watchit
Expertenforum

2. Veröffentlichung

Decision-Making with Artificial Intelligence in the Social Context

Responsibility, Accountability and Public Perception with Examples from the Banking Industry

Udo Milkau and Jürgen Bott

DHBW
Duale Hochschule
Baden-Württemberg
Mosbach

Impressum

DHBW Mosbach
Lohrtalweg 10
74821 Mosbach

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

Decision-Making with Artificial Intelligence in the Social Context

Responsibility, Accountability and Public Perception with Examples from the Banking Industry

von Udo Milkau und Jürgen Bott

Herausgeber:
Jens Saffenreuther
Dirk Saller
Wolf Wössner

Mosbach, im August 2019

Decision-Making with Artificial Intelligence in the Social Context

Responsibility, Accountability and Public Perception
with Examples from the Banking Industry

Udo Milkau and Jürgen Bott

Udo Milkau *received his PhD at Goethe University, Frankfurt, and worked as a research scientist at major European research centres, including CERN, CEA de Saclay and GSI, and has been a part-time lecturer at Goethe University Frankfurt and Frankfurt School of Finance and Management. He is Chief Digital Officer, Transaction Banking at DZ BANK, Frankfurt, and is chairman of the Digitalisation Working Group and member of the Payments Services Working Group of the European Association of Co-operative Banks (EACB) in Brussels.*

Jürgen Bott *is professor of finance management at the University of Applied Sciences in Kaiserslautern. As visiting professor and guest lecturer, he has associations with several other universities and business schools, e.g. IESE in Barcelona. He studied business administration at the Julius Echter University of Würzburg, and statistics and operations research at Cornell University (New York). He received his doctorate degree from Goethe University Frankfurt. Before he started his academic career, he worked with J. P. Morgan, Deutsche Bundesbank and McKinsey & Company. He was involved in projects with the International Monetary Fund and the European Commission (EC) and is an academic adviser to the EC, helping to prepare legislative acts or policy initiatives on banking issues.*

Abstract

One can sense a perception in public and political dialogue that artificial intelligence is regarded to make decisions based on its own will with an impact on people and society. The current attribution of human characteristics to technical tools seems to be a legacy of fictional ideas such as Isaac Asimov's "Three Laws of Robotics" and requires a more detailed analysis. Especially the blending of the terms ethics / fairness / justice together with artificial intelligence illustrates the perception of technology with human features, which may originate from the original claim of "ability to perform similar to human beings".

However, the current discussion about artificial intelligence falls short, not only due to the missing free will and own choice of those technological agents, but as decision-making lacks an overall model for the decision-making process, which is embedded in a social context. With such a process model, the question of "ethics of artificial intelligence" can be discussed with the selected issues of autonomy, fairness and explainability. No tool, algorithm or robot can solve the challenges of decision-making in case of ethical dilemmas or redress the history of social discriminations. However, the public discussion about technical agents executing pre-defined instruction has a tendency towards anthropomorphism, which requires a better insight for the chain of (i) responsibility for the decision, (ii) accountability for execution, and (iii) perception in the society.

Introduction

The vibrant discussion about "ethics of artificial intelligence" and "algorithmic fairness" in media and academic writings might veil that the key issue is the process of decision-making itself and the impact of decisions on people and the society. The current debate places special emphasis on one single element of decision-making: the "instructed agent" such as a piece of software, a robot, but even a human being with a manual. However, it dismisses the embeddedness of any decision-making in a wider social context. Anthropomorphisation of machines may fit the *Zeitgeist*, and e.g. Northeastern University's Roundtable "Justice and Fairness in Data Use and Machine Learning" discussed the relationship between Artificial Intelligence (AI) and "*[...] fundamental questions about bias, fairness, and even justice still need to be answered if we are to solve this problem*." (Northeastern, 2019). However, novel technology is never "neutral", but always Janus-faced with opportunities and risks[1].

A careful risk assessment is required for immediate and long-term consequences based on the best available knowledge[2]. However, there is a current vogue about "*bias, fairness, and even justice*" of AI and algorithms[3] (for an introduction see: Olhede and Wolfe, 2018). This exceeds quantitative risk assessment of technologies (Aven, 2012), as it blends the bottom-up approach of statistical estimations of risks with a top-down dispute about ethical and social values. According to Max Weber (1922), the question about the tangible impact of any new technology in a social context relates to "*Verantwortungsethik"* (ethic of responsibility), whereas the discussion about generally agreed believes relates to *"Gesinnungsethik"* (ethic of conviction).

In this paper, we understand ethic as the right way of and responsibility[4] for decision-making. With that background, we discuss the process of decision-making with "instructed agents" (a "written" software code, a "trained" piece of AI, or an "assigned" employee with a work manual[5]) as a basis for the general debate about opportunities and risk of algorithms and artificial intelligence.

The question of "Ethics of Machines" seems to assign some human capability to technical agents without free will and own choice. An "instructed agent" can just execute pre-defined instruction; they are not "autonomous" in the sense personal responsibility for the consequences[6]. Additionally, the imagination of autonomous agents raises the issue of "fairness", but what does fairness of a statistical classifier - such as most currently implemented AI systems - really imply?

---

[1] Even the Neolithic Revolution had a negative impact on human health (Latham, 2013).

[2] The issue of a general risk assessment of novel technologies is beyond the scope of this paper and the reader is referred to e.g. Aven (2012) and Fischhoff (2015). Any well-intended novel technology can cause public benefits such as higher security or better living, but also has the risk of inadvertent negative - and often path-dependent - consequences despite all possible human precaution.

[3] The term "algorithm", which simply means a set of instructions to deal with a problem, derives from the ninth-century scholar al-Chwārizmī - or in Latin "Algorismi".

[4] A detailed discussion about individual responsibility was given by Domènec Melé (2009).

[5] The economic rationale for a substitution of human agents doing manual work by technical agents is an interesting question, which is discussed in Ajay Agrawal (2019) looking at the "cost of prediction" in decision-making under uncertainty.

[6] A human agent can deviate from the pre-defined instructions, but for the scope of this paper, no misconduct will be assumed and legitimate instructions are to be followed.

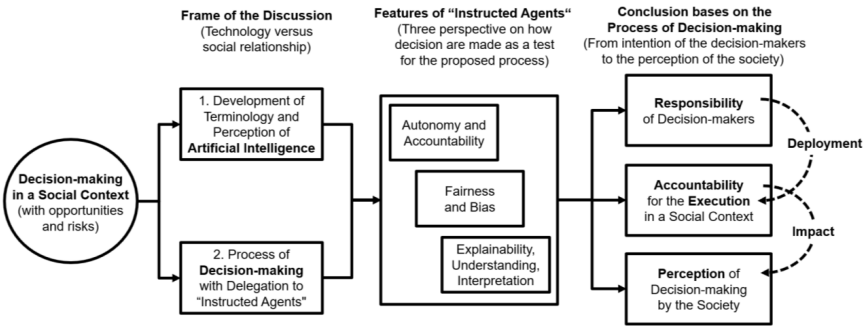

Figure 1: Decision-making in a social context with development of the proposed process model, which is tested with three different perspectives. The conclusion that decision-making required a separation between (i) the responsibility, (ii) accountability for the execution, and (iii) the perception in the society.

This paper elaborates what the basis for a discussion about an "Ethics of Machines" could be. It begins with the two starting points: the historical legacy of AI and the process of delegated execution of decisions (see Fig. 1). Three selected issues are used to test the proposed process model: the first question about the "autonomy and accountability" of a robot leads to the issue of moral dilemma in autonomous decision-making with the need to clarify where the intention comes from. The second question about "fairness and bias" leads to the social dilemma as all our experience is mirroring the actual situation of the real world with all existing unfairness and discriminations. As Cassie Kozyrkov (2019) said [quote]: "*Bias doesn't come from AI algorithms, it comes from people*." The third issue is the ambiguity of "explainability vs. understanding vs. interpretation", as the significance of an explanation depends on the background of the receiver and the social context.

Of course, every decision made by humans or machines in a commercial relationship requires explainability, but the concrete way of explainability depends strongly on the receiver of the explanation. Correspondingly, "algorithmic risk" arises as misunderstanding about the capabilities of algorithms, statistical classifiers or predefined instructions in general.

The discussion of these issues requires an intersection of different perspectives: from philosophy via technology and statistics to economics and sociology. The paper will review the relevant contributions and put them into the overall framework of a proposed process model.

This schematic model integrates the process of decision-making - with the instructed agents executing pre-defined programs with the intensions of the programmer[7] - into the social context with feedback based on the public perception. Within this paper, the example of a loan approval in a bank based on credit-scoring will be used to illustrate the discussion with a non-trivial example from the real world. The paper concludes that the current discussion about algorithms and AI tells us more about decision-making in the social context than about specifying requirements for technical AI tools.

The Development of the Term "Artificial Intelligence"

As shown in Fig. 2, the development of AI oscillated between two poles. The one side is characterised by the bottom-up concept, as introduced by Norbert Wiener already in the 1940s, based on samples of recorded data (as representation of "experience"), analysis of correlations ("patterns"), and statistical estimations ("predictions") about the future development of a system. The other side is the top-down approach represented by the claim of John McCarthy at the seminal Dartmouth Summer Research Project on Artificial Intelligence in 1956 that AI stands for a systems with an ability to perform similar to human beings.

The top-down approach dominated until the so-called "AI winter" with failure of most implementations and especially of so-called expert systems. The bottom-up approach nearly got missing after the critics of Minsky and Papert, but revitalized with the increase of computer power and data in the 2000s.

The claim of the top-down approach to "*perform similar to human beings*" jumped over figuratively to the bottom-up approach with the labelling "*learning*" for an optimising function in artificial neutral networks (ANN). Whilst the top-down approach aimed at an emulation of human "*understanding*", the bottom-up approach[8] always focussed on correlations in samples of recorded data to perform predictions for future developments (or more precisely: statistical estimations).

The latest achievement in AI with so-called "Deep Learning" caused even more confusion in the public reception. The recent example of the ability of "AlphaZero" (Campell, 2019) to train Chess or Go simply by computer programs playing games against each other (but more than ever played by humans in history) lend credence to a public imagination of an "autonomy" of such technical tools.

With few exceptions, contemporary AI executes pre-defined instructions on behalf of the original developers. If a tool is designed to "learn" chess and to win based on the "experience" of repeated games between computers, it will exactly follow those instructions. However, AI tools do not perform more "autonomy" (i.e. free will and own choice) than the famous Hollerith Machine used to tabulate the 1890 U.S. census data based on punch-cards (U.S. Census Bureau, not dated).

