. Oudeyer, O. Sigaud, P. Fournier, and M. Chetouani, 2019, "Curious: intrinsically motivated modular multi-goal reinforcement learning," proceedings of the 36th International Conference on Machine Learning 2019, 9-15 June, Long Beach, CA

Coyle, D., and A. Weller, 2020, '"Explaining" machine learning reveals policy challenges," Science 368/6498, 1433-1434

Domingos, P., 2012, "A few useful things to know about machine learning," Communications of the ACM 55/10, 78-87

EBA, 2019, "EBA consumer trends report 2018/19," European Banking Authority, February 20,

Ebers, M., 2020, "Regulating AI and robotics: ethical and legal challenges," in Ebers, M., and S. Navarro, Algorithms and law, Cambridge University Press

Erdmann, M., 2020, "Deep learning," Physik Journal 19/4, 31-36

European Commission, 2020a, "A European approach to excellence and trust," white paper on artificial intelligence, https://bit.ly/3kOElnl

European Commission, 2020b, "Summary report on the open public consultation on the white paper on artificial intelligence," https://bit.ly/3atYccl

European Council, 2004, "Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services", Official Journal of the European Union, https://bit.ly/3iHH9qj

European Parliament, 2016, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation 'GDPR')," Official Journal of the European Union, https://bit.ly/3kSessr

Eykholt, K., I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, 2018, "Robust physical-world attacks on deep learning visual classification," in 2018 IEEE/CVF conference on computer vision and pattern recognition, Salt Lake City

Fourcade, M., and K. Healy, 2017, "Seeing like a market," Socio-Economic Review 15/1, 9–29

Frankfurt, H. 1987, "Equality as a moral ideal," Ethics 98/1, 21-43

G20, 2019, "G20 AI principles," G20 ministerial meeting on trade and digital economy, Tsubuka, Japan,

Gambacorta, L., Y. Huang, H. Qiu, and J. Wang, 2019, "How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm," BIS working papers no. 834

Gill, C., 2020, "WeChat launches personal credit rating for 600m users," Asia Times Financial, May 12, https://bit.ly/3keqllx

Gunning, D., M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, 2019, "XAI - Explainable artificial intelligence," Science Robotics 4/37: eaay7120

Hao, K., 2019, "This is how AI bias really happens—and why it's so hard to fix," MIT Technology Review, February 4, https://bit.ly/30Ya6rT

Hayek, F. A. von, 1976, Law legislation and liberty: the mirage of social justice, University of Chicago Press

HHI, 2019, "Der Blick in Neuronale Netze," Forschung Kompakt, Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut, https://bit.ly/3kIJ1Rk

Hu, L., and Y. Chen, 2020, "Fair classification and social welfare," FAT* '20: proceedings of the 2020 conference on fairness, accountability, and transparency, ACM, New York

Hutson, M., 2018, "AI researchers allege that machine learning is alchemy," Science, May 3, https://bit.ly/312EI14

Johnson, D. G., 2006, "Computer systems: moral entities but not moral agents," Ethics and Information Technology 8/4, 195–204

Kaplan, S. J., J. J. King, and D. Sagalowicz, 1988, "Knowledge based processor for application programs using conventional data processing capabilities," United States Patent no. 4783752

Kiviat, B., 2019a, "Credit scoring in the United States," Economic Sociology: The European Electronic Newsletter 21/1, 33-42

Kiviat, B., 2019b, "The moral limits of predictive practices: the case of credit-based insurance scores," American Sociological Review 84/6, 1134–1158

Kleinberg, J., S. Mullainathan, and M. Raghavan, 2017, "Inherent trade-offs in the fair determination of risk scores," 8th Innovations in

theoretical computer science conference (ITCS 2017), https://bit.ly/2PVaH7g

Kompella, V. R., M. Luciw, M. Stollenga, L. Pape, and J. Schmidhuber, 2012, "Autonomous learning of abstractions using curiosity-driven modular incremental slow feature analysis," 2012 IEEE International conference on development and learning and epigenetic robotics (ICDL), San Diego, CA

Kozyrkov, C., 2019, "What is AI bias?" Towards Data Science, https://bit.ly/3g1OsqX

Ladd, H. F., 1998, "Evidence on discrimination in mortgage lending," Journal of Economic Perspectives 12/2, 41–62

Langenbucher, K., 2020, "Responsible A.I. credit scoring - a legal framework," European Business Law Review 31:4, 527 – 572

Lapuschkin, S., S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, 2019, "Unmasking Clever Hans predictors and assessing what machines really learn," Nature Communications10:1096

Luhmann, N., 1968, "Vertrauen: Ein Mechanismus der Reduktion sozialer Komplexität," Enke Verlag

MacCarthy, M., 2019, "Fairness in algorithmic decision-making," The Brookings Institution's artificial intelligence and emerging technology (AIET) initiative, https://brook.gs/3fY21rD

McCarthy, J., M. Minsky, N. Rochester, and C. E. Shannon, 1955, "A proposal for the Dartmouth summer research project on artificial intelligence," August, https://bit.ly/3iEJimu