[7] This paper assumes that the "programmer" can be responsible and is not a hired peon via Amazon Mechanical Turk or similar tools, which would simply shift the discussion by one level.
[8] Some current scholars limit the definition to a bottom-up perspective e.g.: "*AI systems are self-training structures of ML predictors that automate and accelerate human tasks*." (Taddy, 2019)

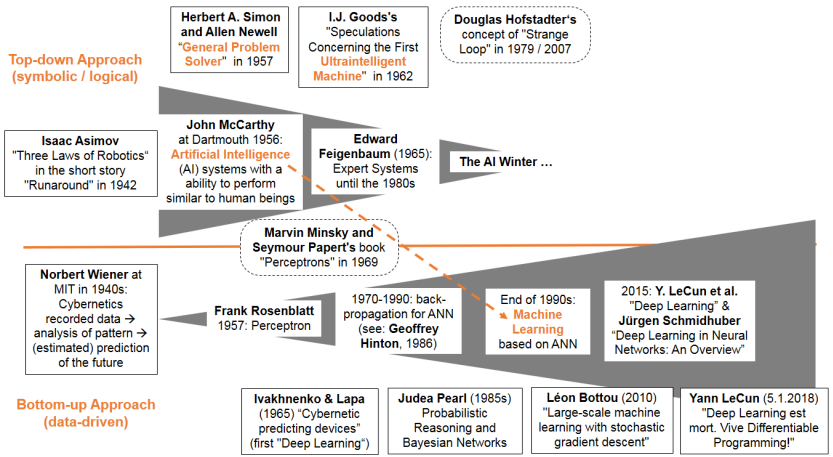

Fig. 2: Timeline of "artificial intelligence"
The book of Marvin Minsky and Seymour Papert marks a turning point towards the symbolic / logic top-down approach (see also Mikel Olazaran, 1996, about the controversy between an "official-history" versus "research-area" perception). The "AI Winter" indicated the time of disillusion with the top-down approach and especially the failure of expert system to achieve success outside narrow niches. The work of Judea Pearl and Léon Bottou based on statistics can be seen as an antagonist to a "General Problem Solver" of H.A. Simon and A. Newell or the "Speculation" of I.J. Good that in the in the twentieth century a "Ultraintelligent Machine" will be build as "the last invention that men need make". The quote by Yann LeCun about "Deep Learning est mort. Vive Differentiable Programming!" indicates some renaissance of programming instead of "learning" in the discussion about AI as statistical classifiers. Douglas Hofstadter (2007) concept of "Strange Loop" dated from 1979 and was elaborated in his book of 2007 as a proposal to explain the difference between human intelligence and machines with statistical classification: i.e. recursion compared to linear processing. Remark: ANN stands for Artificial Neutral Networks

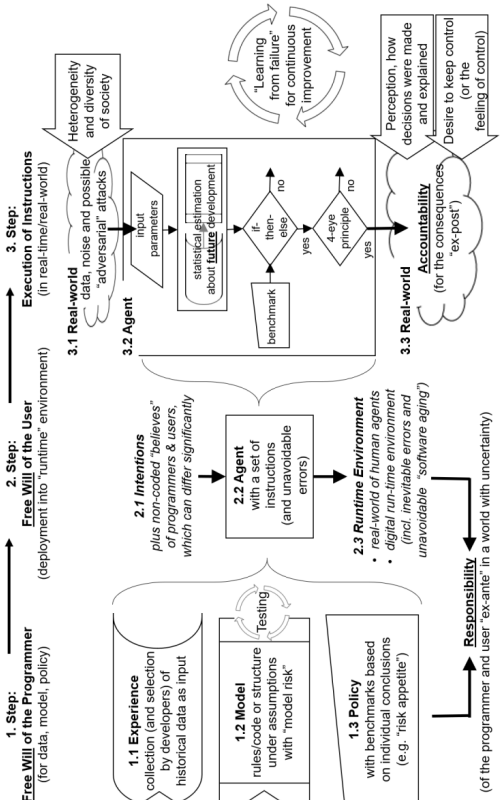

Fig. 3: The process of decision-making with delegation to instructed agents in the social context (schematic simplification, detailed explanation see text).

However, misunderstandings found the way into statement like the European Parliament resolution of 12 February 2019 on artificial intelligence and robotics (European Parliament, 2019) saying [quote, emphasis by the authors]:

"155. Believes that artificial intelligence, especially systems with built-in autonomy, including the capability … and the possibility of self-learning or even evolving to self-modify, should be subject to robust principles; …;"

As a matter of fact, another report (European Parliament, 2017) called for an "electronic personality"[9] for "autonomous robots" [quote, emphasis by the authors]:

"59. f) creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently;"

This misunderstanding underlines a deficit in communication about AI to public stakeholders, politicians and citizens. With simplification, one can apply the following hierarchy to AI for the time being:

- **Weak AI** (or Artificial Narrow Intelligence, ANI) covers nearly all contemporary AI systems, which can solve pre-defined specific problems[10] (Parikh et al., 2019). Corresponding to Pearl (2018) such systems are "able to fit a function to a collection of historical data points".
- **Universal AI** (or UAI, coined by Hutter, 2005) is capable of transferring one solution to new problems. Causal[11] Inference (see: Pearl, 2016), Machine Reasoning (see: Boos, 2018) and Curiosity-driven Learning (see: Schmidhuber and co-workers; e.g. Kompella, 2012) run in this direction, but are still software based on mathematical approaches as graph theory.
- **Cognitive Computing** "r*efers to systems that learn at scale, reason with purpose, and naturally interact with humans […] cognitive systems can make sense of the 80 percent of the world's data that computer scientists call "unstructured […] None of this involves either sentience or autonomy on the part of machines.*" (Quote taken from: Kelly, 2015) One can argue about the actuality of cognitive computing, but it is far away from "sentience or autonomy".
- **Strong AI** (or Artificial General Intelligence, AGI) in the sense of John McCarthy's "*ability to perform similar to human beings*"[12] and Simon/Newell's "*General Problem Solver*" (Newell et al., 1959) is still beyond our current approaches in computer sciences[13].

[9] As a remark, there are few examples of granting things the status of a legal person like Whanganui River: "[the Whanganui River] *Te Awa Tupua will have its own legal identity with all the corresponding rights, duties and liabilities of a legal person*." (New Zealand Government, 2017)

[10] Therefore current (weak) AI is suitable for games (with pre-defined rules), has tremendous problems with everyday situations (such as for autonomous vehicles), and struggles with decision-making under uncertainty (e.g. in a commercial context).

[11] Concerning causal models see also: Lake et al. (2016).

[12] Currently, there is a controversy on the ability of current AI system to perform creativity in art. While Elgammal (2019) stated [quote]: "Advanced algorithms are using machine learning to create art autonomously.", Ullman (2019) takes the opposite position [quote]: "If we allow [...] to treat machine "cre-

- **Superintelligence** is the title of a New York Times bestseller by Nick Bostrom in 2014, rehashing I.J. Good's speculations about "ultra-intelligent machines" of 1962, based on the simple assumption that future developments of computation power would result in what he dubs "speed superintelligence".
- **Reinstating AI** is a new concept developed by Acemoglu and Restrepo (2019) to [quote]: "*restructure the production process in a way that creates many new, high-productivity tasks for labor. […] Recent technological change has been biased towards automation, with insufficient focus on creating new tasks where labor can be productively employed. [...] but this might mean missing out on the promise of the 'right' kind of AI with better economic and social outcomes".*

Sometime in the future - or even tomorrow, if a genius will invent such a machine - AI might exhibit creativity, understanding, dreaming and even the ability to forget, but today nobody knows how to achieve those visions elsewhere than in science fiction literature.

Decision-making with delegation to instructed agents

The contemporary "*fundamental questions about bias, fairness, and even justice*" require a deeper look on the process how delegation to "instructed agents" works[14]. In Fig. 3, a process flow of decision-making is illustrated with three times three steps, which can have deviations in real implementations. Developer and user can be the same person; the "learning" can be done iteratively with new samples of data after initial deployment; or a benchmark can already be fed into the training procedure. For the scope of this paper, decision-making is regarded as a binary decision ("if-then-else") with a "yes" or a "no" and related impact on other actors[15].

ativity" as a substitute for our own, then machines will indeed come to seem incomprehensibly superior to us. But that is because we will have lost track of the fundamental role that creativity plays in being human."

[13] One way for new concepts is the development of "neuromorphic" hardware (see e.g. Davis, 2018; Ullman, 2019; or Feldmann et al., 2019).

[14] This approach skips many different definitions of "algorithms", although these discussions show some connection to decision-making as a process (see e.g.: Dourish, 2016 and Seaver, 2017).

[15] This has to be distinguished from the "choice under uncertainty" question in economics with the optimisation of the von Neumann-Morgenstern utility function with conditional probability (i.e. given a prediction signal) for a number of different choices with own utility.

The following schematic process for decision-making can be a guiding line[16]:

1. Decision-makers (programmer, developer, instructor etc.) with a free will and responsibility:

   1.1 Any decision-making requires experience or (incomplete[17,18]) knowledge about the past to estimate the future risk linked to the decision (Milkau, 2017). The collection provides a sample of data with a statistical distribution, which can be used either as input for traditional statistical systems or for training of "weak" AI[19]. The selection of data depends on the question to be decided, but also on the availability of data. These aggregated experiences are always embedded in the social context of the real-world and mirror the historical development including any bias or unfairness in a society[20]. Additionally, the collection of samples is done from the point of view of the person in charge for data collection, with a limited perspective.

   1.2 The programmer develops a model* under assumptions about the context of later operations. The models to be deployed are a result of an "experimental" search process with different potential alternatives, different parametrisations and a selection of the final model based on quality criteria[21] (analogue to "least square fit" in statistics).
   Models inevitably have residual errors not detected despite testing. No model, no (technical) system, no software and no piece of AI will be free of errors and requires appropriate calibration, testing, and review. Additionally, these models are related to the existing social norms and guidelines: from legislation and regulation to moral values of the creators.
   *) All models used e.g. in a bank are based on assumptions and show a model risk.

[16] It is beyond the scope of this paper to discuss AI tools provided by companies for their customers such as e.g. Bank of America's Erica avatar (Bank of America, 2019).
[17] For this paper, incomplete knowledge will be defined as „invincible ignorance" in the sense of Domènec Melé (2009) due to natural limitations to be distinguished from negligence or intention.
[18] It would be beyond the scope of this paper to discuss the impact of the current trend to post-factual narratives in the society and politics.
[19] Bücher et al (2017) discussed the issue that - in the perception of decision-makers - data can develop a "transition to independence" with a self-relief of decision-makers from responsibility.
[20] Self-reinforcing developments are an integral part of social systems (including e.g. a bank). Hiring new (young) job candidates based on a „fit" to the existing experience with former hiring and the current composition in staff and management comes with the problem of amplifying a trend to the average (or even to mediocrity). However, this has nothing to do with algorithms or AI tools.
[21] Recently Lapuschkin et al. (2019) proposed a new Spectral Relevance Analysis to characterize and validate the behavior of nonlinear machine learning. This assessment whether a "learned" model delivers reliable results for the original problem revealed strong dependence on the structure of the input data (i.e. hidden text information in pictures or associated structures such as "rails" as main detection element for "trains"). However, Kickingereder et al. (2019) show that machine learning methods carefully trained on standard magnetic resonance imaging (MRI) are more reliable and precise than established radiological methods in the treatment of brain tumors. This was an important first step towards automated high-throughput analysis of medical image data of brain tumors.

This requires an understanding of the model, its assumption, its limitation, and how the model will be interpreted by decision-makers. Recently, German banking supervision authority BaFin (Bartels and Deckers, 2019) proposed to run two different and independent models in parallel (e.g. a rule-based vs. an AI-based model).

1.3 Decision-makers (e.g. executives in a bank) have the responsibility to define a policy for decision-making & risk-taking (e.g. credit policy with a risk appetite of a bank*) and to specify:
1.3.1 Which variable(s) are relevant for decision-making?
1.3.2 Which benchmark (threshold to be compared to) should be used?
1.3.3. Which tolerances (accuracy or misclassication) and forecasting errors are acceptable?
*) The risk appetite specifies the accepted amount of estimated future losses (expected losses and unexpected losses) derived from statistical distributions of past defaults.

2. Deployment into the runtime environment for execution in the real-world by a "user":

2.1 The ex-ante intention of the programmer is implemented in an algorithm (manual for human workers, traditional code or a ANN), but is always associated with non-coded "beliefs" of the programmer about the future run-time environment, which consists of the technical environment but also of the social context concerning the consequences of the decision-making. Practically, the assumption of the programmers do not coincide generating "unintended" use of computer systems up to fatal accidents, when assumptions contradict.

2.2 The "instructed agent" can be either a human worker (with an "instruction manual") or a technical agent (with a set of instructions, i.e. computer code). Due to the bounded rationality of human beings (see: Simon, 1957) but likewise computers, a complete prediction of future scenarios is impossible except for trivial instructions (e.g. the famous computer program "print 'Hello World'"). Hence, instructions will never be all comprehensive and - sooner or later - some errors or inconsistencies will occur.

2.3 The "instructed agent" will operate in an assumed[22], but not controllable context of (a) the real-world incl. human actors and (b) the digital runtime environment with inevitable errors due to unpredictable situations and unavoidable software aging due to the interaction of very different software layers with non-aligned version/update/patches (Parnas, 1994).

*) The issue of a general risk assessment of novel technologies is beyond the scope of this paper and the reader is referred to e.g. Aven (2012) and Fischhoff (2015).

3. Execution of the instructions in real-time in the real-world with an accountability for explanation:

3.1 As an "instructed agent" will receive input, which will be messy and consists of source data plus noise due to uncalculated effects from the changing environment. Additionally, there are known strategies for so-called adversarial attacks.

[22] All systems deployed by human actors – whether manual procedures or technical tools – will at some point in time deviate from the original assumptions, plans or concepts and will show errors, inconsistencies and/or unintentional collateral effects.

With carefully implemented changes, these adversarial attacks can fool especially AI-based pattern recognition (Biggio and Roli, 2018; Finlayson et al., 2019; Thys et al., 2019). Data about the real-world will represent the actual situation of the society with all the heterogeneity, diversity and legacy (Rice and Swesnik, 2013), but also face the danger of intended manipulations by well-designed attacks[23].

3.2 The agent will execute the pre-defined instructions based on the input variables, calculate corresponding statistical probabilities, validate the "if-then-else" instruction, and - typically in a banking environment - apply a "4-eye-principle" as a second validation step (simplified in Fig. 3, as this step could be a second agent or an independent sub-process). Furthermore, execution processes are part of feedback loops for continuous improvement to achieve a "learning from failure" (Edmondson, 2011).

3.3 The impact of the executed instructions relates to accountability, i.e. explainability of the decision-making to the social context for consequences of and liabilities for the delegated decisions (e.g. approval of a loan or not). The decision-making has a consequence for the specific case, but likewise an impact on the social context:
(i) the social perception how the decision was made and how decisions were explained,
(ii) the social desire to keep control or - in reality - have the perceived feeling of control, and sometimes only an illusion of control

The decision-making process starts with responsibility of the programmers/authors/creators of the "instructed agents" and ends with the accountability of the (legal) entity executing the decision-making process for the impact of the decisions. Although responsibility and accountability[24] may be used synonymously, there is a difference as elaborated by Domènec Melé (2009): While responsibility is an individual obligation of the creators in charge for a certain decision-making (process), the accountability for the explanation of the decision-making and the consequences stays with the legal entity accountable for such decisions towards the society.

This model process can be tested with the three following cases.

How much autonomy has a robot?

Since more than one decade, a discussion has evolved about "autonomy" of technical systems and their role as "moral agents. A survey about the different perspectives can be found in Amanda Sharkey's review "*Can we program or train robots to be good?*" (Sharkey, 2017). The arguments in this paper support the point of view taken especially by Deborah G. Johnson in her seminal work (Johnson, 2006)

"*Computer systems: Moral entities but not moral agents*",

but also a recent contribution by Christoph Markschies (Markschies, 2019).

[23] Especially when AI systems are linked together like in autonomous cars (see e.g.: Tencent 2019).
[24] The European High-Level Expert Group on AI (AI HLEG, 2019) provided a different definition with (i) "transparency" including "explainability" versus (ii) "accountability" in the sense of documentation.

In the process of delegated decision-making, an "instructed agent" can execute exclusively the instructions, which were pre-defined by the programmer and deployed by the user[25], although they may include statistical estimations of probabilities.

Fig. 4 illustrates the issue of autonomy of such an "instructed agent" in the simplified case of buying a bottle of coke. The remote robot does not require any human control and will be able to execute if-then-else instructions, but cannot make decisions without an own will and an own intention. One can derive from the classical legal question whether machines can enter into a contract (with the answer that they are always acting "on behalf" of the owner) a situation with three steps:

1. A human being buys a coke at a kiosk.
2. A human being buys a coke at a vending machine (i.e. a robot provided by an owner).
3. An "instructed agent", i.e. robot, buys a coke at a vending machine, so that two machines interact, but only on behalf of both owners, who provided instructions ex-ante (representing their intentions).

Such robots have no own will and no responsibility[26], which is always with the owner of the robots. This schematic scenario also holds true for a situation with "autonomous" vehicles, which execute instructions in real-time. Of course, an action conducted by a piece of software within milliseconds without any human interference resembles some "autonomy", but those two cases merely differ in the time-scale of execution, in which the predefined instructions are executed. Another factor for the perception of "autonomy" is remoteness in the sense of spatially separation of actors and controllers.

Granting an online loan in real-time to a consumer can create an impression of an "autonomous" system. However, this is just a fast computer program following the same instructions a human being would execute according to the loan manual (but much slower). With the ongoing development in Natural Language Processing (NLP, as a domain of AI), such a loan application will be possible made by a "chat-bot" (text recognition) or a personal assistant (such as Apple's Siri, Microsoft's Cortana, Amazon's Alexa et cetera). Nevertheless, also this so-called "conversational banking" does not indicate "autonomy" of these systems, which simply enter requested data into a loan calculation program.

[25] This includes pre-defined rules e.g. in the case of self-driving cars how to select cruise speed, because simple optimisation algorithms cannot balance speed versus risk: In a traditional perspective, risk minimising would result in an useless car with no (or minimal) speed at all, but (advance) self-driving cars could reduce "normal accidents" by distracted or fatigued drivers while introducing new types of accidents in solitude situations, for which "learned" systems have no solution compared with human intuition based on heuristics.

[26] Although „robots" appeared in the literature first in the 1920s with the science fiction play "Rossumovi Univerzální Roboti" by the Czech writer Karel Čapek in 1920 (see also Wagnerová, 2019)) and later in the novel "Metropolis" by the German writer Thea von Harbou in 1925, the topos has a long legacy starting from the ancient Greek guardian of Crete „Talos" via the golem narrative of Judah Loew ben Bezalel, the late 16th century rabbi of Prague, to Johann Wolfgang von Goethe's Zauberlehrling (The Sorcerer's Apprentice).

Real-time execution and remoteness augment a perception of “autonomy” and lead to enact humans and machines as similar (see especially the seminal book of Lucy Suchman, 2007, about “Human-Machine Reconfigurations”). Although mere tools, the attribution of robots as “autonomous” agents (such as cars, trains, drones, or robots) or “automated” decision-making illustrates the embeddedness of technology in a social context. This becomes even more important in cases of “moral dilemma”, i.e. stylized situations for which there are no existing guidelines for human decision-making[27]. One typical example[28] would be an emergency room with more patients in life-threatening conditions than available medical resources (doctors, rooms, equipment): Who is to be treated first, when actual resource allocation (due to unplanned logistics bottleneck or intended financial planning) collides with human dignity?

No machine, robot or AI, can decide a moral dilemma until the society achieves an agreement about prioritisation of values in these situations. A comprehensive summary was given by Joanna J. Bryson (Bryson, 2018) [quote]: *"The questions of robot or AI Ethics are difficult to resolve not because of the nature of intelligent technology, but because of the nature of Ethics. As with all normative considerations, AI ethics requires that we decide what 'really' matters - our most fundamental priorities."*

Recently, "The Moral Machine Experiment" (Awad, 2018)[29] conducted an international internet survey about preferences in case of a stylized dilemma and analysed the response. Notwithstanding the title, the experiment tested the socially accepted principles for ethical dilemma. With 40 million responses in ten languages from people in 233 countries and territories, this analysis found significant differences based on respondents’ demographics and reported cultural-based ethical variation. Contrary to the opinion of Dewitt (Dewitt et al., 2019), the experiment with its stylized context was not designed as starting point for policymaking concerning “moral machines”[30], but revealed a significant cultural dependency concerning socially agreed principles[31,32,33].