Matsakis, L., 2019, "How the west got China's social credit system wrong," Wired, July 29, https://bit.ly/3ilR3b9

Milkau, U., and J. Bott, 2019, "Decision-making with artificial intelligence in the social context," WatchIT Nr. 2, DHBW Mosbach, Germany

NTSB, 2019, "Collison between vehicle controlled by Developmental Automated Driving System and pedestrian," National Transportation Safety Board, Accident Report, NTSB/HAR-19/03, PB2019-101402

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan, 2019, "Dissecting racial bias in an algorithm used to manage the health of populations," Science 366(6464), 447-453

OECD, 2019, "Recommendation of the Council on Artificial Intelligence," Organisation for Economic Co-operation and Development, OECD/LEGAL/0449, https://bit.ly/310KYB1

Parnas, D. L., 1994, "Software aging," 16th International conference on software engineering, 1994, Proceedings, ICSE-16

Patel, N. V., 2020, "Are we making spacecraft too autonomous?" MIT Technology Review, https://bit.ly/2Y4bGH1

Pearl, J., M. Glymour, and N. P. Jewell, 2016, Causal inference in statistics - a primer, John Wiley & Sons

Pearl, J., and D. Mackenzie, 2018, The book of why, Basic Books/Hachette Book Group

Quillian, L., J. J. Lee, and B. Honoré, 2020, "Racial discrimination in the U.S. housing and mortgage lending markets: a quantitative review of trends, 1976–2016," Race and Social Problems 12, 13–28

Ribeiro, M. T., S. Singh, and C. Guestrin, 2016, "Local interpretable model-agnostic explanations (LIME): an introduction," O'Reilly online learning, https://bit.ly/3iK3qDN

Rosa, E., A. McCright, and O. Renn, 2014, Risk society revisited: social theory and governance, Temple University Press

Rosenblatt, F., 1957, "The perceptron - a perceiving and recognizing automaton," Report 85-460-1, Cornell Aeronautical Laboratory, https://bit.ly/3kTuWAS

Samek, W., G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, 2019, "Explainable AI: interpreting, explaining and visualizing deep learning," Lecture notes in artificial intelligence no. 11700, Springer International Publishing

Schmidhuber, J., 2015, "Deep learning in neural networks: an overview," Neural Networks 61, 85-117

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 2017, "Grad-CAM: visual explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International conference on computer vision (ICCV), Venice

Siddiqui, M. S. A., D. Mercier, M. Munir, A. Dengel, and S. Ahmed, 2020, "TSViz: demystification of deep learning models for time-series analysis," in Computing Research Repository eprint Journal, https://bit.ly/2FxqU0R

Simon, H., 1991, "Bounded rationality and organizational learning," Organization Science 2/1, 125–134

U.N., 2011, "Guiding principles on business and human rights - implementing the United Nations "Protect, respect and remedy" framework," United Nations, New York and Geneva

Wang, L., and A. Wong, 2020, "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images," arXiv:2003.09871v2

Wei, J. W., L. J. Tafe, V. A. Linnik, L. J. Vaickus, N. Tomita, and S. Hassanpour, 2019, "Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks," Scientific Reports 4;9:1, 3358, https://go.nature.com/348gfU6

Xiaoxuan, L., L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J. R. Ledsam, M. K. Schmid, K. Balaskas, E. J. Topol, L. M. Bachmann, P. A. Keane, and A. K. Denniston, 2019, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," Lancet Digital Health1: e271–97, https://bit.ly/3kPgJEV

Yona, G., 2017, "A gentle introduction to the discussion on algorithmic fairness," Towards Data Science, https://bit.ly/3asqpjK

Zuiderveen Borgesius, F., 2018, "Discrimination, artificial intelligence, and algorithmic decision-making," Council of Europe, Directorate General of Democracy, https://bit.ly/2Cu0Odl

ABOUT CAPCO

Capco is a global technology and management consultancy dedicated to the financial services industry. Our professionals combine innovative thinking with unrivalled industry knowledge to offer our clients consulting expertise, complex technology and package integration, transformation delivery, and managed services, to move their organizations forward.

Through our collaborative and efficient approach, we help our clients successfully innovate, increase revenue, manage risk and regulatory change, reduce costs, and enhance controls. We specialize primarily in banking, capital markets, wealth and asset management and insurance. We also have an energy consulting practice in the US. We serve our clients from offices in leading financial centers across the Americas, Europe, and Asia Pacific.

WORLDWIDE OFFICES

APAC
Bangalore
Bangkok
Gurgaon
Hong Kong
Kuala Lumpur
Mumbai
Pune
Singapore

EUROPE
Berlin
Bratislava
Brussels
Dusseldorf
Edinburgh
Frankfurt
Geneva
London
Munich
Paris
Vienna
Warsaw
Zurich

NORTH AMERICA
Charlotte
Chicago
Dallas
Hartford
Houston
New York
Orlando
Toronto
Tysons Corner
Washington, DC

SOUTH AMERICA
São Paulo

WWW.CAPCO.COM

CAPCO