[27] Whereas highway codes usually regulate that a driver has always to drive with speed adapted to the situation and be able to stop a car before an accident, "The Moral Machine Experiment" assumes a situation, in which a driver (human or AI) cannot stop and has to decide between two casualties.
[28] This paper dismisses the often-used case of an “unavoidable” accident (trolley, train, car, or autonomous vehicle) as the emergency room scenario has much higher probability and concerns many people in medical care.
[29] This experiment is to be distinguished from "The Moral Choice Machine" of Jentzsch et al. (2019), which extract deontological ethical reasoning about conduct from human texts.
[30] One can imagine a plethora of theoretically possible dilemma situations e.g. for autonomous vehicles with unavoidable accidents, but only with one simple result that there can be situations without a clear moral guideline for human decision-making (and always given a certain cultural background).
[31] This study triggers discussion about a “common morality", as e.g. promoted by Beauchamp (2003).
[32] An open issue of this study is that the research asked individuals in isolation (online) and with a stylised questionnaire, which might derive from social processes of deliberation (see e.g.: Dryzek, Bächtiger et al., 2019).
[33] See e.g. Whitehouse (2019) for a relationship "Complex societies precede moralizing gods throughout world history".

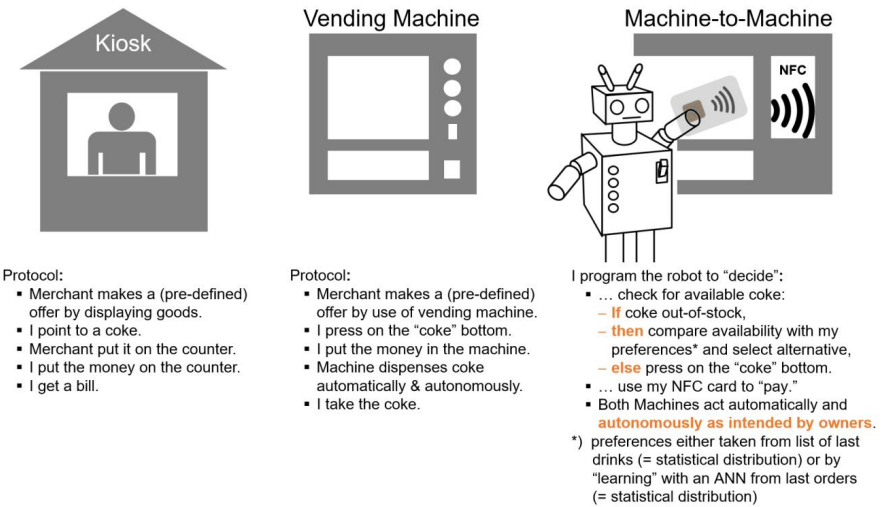

Fig. 4: Autonomy of a robot with "if-then-else" as intended by the programmer

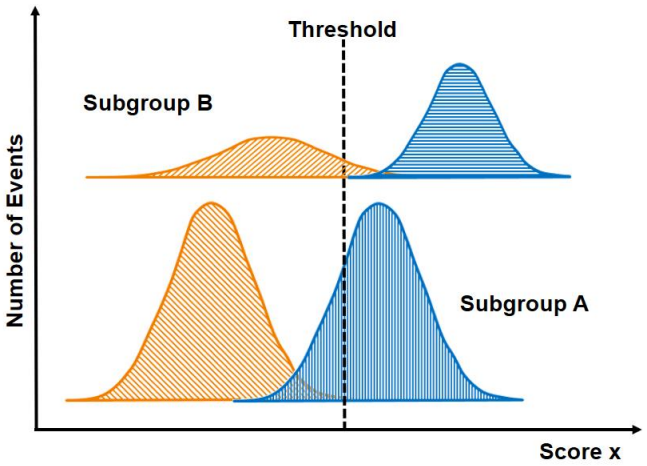

Fig. 5: A statistical classifier with a one-dimensional score value x (i.e. an estimated probability) applied for a distribution with negative (left) and positive (right) events and two different subgroups A and B. The subgroups differ by a "hidden" parameter, which is not included in the scoring. In this schematic example, a choice of the threshold to optimise the false positive / false negative ratio will result in different misclassification (false positive / false negative) ratios if the subgroups would be analysed separately (i.e. by using the hidden parameter as an additional variable). For the illustrative threshold in the graph, subgroup A will show nearly no false positive, but significant false negative results, while in subgroup B will show nearly no false negative, but some false positive result

	Normative approach	Positive perspective	Negative examples
Cultural context	Dependency as illustrated in "The Moral Machine Experiment" (Awad, 2018)		
Legal system "iustitia" (with a dynamic interaction with the development of societies)	Human rights and dignity	• Non-discrimination • Personal rights • Property rights • Freedom of contract	Failed states and abuse of political power, despotism, dictatorship, slavery, repression of minorities, kleptocracy, perversion of justice etc.
Society and economy (as a search process)	Open society and market economy	• Diversity / heterogeneity • Search process of market • Individual responsibility • Social protection	Historical development and legacy of discrimination and unfairness Governmental intervention into a free market economy Paternalism and "nudging"
Individual decisions (under uncertainty and with bounded rationality due to limited capabilities of men and machines)	Moral guidelines in case of dilemmas ("right or wrong conduct")	• Individual preferences and perceptions of risk • Individual decision with responsibility for outcome	Misconduct by individuals, groups, or organisations Abuse of knowledge / information asymmetry or market power
Applied Technology (such as "instructed agents")	Due diligence for development and deployment	• Risk assessment of (new) technologies or methods • Risk protection and safety procedures	(Inadvertent) consequences of complex systems incl. e.g. "software aging" or "moving processes"
Natural errors (or, respectively, protection against)	Resilience and redundancies	• Tests, tests, tests, … • … and the understanding that no (non-trivial) software or system will be fully error-free	Aggregation of many (tested) components in complex systems without a design for resilience

Table 1: The relevant levels for decision-making processes: cultural context, legal system ("iustitia"), society and economy, individual decision-making, applied technology and protection against natural errors (as errors cannot be avoided completely, but only reduced).

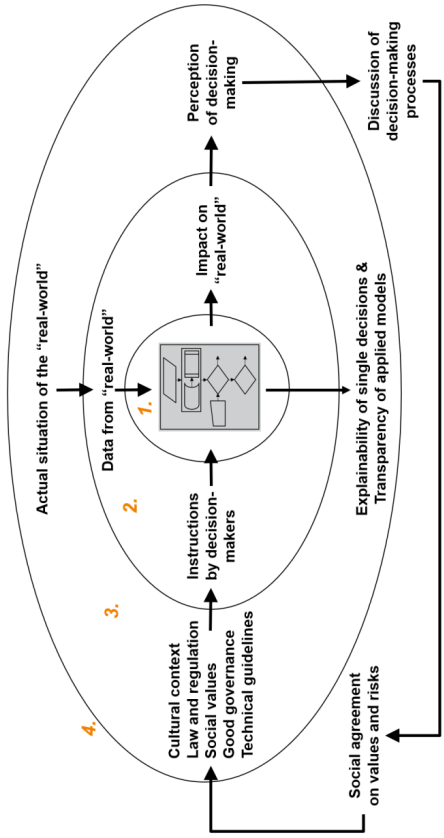

Fig. 6: Simplification of decision-making as a system and the embeddedness in the real-world with four iterations and feedback-loop:

1. *Instructed agent,*
2. *Decision-making process in a corporation (never purely rational),*
3. *Social context with the relevant framework and the perception of the outcome by the society,*
4. *Development of the social agreements in a dynamic discussion of values and risks*

From "moral" dilemma to "social" dilemma in decision-making

As human beings, we have to decide on priorities, principles, or policies. An example could be the decision about a credit risk policy in a bank with clear priorities about the "risk appetite" and consequently definition of parameters for loan approval. However, such responsible decision-making requires an understanding of the impact of the implementation, the social context, and the consequences of our decision to use such tools.

Every decision we make, requires knowledge about the past and responsibility for an uncertain future. The cohesion of knowledge (past), decision (present) and responsibility (future) connects the individual decision-making to the social context and historical development.

The basis for decision-making is our knowledge, which can be structured in four levels of technicality:

- qualitative heuristics (developed during the evolution of human beings)[34]
- quantitative data-set (as recorded representation of the world)
- statistics with the probability calculus (as a codification of experience)
- applied statistics (such as correlations or statistical classifications)

The more the society has been moving from pure heuristics (Gigerenzer et al., 2011) to statistical tools to extract some insight from experience for future application, the more essential an in-depth understanding of those tools is. It is far beyond the scope of this paper to elaborate on the common misunderstandings in society about statistics, and a very good introduction may be the book of Pearl et al. (Pearl, 2016). However, three examples could illustrate the problems people have with statistics:

- A cancer test of a mid-age female person without symptoms or an obvious risk factors will come with false positive results. Due to low probability of cancer but generic uncertainties in the test procedure, only 3 out of 363 women with "positive" test results have really the disease, i.e. there are many "false positive" results for healthy patients (see Pearl, 2018 for details). Such "false positive" results could cause more psychological harm than exhibit help - i.e. mass screening is rather ambivalent and does not automatically achieve an overall benefit.
- The allocation of university places to applicants can be done by a matching process (in the sense of Alvin E. Roth's market design) with an optimal match of the applicants' preferences to available places. Nevertheless, applicants from households in regions with lower average income usually cannot finance studying at faraway places and will prefer universities in the neighbourhood. An optimised match based only on personal preferences will result in a clustering of students from households with lower average income at "poorer" universities, although the matching algorithm does not include "household income" at all.

[34] Recently, McDuff and Kapoor (2019) proposed "visceral machines", which apply human reactions in simulated driving situations (e.g. "fear" measured by pulse amplitude) as incentive in reinforced learning of ANN, as human heuristic reactions provide better learning results compared to physical parameters (such as distance to crash).

University scholarship programmes can improve this social situation practically, but not a centrally "planned" allocation of students to faraway universities against their individual preferences, which would at best achieve abstract objectives of an ideal society.

- A credit scoring in a bank cannot be compared to a loan from a personal friend. While the latter depends on an individual relationship with long-term "experience", a credit scoring is always a statistical calculation for a probable (future) loss based on aggregated historical data of the borrowers[35]. The score value, i.e. an estimation for future defaults under uncertainties, is compared to the pre-defined risk appetite of the bank. Additionally, a general code of conduct will restrict lending to a potential borrower, who cannot be assumed to be able to pay back the loan. Finally, a credit decision is independent, whether the technical approval process is done by a human "instructed" bank employee or an "instructed" technical tool.

A very special kind of dilemma emerges when the dataset consists of two (or more) subgroups, such as people with different „sensitive data" as e.g. defined in the General Data Protection Regulation (European Parliament, 2016): racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, genetic data, biometric data, data concerning health or data concerning a natural person's sex life or sexual orientation. As simplified in Fig. 5, two subgroups may have different probability distribution of "positive" and "negative" data points, i.e. with different mean values and variances. As the base difference between the subgroups based on sensitive data is not considered, an analysis will be made for a single "sub-group blind" distribution with only one threshold. It is likely that different "hidden" subgroups would show different misclassification rates[36], if and only if re-analysed with the (originally not applied) sensitive attribute as control variable.

As shown in Fig. 5, subgroup A will show nearly no false positive, but significant false negative results, while subgroup B will show nearly no false negative, but some false positive result.[37]

Given the assumption in this paper that fraudulent manipulation and/or intended discrimination, e.g. by using sensitive attributes (or "disparate treatment") is taken care of by law and order, a "social dilemma" results from an avoidance of sensitive attributes with an "unfairness" of the outcome (see: e.g. Kleinberg et al., 2017). Scoring in a heterogeneous world will also be heterogeneous when "hidden" subgroups are analysed ex-post ("disparate impact"). An alternative proposal (Kleinberg et al., 2018) to set different thresholds for different subgroups to "promote fairness" could increase the social dilemma. On the one side, some (Which? How many? Who decides?) sensitive attributes have to be included or reconstructed from other features. On the other side, the "adjusted thresholds" or "calibrated ratios" would require a kind of social planner, who decides about a "fair" criteria, adjustments, or calibrations. This leads directly to the danger of centralized planning for every decision-making to "correct" the existing social structure.

[35] A recent study (Frost et al, 2019) found support for the hypothesis that BigTech lenders have an information advantage in credit assessment relative to a traditional credit bureau.

[36] A good explanation was given by Gummadi (2018) with FPR: false positive rate, FNR: false negative rate, FDR: false discovery rate, and FON: false omission rate; and the ratios FPR / FNR versus FDR / FOR.

37 see also: Kleinberg et al. (2017)

The danger could become more severe when more sub-groups or sub-sub-groups defined by combinations of sensitive attributes would be taken into account. In an extreme scenario, the whole data set could fragment into "groups" of size one, but with different sensitive characteristics. Each "group" of size one would require an own threshold, which either results in pre-selection by the planner or in no differentiation at all.

Additionally, no input can represent an absolute "truth". The input data samples are a "measurement", i.e. recording data in a certain predefined scope - always from a subjective perspective, as neither human beings nor technical systems can assess the whole world. There is neither absolute truth in a real-world of (physical) measurements nor in (digital) data samples. Measurements in sciences or data samples for decisions are always dependent on the "research question" or the "experience" of the collector of the data (or the programmer of an instruction or of the user deploying an agent or the outside spectator perceiving an impact et cetera[38]). In a recent communication, the European Commission (2019) summarized [quote]:

> *"When data is gathered, it may reflect socially constructed biases, or contain inaccuracies, errors and mistakes."*

This so-called "sampling bias" (although it is no "bias", but simply "sample") does not imply inaccuracy of statistical classifications, but a dependency on the selection made by the decision-maker[39] and/or a statistical error. However, there is a growing concern about the "fairness", discrimination, disparate treatment / disparate impact, biased data, or even "*Justice and Fairness in Data Use and Machine Learning*" like at Northeastern University's Information Ethics Roundtable[40].

This "social" dilemma results from the tension between two perspectives:

Perception of fairness (see e.g. Grgić-Hlača, 2018) in a society and a wish for redress of historical social unfairness (sometimes called "bias correction"[41]) - with the fundamental problem that "fairness", "social justice" or "distributive justice" depend on individual perspectives.

[38] As all human actors along the decision-making process, including the actors in the (outside) social context, are human beings with bounded rationality, one could appoint a "bias" to every step in a decision-making process. There seems to be a tendency to develop continuously longer and longer lists of "biases" (see e.g. the arbitrary taxonomies by: Danks and London, 2017; and Silva and Kenney, 2018). We all are "biased" by our social context and (path-dependent) personal legacy.

[39] This is also the reason why transporting AI models across different context is critical: For example, the patients in a metropolitan hospital in the USA have different behaviours compared to a county hospital in Germany (i.e. different statistical "confounders", but, additionally, the medical guidelines in the USA and Germany differ, and so do the therapies (see e.g. Balzter, 2018)

[40] A good starting point for an overview of the current discussion can be found at the Social Computing Systems of Krishna P. Gummadi at Max Planck Institute for Software Systems (MPI-SWS; people.mpi-sws.org/~gummadi).

[41] see e.g. Yochai Benkler (2019) claiming [quote]: "*Because algorithms are trained on existing data that reflect social inequalities, they risk perpetuating systemic injustice unless people consciously design countervailing measures.*"

The confrontational discourse about “justice” or “fairness” (except of the legal definition, how “iustitia” works) started with the Scholastics of the School of Salamanca, culminated with the well-known debate between Rawls (1971) and Nozick (1974)/v. Hayek (1976), and continues as antagonism of benevolent theorists and pragmatic realists to this day.

- Statistical distributions with “generic” statistical errors and the impracticality to achieve diverging optimisations simultaneously, if one does not accept either a significant loss of accuracy and/or an additional external objective of an “enforcement of algorithmic fairness” beyond simple statistics.

No machine, algorithm, or AI can solve the conflict between norms and values concerning the social impact of a decision-making and an individual decision-making process with responsibilities for the specific consequences. Of course, this excludes the conscious use of biased data or using sensitive data against the laws, which is a real problem, but one of human beings.

The discussion about “fairness” of AI sometimes has a tendency to start with too simplified imaginations how decision-making is made in an economic context. One example is the approval of a loan and the credit scoring process.

Illustrative, one can look to the narrative in Schröder (2019) that a loan approval could be based on place of residence and people living in an area with historically more defaults would be discriminated. The real process is more sophisticated and does apply a combination of rule-based methods (i.e. methods based on causal models), data records and statistical estimations. Typically, a bank will:

Calculate rules with personal data (e.g. available income of household vs. monthly repayment),

- Use external information sources (e.g. from credit agencies such as Schufa in Germany with proprietary scoring systems based on the credit history of customers of banks, telecommunication providers and other firms with credit risk) and
- Apply internal scoring models (with the statistical estimation of future default probability based on the historical loan portfolio of the bank, which can include AI applications).
- Additionally, in the case of the car loan or a mortgage, the collateral will be taken into account.

Within the combination of causal, historical and statistical models, simple pattern recognition by AI tools would merely be one part. Furthermore, loan approval is the archetype for the 4-eyes principle with a two-step approval process by two agents to avoid errors, bias, or fraud as well – whether the agents are humans and/or machines. This two-step decision-making process can be regarded as a secure application of the proposal of Valera et al. (2019) to enhancing the accuracy and fairness of (human) decision making with a delegation to a pool of experts. Furthermore, there are technical arrangements such as aeroplane navigation systems with an implementation of a “voting” mechanism of three independent systems in parallel (plus back-up by the human pilot in case of no majority result as last resort).

Any bank with a sustainable business model and good conduct will take a double responsibility for (i) a sound internal credit risk management and (ii) good conduct towards the customer, i.e. provide no loan, which would drive a customer into financial distress. The data of a loan portfolio will inevitablely comprise structures of the society like income distribution or regional distribution of economic conditions and security of employment (including statistical confounders and spurious association).

Nonetheless, the actual credit decision of the bank results from statistical estimation of a probability of default with a "demographic blindness". Respectively, a refusal to approve a loan can be explainable and is caused typically by insufficient capabilities for repayment, former financial embarrassment and/or high probability for a default.

Because the business of a bank is *inter alia* to provide loans to the society and, respectively, take credit risk on its own book, a decision-making for a loan approval is based on the data relevant for credit risk management. Banks cannot promote a change in the economic heterogeneity in a society as an objective beyond this scope of the business model. Social changes are tasks of governments (e.g. with redistribution of taxes) or of promotional banks owned by governments or supra-national bodies. However, the subprime mortgage crisis in the U.S. illustrated that a governmental objective (here: to provide an own house for every citizen) can trigger a chain reaction with unintended collateral damages in the long-run due to a complex amalgamation of causes. However, decisions in a commercial bank can be made alone with data (statistically) significant for the generic business model.

A causation from sensitive data such as "data concerning a natural person's sex" (as defined in the GDPR) to income and to the ability to repay a loan can exist. Therefore, "equal pay" is a social question and a political objective. Nonetheless, the GDPR limits explicitely the processing of such sensitive data to justified cases, and additional social and/or political criteria render a risk-based decision-making unfeasible in the sense of a social dilemma beyond the responsibility of a bank[42].

The challenge of explainability versus understanding versus interpretation

On the one side, Matthew Hutson (2018) asked recently: "Has artificial intelligence become alchemy?" as programmers of the most sophisticated AI systems develop with "trial and error" and cannot predict from first principles, which design or parametrisation will be successful. Nonetheless, for a successfully implemented system, input-to-output relations exist including statistical errors. Taking one recent example, machine learning carefully trained on standard magnetic resonance imaging (MRI[43]) are more reliable and precise than established radiological methods in the treatment of brain tumours (Kickingereder et al., 2019).

[42] The well-known newspaper article of Milton Friedman (1979) "The Social Responsibility of Business is to Increase its Profits" describes the different objectives of a firm, the shareholders, the government and the society and is applicable to this day.

[43] MRI itself is a highly sophisticated medical imaging technique to create pictures of the anatomy or the physiological processes: a medical application of nuclear magnetic resonance (NMR) in combination with computer-based data and image analysis. Interestingly, no patient - very with few exceptions - would be able to understand this complex technology, but the patients trust the doctors that they understand it.

This is an essential first step towards the automated high-throughput analysis of medical image data of brain tumours, although the "internal" structure and parametrisation of the artificial neural network (ANN) may be highly complex. However, there is a clear and tested input-to-output relation with better (statistical) results compared to rational analysis.

The perspective of practitioners is that every decision made by professionals or machines requires explainability. Otherwise, our social and commercial systems would be faced with arbitrariness but, respectively, with a lack of accountability of the economic agents.

Every decision-maker in a firm is to take the responsibility for the decision, which rather trivially includes the task to explain the "why" and the "how". Of course, nobody can be forced in a free market economy to enter into a contractual relationship (except for some situation with governmental regulation). However, this is far from a discussion about the use of technology.

On the other side, current[44] AI systems are focussed and restricted. Even advanced implementations such as AlphaZero are limited to the cases they render useful, and according to Campell (2019) AlphaZero makes [quote] "*advantage of the TPU hardware that AlphaZero has been designed to use […] fully observable […] zero-sum, deterministic, static, and discrete, all of which makes it easier to perfectly simulate*".

Playing games may not be a prototype for decision-making, but this example illustrates that even in fully deterministic games (e.g. Chess or Go) with a clear problem ("to win"), no final strategy exists due to the tremendous number of possible moves. Likewise, nobody can "explain" the reasons behind the moves e.g. in the last World Chess Championship 2018 between Magnus Carlsen and his challenger Fabiano Caruana, in which the traditional match ended with 12 consecutive draws and rapid chess was used as a tie-breaker, with Carlsen winning three consecutive games to retain his title. We may understand the rules of chess, but we wonder about the moves both players made.

There is a subtle difference between the explainability of a model and the explainability of a single case, single decision, or single move. If the abilities of AlphaZero or Magnus Carlsen to make a specific decision are compared, machine and human beings are "black boxes" for the rest of the world, although the problem is 100% deterministic. In a chess class, the teacher will explain moves, players can learn different strategies, but a world champion is "unexplainable". Therefore, the challenge of explainability depends on the questions asked, the level of knowledge to "understand" the explanation, and the ability to "interpret" the results. There are different levels of explainability:

1. a subject matter level to explain the dependence of an output on the input
2. an understanding of the decision-making process incl. the accountability of executives
3. the interpretation of the decision-making processes as perceived outside-in from the society

[44] It is out of the scope of this paper to discuss the application of quantum computing to AI (see e.g.: Havlíček, 2019), but such developments - requiring an understanding of quantum mechanics - add an additional layer of complexity to the challenge of explainability, understanding and interpretation.

Although some approaches for explainability on level 1 exist even for complex tasks such as e.g. pattern recognition in pictures with AL[45], these "explanations" could be hard to tell to an executive (on level 2) and would have no meaning for a customer (level 3).

Public perception of decision-making between fear and trust

A headline in The Guardian "*Computer says no: why making AIs fair, accountable and transparent is crucial*" (Guardian, 2017) can be regarded as a key in the whole discussion about the public perception of decisions made by computers. There is this crucial social requirement to regard AI systems as "transparent", so that people can trust in the decision-making, in the same way they would trust other people. Indirectly, this requirement can be found in the European General Data Protection Regulation (GDPR, 2016).

The GDPR highlights explicitly a distinction between the human procession of instruction (e.g. according to a loan manual) and the "automation" by an instructed technical agent executing the same procedure but by software than by paper [quote, emphasis by the author]:

> *General Data Protection Regulation - Recital (71)*
> *The data subject should have the right not to be subject to a decision, […] which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. […]*
> *Article 22. Automated individual decision making, including profiling*
> *1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*
> *2. Paragraph 1 shall not apply if the decision:*
> *(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;*
> *(b) is authorised by Union or Member State law [...] or*
> *(c) is based on the data subject's explicit consent.*

Additionally, the Article 29 Data Protection Working Party (Art. 29 WP; 2017) published "Guidelines on Automated individual decision-making and Profiling …" [quote, emphasis by the author]:

> *Article 22(1) sets out a general prohibition on solely automated individual decision with a significant effect, as described above. …*
> *As described in the WP29 Opinion on legitimate interest, necessity should be interpreted narrowly. The controller must be able to show that this [i.e. automated] profiling is necessary, taking into account whether a less privacy-intrusive method could be adopted. If other less intrusive means to achieve the same goal exist, then the profiling would not be 'necessary'.*

[45] e.g.: the DARPA project (www.darpa.mil/program/explainable-artificial-intelligence); Ribeiro (2016) for the LIME algorithm; the DeCoDeML research network (https://www.tu-darmstadt.de/universitaet/aktuelles_meldungen/news_details_221312.en.jsp), or DreamQuark' as commercial AI tool "Brain" (https://www.dreamquark.com/)

The GDPR and the Guideline express a clear differentiation between a human decision-making and an automated (technical) process of decision-making. Especially, automated decision-making is regarded as a "*privacy-intrusive method"* compared to the same set of instructions executed manually. The GDPR and the Guideline do not provide any justification why the different ways executing the same set of instructions (i.e. manually versus technically) are regarded differently. Nonetheless, the regulator and the Art. 29 WP (now followed by the European Data Protection Board) point out a difference between technology and manual processing. Although speculative, there appears to be one explanation that people are regarded as "trustful", while machines are seen as "opaque".

Pure technological approaches of "explainability" do not address the missing trust in the society to any kind of algorithmic decision-making. The most important issue would be to regain "trust as a reduction of complexity" (as elaborated by Luhmann, 1968) in the society by a process of open communication about technologies and the consequences of decision-making processes on the level of common sense. Trusting other people is always a risky thing, as nobody is fully predictable.

Trust in decision-making requires long-term trustful conduct and clear communication about the process. This includes transparency how, e.g. interest rates for a loan are calculated. Of course, the height of the interest rates belongs to the freedom of contract and depends on the risk appetite of a bank. Nonetheless, people want some control (or at least some feeling that they have control), when they fear to be pulled over the barrel by some "black box" they do not understand. Although this is no rigorous scientific argumentation, ethical conduct and good communications are essential for building trust, especially when it comes to situations, in which people have some subliminal anxiety.

An anecdote about innovations and adoption of technology

Although it is only an anecdote and with very different time scales, the transition from sailing ships to steamboats illustrates the different steps of perceptions (by the public) and adaption (by users) of technology. The development of steamboats took a long time, and a first model steamboat was constructed by the French physicist Denis Papin already in 1707. Unfortunately, this vessel was seized and destroyed on its maiden voyage on the river Weser by local boatmen with the fear it would destroy their occupation.

It took more than 75 year - and only after James Watt developed an improved steam engine in 1769 – to the first steam-powered ship Pyroscaphe built in France in 1783 by Marquis Claude François Jouffroy d'Abbans, which was a paddle steamer travelling on the river Saône for some fifteen minutes before it failed. However, one hundred years after James Watt, the Cutty Sark was built in Scotland in 1869 as one of the fastest - and one of the last - tea clipper sailboats in history. Although sailboats could not compete to steamships commercially then, many executives were in fear of the new technology, maintained the legacy and ignored innovation with a change from wind to steam.

Ironically, the Cutty Sark outlived the steamboat age as training ship into the time of modern shipping engines until 1954. Although contemporary steamships could not outrace a Cutty Sark, the synergy of technological improvements in steam technology and the opening of the Suez Canal (also in 1869) made steamships to dominate the sailing route to India (tea) and Australia (wool).

The Cutty Sark is a warning that sophisticate optimization of legacy technology cannot compete in a market economy when new technology and new channels (literally) provide more benefit for the customers. Likewise, the anecdote illustrates that "technophobia" can be found at different phases of development of technology: at the beginning of novel techniques, but also at maturity.

Decision-making as a process in a multi-level context

The three selected issues illustrate that (i) decision-making is always embedded in the social context and (ii) this complexity of decision-making leads to manifold misunderstandings how decisions are made (either by humans or by machines) and how tools and especially statistics work. Additionally, the tremendous development of technical "instructed agents" mislead public understanding with (a) an anthropomorphisation of machines, but also (b) a subliminal fear of losing control to machines as portrayed in sciences fiction.

Following Deborah G. Johnson, instructed agents are (passive) moral and social entities, but neither (active) moral nor social agents. The hypnotising metaphor of an "ethics of algorithm" falls short of the process of decision-making with delegation, the issue that any social acceptance of technology cannot be "implemented", but only accommodated, and the reality of any decision-making process in the legal, social, individual, technical and error-prone context.

Based on work by Martin Rhonheimer (2015), Table 1 illustrates the five levels, which are relevant for decision-making: the legal system („iustitia"), society and economy, the individual sphere of decision-making, technology and finally protection against errors. Each level has a corresponding normative approach with agreed values, although the "Moral Machine" experiment make the cultural context transparent and, consequently, an additional level of cultural context is supplemented (taking into account that Rhonheimer put the focus on the Judeo-Christian and subsequent European context):

- Cultural context as general environment of life
- Legal system („iustitia") with a dynamic interaction with the development of societies an evolution over time according to the status of the real-world
- Developments in society and especially in economy as a search process for best allocation
- Individual decisions with always incomplete contracts, bounded rationality and uncertainty about the future (see especially: Williamson, 1991)
- Technology as part of an overall "socio-technical" system
- Protection against errors - as a key feature of any technology, when applied by men

Any kind of decision-making is a process embedded in the social context. The current discussion about the ability of AI and/or robots falls short, not only due to the missing free will of the technical agents, but especially as the process of decision-making is compressed on one element (an AI tool) of a long chain and the discussion dismisses the importance of the context:

- No technical tool - including AI and machine learning - can remove bias in society, correct human imprints of historic bias in language (see e.g.: Caliskan et al., 2017), revise historical deficiencies of fairness, or help to define moral guideline for decision-making in case of a dilemma.
- Vice versa, if one wants to tackle the roots for dilemmas (such as limited resources in medical services), to remedy unequal chances in society (e.g. “equal pay”), or to fight against discrimination (intended and unconscious), technology[46] is the wrong place to start with.
- The contemporary public discussion about AI and machine learning systems misunderstands termini technici such as “learning” or “intelligence” and attributes human abilities and human behaviour to statistical classifiers.
- Vice versa, guidelines for decision-making could be aligned to standards such as Beauchamp and Childress (1977) "*Principles of Biomedical Ethics*", but with the same focus to guide the human professional-patient relationship in health care (respect for autonomy, non-maleficence, beneficence, and justice) in the tradition of the Hippocratic Oath[47].
- This blending of guidelines for human decision-makers with the technical execution of predefined instructions generates mistrust against the current technological development. This indicates a gap in communication to public stakeholders, but also misunderstandings in public agencies (as an example see e.g. Mihalis, 2018). Approaches that current “autonomous” systems establish new socio-technical phenomena, which in turn require new ethical concepts (see e.g. Simon, 2019), may support a process-oriented perspective, but have the tendency to attribute human-like features to mere technical tools.
- The concept of decision-making as a process in a social context illustrated a number of starting points, where particular improvements can be made: from cockpits to get insight into the technological tools to corporate governance guidelines with clear definition of responsibility and accountability. Nonetheless, the key will be trust building by an open and easy to understand communication of the whole process and the interdependency with the social context.

[46] Any machine-learning is prone to incorporating the biases of the society, which will creep in the data-sets used to train the AI tools, but also any traditional statistical analysis tools. Although beyond the scope of this paper, current concepts of distributed deep learning, i.e. local use of data e.g. from mobile devices without accessing the raw data centrally (see e.g. Vepakomma, 2018) could provide some benefits for unbiased data, but add additional layers to the training.

[47] It should be remarked that the Hippocratic Oath says [quote] "help the sick according to my ability and judgment". Thus, physicians have the responsibility for their decisions and cannot justify themselves by reference to wrong information, books with “printing errors” or “black box” algorithms.

How humans treat robots and algorithms

People tend to project human-like qualities onto “robots”, “chat-bots”, or “PA’s” (personal assistance, such as Siri, Cortana, Alexa & Co.). As increasingly technology - based on different forms of AI - interacts with humans, this anthropomorphism gains centre stage compared to internal features of AI systems.

We are going to ascribe agency to these AI systems and treat them as "moral machines", "social actors", or "trustful companions". Kate Darling (2017) did an intriguing experiment with framing. For a small robotic toy different, two narratives were used: either the robot was "Frank" with a personal story behind or the robot was introduced as a non-personal object. To cut a long story short, people treated "Frank" significantly more human-like compared to the non-personal robot.

Similar reactions of humans in the interaction with robots are reported for "moving" robots (e.g. by Kate Darling at the RE: MARS conference in 2019) or robots with a human-like language (i.e. with delays, with introduced "ah's", or with some "emotional" intonation).

With much simplification, one can imagine the following hierarchy:

- Mechanical loom (programmable with perforated cards as an invention by Jean-Charles Jacquard in 1804)
- Combined harvester (today with GPS, navigation systems et cetera)
- Software for commercial use (“accounting”)
- Pattern recognition (as domain of AI)
- Automated processing of a loan application
- Autonomous vehicle or drone
- Toy robot with a name (“Frank”)
- Moving robot with natural language capabilities (“Hello, who is there?”)
- The humanoid robot "Sophia "developed by Hanson Robotics with a first appearance on stage at South by Southwest Festival in March 2016

This is a list of technologies sorted by advanced features, but no hierarchy in the sense of increasing “personality”. These tools can help (or replace!) humans to execute daily jobs, and they all do not differ in principle, as they are tools designed by a “programmer” and have no own intention, will or choice. However, there is a difference from an outside point of view how they are perceived and trodden by people! The more human-like the robots appear (ability to move, ability to speak, name), the more they are treated as agents with their personality.

Antagonistic to this anthropomorphism, one can regard the recently introduced concept of algorithmic risk as to the risk that algorithms are perceived negatively from an outside-in point of view. From a traditional “internal” perspective, algorithms do not differ from other (computer) technologies and have operational risk (due to errors, fraud, or misconduct) and model risk (due to assumptions and limitations of the implemented concepts).

The new “algorithmic risk” can be regarded as the risk of a well working algorithm (as intended and within the defined quality criteria such as, e.g. statistical error of prediction), but with an impact on the context which is regarded as a violation of ethical, social or political norms by external stakeholders. Similar to reputational risk, it matters whether the algorithm - or more precisely the responsible decision-makers - is accused of a transgression publicly.

The best known example might be the COMPAS cases (a software for Correctional Offender Management Profiling for Alternative Sanctions), in which a first analysis argued that for black defendants the estimated likelihood for recidivism was higher than for white defendants, while white defendants were more likely to be incorrectly flagged as low risk, while the input data did not (sic!) include an individual’s race!

This publication sends shockwaves across the media is still quoted in the discussion about “algorithmic fairness”, although the analysis itself was “biased”. Although Gummadi (2018) elaborated that this recidivism prediction tool suffers from the problem that base recidivism rates for different races differ and therefore [quote]: “no non-trivial solution to achieve similar FPR, FNR, FDR, FOR” rates[48] exist. In other words, it depends on the - ex-post - chosen fairness measures, whether this tool can be regarded as fair or as unfair. Additionally, Dressel and Farid (2018) showed that this tool is no more accurate than predictions made by people and that a simple linear predictor provided with only two features could be nearly equivalent. All-in-all, the tool can be regarded as poorly programmed software, which requires much mathematical understanding to be used correctly. The tool is not “unfair”, but there is the social problem that base recidivism rates differ for diverse races. Nonetheless, the discussion about “fairness” was focussed on the algorithm, but neither on the historical development of the society nor on the naïve way the algorithm was applied in the justice system instead of human judgement.

Conclusion

As the comic of the robot in Fig. 4 illustrates, we can delegate a pre-defined task to an “instructed agent”, which will act on behalf of the programmer and/or user without anything like an individuality, free will or own choice. Such robots, machines or AI tools are simple entities in an overall process of decision-making. The “instructed agent” represents the intention of the programmer and the user, who deploys the agent, and the responsibility for decision-making under uncertainty remains always with the programmer or user, who depend on their subjective experience.

Furthermore, execution of the decision-making is part of the fabric of society, which reflects - unfortunately - the actual matrix of the world including prejudices, stereotypes and bias in the society with historical patterns of discrimination and exclusion. Concerning any decision-making, it has to be decided whether experience (i.e. a sample of data) about the actual conditions should be used, or whether an ideal “utopia” of a social planner is to be taken as guideline. Nevertheless, a user deploying an “instructed agent” (human, technical, or AI-based) has an accountability for explanation of the impact of its decision-making to the public.

[48] FPR: false positive rate, FNR: false negative rate, FDR: false discovery rate, and FOR: false omission rate

Attempts to focus only on "instructed agents" without the embeddedness into the context fall short. Either they are going in the direction to overload the agents and [quote]: "*conceptualize algorithms as value-laden, rather than neutral, in that algorithms create moral consequences, reinforce or undercut ethical principles, and enable or diminish stakeholder rights and dignity*" (Martin, 2017). Alternatively, they define the whole system as [quote]: "*Algorithms as culture*" (Seaver, 2017). Both ways of attribution of human responsibilities to machines is diluting the key issue that decision-making process has to be communicated properly to the society.

As David Beer (2017) elaborated [quote]: "*The notion of the 'algorithm' is now taking on its own force, as a kind of evocative shorthand for the power and potential of calculative systems that can think more quickly, more comprehensively and more accurately than humans. As well as understanding the integration of algorithms, we need to understand the way that this term is incorporated into organisational, institutional and everyday understandings*." Even more provocative, Mona Sloane (2019) argued [quote]: "*that the hype around 'ethics' as panacea for remedying algorithmic discrimination is a smokescreen for carrying on with business as usual*."

Recent proposals (see e.g. Kleinberg et al., 2019) discussed the active use of computerized algorithms to detect possible bias or discrimination as documented computer programs (whether traditional rule-base or advances ANN) are available for inspection compared to "gut decision" made by human beings. However, this has nothing to do with "intelligence" but with available documentation (= training data plus computer code, even if presented as a parametrised neural network).

Starting with the original wording of "intelligence" and "learning", the discussion about AI in decision-making faces the danger of misunderstanding the capabilities and limitations of this technology, which become manifest in combined terms like "Moral Machines", "Algorithmic Fairness", "Trustworthy AI", or "Good AI Society" (Floridi, 2018).

Such notions generate ambiguity, because they introduce an anthropomorphisation[49,50] of machines. Recently, Rich and Gureckis (2019) summarized [quote]: "*machine systems … share many of the same limitations that frequently inhibit human judgement, for many of the same reasons*." And we can take "*[l]essons for artificial intelligence from the study of natural stupidity*".

[49] On the other hand, a recent study by Nijssen et al. (2019) revealed that an anthropomorphic appearance of robots lures people to attribute affective states to them and, in a stylize dilemma situation, people are even reluctant to sacrifice them in order to save humans.

[50] There are also ideas to "correct" the decision made by human judges with AI, as proposed by Chen (2019): "*By predicting judicial decisions [...] machine learning offers an approach to detecting when judges most likely to allow extra legal biases to influence their decision making.*" This would contradict the legal tradition to have humans judging humans with the ability to consider the whole matrix of the world behind a given case, and would start to introduce "a rule of the machines".

The proposed process with (i) decision-makers with responsibility, (ii) accountability for the impact of the execution of instructions by "instructed agents", and (iii) the perception in the society can help to clarify the discussion. Especially the chain of general responsibility of the decision-makers, deployment of the "instructed agents", and perception of the impact by public stakeholders can assist in the discussion about existing bias in the society, misconduct in doing business, or the de-coupling of responsibility from decision-making (especially within large organisational structures).

Essential questions of our present society, the historical legacy and the future development should focus on the decision-makers, but not first and foremost on the technical auxiliaries[51].

Potential Conflict of Interest

The views expressed in this paper are those of the authors and not necessarily those of the organisations mentioned.

Acknowledgments

The authors want to thank Stefanie Büchner, Hans-Christian "Chris" Boos and Ritva Tikkanen for valuable discussions and comments.

[51] A recent paper by Rahwan, Cebrian, Obradovich, et al. (2019) even defined a new field for scientific studies [quote]: "*machine behaviour: the scientific study of behaviour exhibited by intelligent machines*". However, currently nothing like an "intelligent" machine even exist, but only computer programs based on rules and/or statistical analysis; although many computer programs may sometime show "strange" behaviour especially when used under conditions not foreseen by the programmers. On the other side, it is well known that even very simple and deterministic "machines", so-called cellular automata (see especially: Wolfram, 2002), can reveal complex and unexpected "behaviour".

References

Acemoglu, Daron and Pascual Restrepo (2019) "The Wrong Kind of AI? Artificial Intelligence and the Future of Labor Demand", 5.3.2019, available at: https://economics.mit.edu/files/16819 (accessed 12.5.2019).

Agrawal, Ajay, Joshua S. Gans and Avi Goldfarb (2019) "Prediction, Judgment, and Complexity: A Theory of Decision Making and Artificial Intelligence", 14.4.2018, in: , A. Agrawal et.al. (Eds.) "The Economics of Artificial Intelligence", University of Chicago Press/NBER, 2019 (drafts available at: https://www.nber.org/books/agra-1; acessed 22.3.2019):

AI HLEG (Independent High-Level Expert Group on Artificial Intelligence - Set up by the European Commission; 2019) "Ethics Guidelines for Trustworthy AI", European Commission, 8.4.2019, available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58477 (accessed 8.4.20199

Asimov, Isaac "(1942) Runaround", Astounding Science Fiction, issue March 1942, in: “I, Robot “, The Isaac Asimov Collection, New York City: Doubleday, 1950, p. 40.

Aven, Terje (2012) "Quantitative Risk Assessment", Cambridge University Press (June 2012).

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.

Bank of America (2019) "Bank of America’s Erica® Surpasses 6 Million Users", Bank of America Newsroom, 21.3.2019 (available at: https://newsroom.bankofamerica.com/press-releases/consumer-banking/bank-americas-ericar-surpasses-6-million-users; accessed 25.3.2019).

Balzter, Sebastian "Im Krankenhaus fällt Watson durch", Frankfurter Allgemeine Zeitung, 3.6.2018.

Bartels, Jörn and Thomas Deckers (2019) "Big Data trifft auf künstliche Intelligenz", in: BaFin "Digitalisierung - Folgen für Finanzmarkt, Aufsicht und Regulierung – Teil II", BaFin Perspektiven, Ausgabe 1-2019, Bundesanstalt für Finanzdienstleistungsaufsicht (28.2.2019).

Beauchamp, Tom L. and James F. Childress (1977) "Principles of Biomedical Ethics", Oxford University Press, 1977; Seventh Edition (October 2012).

Beauchamp, Tom L. (2003) "A defense of the common morality", Kennedy Institute of Ethics Journal, Vol. 13/3, pp. 259-274.

Beer, David (2017) "The social power of algorithms", Information, Communication & Society, Vol. 20/1, pp. 1-13.

Bengio, Yoshua (2019) "AI pioneer: ‘The dangers of abuse are very real’", interview in Nature News Q&A 4.4.2019), available at: https://www.nature.com/articles/d41586-019-00505-2 (accessed: 6.4.2019).

Benkler, Yochai (2019) “Don’t let industry write the rules for AI”, Nature, Vol. 569, 9.5.2019, p. 161.

Berenbach, Brian and Manfred Broy (2009) "Professional and Ethical Dilemmas in Software Engineering", Computer, Vol. Feb. 2009, Published by the IEEE Computer Society, pp. 74-80.

Biggio, Battista and Fabio Roli (2018) "Wild patterns: Ten years after the rise of adversarial machine learning", Pattern Recognition, Vol. 84, pp. 317-331 (December 2018).

Boos, Hans-Christian (2018) “Artificial Intelligence and the Future of Business”, TEDxWHU, 26.3.2018 (available at: https://www.youtube.com/watch?v=5NlmdoHzmrw, accessed 27.2.2019).

Bryson, Joanna J. (2018) "Patiency is not a virtue: the design of intelligent systems and systems of ethics", Ethics and Information Technology, Vol. 20, pp.15–26 (16.2.2018).

Büchner, Stefanie, Stefan Kühl, Judith Muster (2017) "Ironie der Digitalisierung - Weswegen Steuerungsphantasien zu kurz greifen", Working Paper 13/2017 (3.7.2017; available at: https://pub.uni-bielefeld.de/download/2913003/2913004/Working-Paper-13_2017-Buechner-Muster-Kuehl-2017-Ironie-der-Digitalisierung-mit-Literatur-03.07.17%281%29.pdf; accessed 22.3.2019).

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan (2017) "Semantics derived automatically from language corpora contain human-like biases", Science, Vol. 356/ 6334, pp. 183-186 (14.4.2017).

Campell, Murry (2018) “Mastering board games”, Sciences, Vol. 362/6419 (7.12.2018).

Chen, Daniel L. (2019) "Machine Learning and the Rule of Law", in: M. Livermore and D. Rockmore (eds.) Computational Analysis of Law, Santa Fe Institute Press, Forthcoming, available at https://ssrn.com/abstract=3302507, 6.1.2019 (accessed: 4.4.2019).

Danks, David and Alex John London (2017) "Algorithmic Bias in Autonomous Systems", IJCAI'17 Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19-25.8.2017, AAAI Press, Palo Alto, CA, pp. 4691-4697.

Darling, Kate (2017) “'Who's Johnny?' Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy,” in: ROBOT ETHICS 2.0 by Patrick Lin, Ryan Jenkins, and Keith Abney (eds.), Oxford University Press, N.Y., pp. 173-192.

Davies, Mike et al. (2018) "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning", IEEE Micro, Vol. 38/1, pp. 82 - 99.

Dewitt, Barry, Baruch Fischhoff and Nils-Eric Sahlin (2019), Correspondence in Nature, Vol. 567, p.31 (7.3.2019)

Dourish, Paul (2017) "Algorithms and their others: Algorithmic culture in context", Big Data & Society, July-Dec. 2017, pp. 1-12.

Dressel, Julia and Hany Farid (2018) "The accuracy, fairness, and limits of predicting recidivism", Science Advances, Vol. 4/1, 17.1.2018, available at: https://advances.sciencemag.org/content/4/1/eaao5580/tab-pdf (accessed: 4.7.2019).

Dryzek, John S., André Bächtiger et al. (2019) "The crisis of democracy and the science of deliberation", Science, Vol. 363/6432, pp. 1144-1146, (15.3.2019).

Edmondson, Amy C. (2011) "Strategies for Learning from Failure", Harvard Business Review, Vol. April 2011.

Elgammal, Ahmed (2019) "AI Is Blurring the Definition of Artist", American Scientist, Vol. 107/1,pp. 18-21 (available at: https://www.americanscientist.org/article/ai-is-blurring-the-definition-of-artist; accessed 20.4.2019).

European Commission (2019) "Building trust in human-centric artificial intelligence", Communication COM(2019) 168, European Commission, 8.4.2019.

European Parliament (2017) “Civil Law Rules on Robotics" (A8-0005/2017), 27.1.2017 (available at: http://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.pdf; accessed 20.2.2019.

European Parliament (2019) “A comprehensive European industrial policy on artificial intelligence and robotics”, 2018/2088(INI)) European Parliament - Committee on Legal Affairs "Report,30.1.2019 (available at: www.europarl.europa.eu/doceo/document/A-8-2019-0019_EN.pdf, accessed: 2.2.2019)

European Union (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), Official Journal of the European Union, 4.5.2016.

Feldmann, J., N. Youngblood, C. D. Wright, H. Bhaskaran and W. H. P. Pernice (2019) "All-optical spiking neurosynaptic networks with self-learning capabilities", Nature, Vol. 569, pp. 208–214.

Finlayson, Samuel G., John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, Isaac S. Kohane (2019) "Adversarial attacks on medical machine learning", Science, Vol. 363/6433, pp. 1287-1289 (22.3.2019)

Fischhoff, Baruch (2015) "Risk Assessment - The realities of risk-cost-benefit analysis", Science, Vol. 350/6260, aaa6516 (30.10.2015).

Floridi, Luciano et al. (2018) “AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", Minds and Machines, Vol. 28, pp. 689–707.

Friedman, Milton (1970) "The Social Responsibility of Business is to Increase its Profits", The New York Times Magazine, September 13, 1970.

Frost, Jon, Leonardo Gambacorta, Yi Huang, Hyun Song Shin and Pablo Zbinden (2019) "BigTech and the changing structure of financial intermediation", BIS Working Papers, No 779, 8.4.2019.

Gigerenzer, Gerd, Ralph Hertwig and Thorsten Pachur (2011) "Heuristics: The Foundation of Adaptive Behavior", Verlag: Oxford University Press (26.5.2011).

Goodman, Bryce and Seth Flaxman (2017) "European Union regulations on algorithmic decision-making and a 'right to explanation'", AI Magazine, Vol 38/3, 2017.

Grgić-Hlača, Nina, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller (2018) "Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction", to appear in the Proceedings of the Web Conference (WWW 2018); available at: arXiv:1802.09548v1, submitted 26.2.2018, accessed 16.3.2019).

2017), Leibniz International Proceedings in Informatics, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany, Article No. 43, pp. 43:1–43:23.

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 at: https://www.nber.org/papers/w25548.pdf (accessed 1.7.2019).

Kompella, Varun Raj, Matthew Luciw, Marijn Stollenga, Leo Pape, and Jürgen 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), 7.-9.11.2012, San Diego, CA, IEEE, pp. 1-8.

Kozyrkov, Cassie (2019) "What is AI bias?", Towards DataCcience, 24.1.2019 (available at: https://towardsdatascience.com/what-is-ai-bias-6606a3bcb814, accessed 11.3.2019).

Kritikos, Mihalis (2018) "What if algorithms could abide by ethical principles?", European Parliamentary Research Service, PE 624.267 – November 2018.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman (2017) "Building machines that learn and think like people", Behavioral and Brain Sciences, Vol. 40, E253 (published online: 24.11.2016).

Lapuschkin, Sebastian, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek and Klaus-Robert Müller (2019) "Unmasking Clever Hans predictors and assessing what machines really learn", Nature Communications, Vol. 10, Article number: 1096 (2019).

Latham, Katherine J. (2013) "Human Health and the Neolithic Revolution: an Overview of Impacts of the Agricultural Transition on Oral Health, Epidemiology, and the Human Body", Nebraska Anthropologist, Vol. 187, pp. 95-102.

Luhmann, Niklas (1968) "Vertrauen: Ein Mechanismus der Reduktion sozialer Komplexität", Enke Verlag, 1968.

McDuff, Daniel and Ashish Kapoor (2019) "Visceral Machines: Risk-Aversion in Reinforcement Learning with Intrinsic Physiological Rewards", to be published as a conference paper at ICLR 2019, Seventh International Conference on Learning Representations, May 6-9, 2019, New Orleans, USA (available at: https://www.microsoft.com/en-us/research/publication/visceral-machines-risk-aversion-in-reinforcement-learning-with-intrinsic-physiological-rewards/; accessed 13.5.2019)

Markschies, Christoph (2019) “Warum sollte man einem Computer vertrauen?”, F.A.Z., 23.2.2019.

Martin, Kirsten "Ethical Implications and Accountability of Algorithms", Journal of Business Ethics, published online: 7.7.2017.

Melé, Domènec (2009) "Business Ethics in Action: Seeking Human Excellence in Organizations", Macmillan International Higher Education, Houndmills, Basinstoke, Hampshire.

Milkau, Udo (2017) “Risk Culture during the Last 2000 Years - From an Aleatory Society to the Illusion of Risk Control”, Int. J. Financial Stud., Vol.5/4.

Minsky, Marvin and Seymour Papert (1969) "Perceptrons: an introduction to computational geometry", MIT Press, MA, USA.

Newell, Allen, Shaw, J.C., and Herbert A. Simon (1959). Report on a general problem-solving program, Proceedings of the International Conference on Information Processing, UNESCO, Paris, 15-20.6.1959. pp. 256–264.

New Zealand Government (2017) "Whanganui River settlement passes third reading", https://www.beehive.govt.nz/release/whanganui-river-settlement-passes-third-reading (accessed: 27.3.2019).

Nijssen, Sari R.R., Barbara C. N. Müller, Rick B. van Baaren, and Markus Paulus (2019) "Saving the Robot or the Human? Robots Who Feel Deserve Moral Care", Social Cognition, Vol. 37/1, pp. 41-52.

Northeastern University (2019) "17th Annual Information Ethics Roundtable: Justice and Fairness in Data Use and Machine Learning", 5.-7.4.2019, (www.northeastern.edu/csshresearch/ethics/information-ethics-roundtable/; accessed 19.2.2019).

Nozick, Robert (1974) “Anarchy, State, and Utopia”, Basic Books, 1974.

Olazaran, Mikel (1996) "A Sociological Study of the Official History of the Perceptrons Controversy", Social Studies of Science, Vol. 26/3, pp. 611-659.

Olhede, S. C. and P. J. Wolfe (2018) "The growing ubiquity of algorithms in society: implications, impacts and innovations", Phil. Trans. R. Soc. A 376: 20170364, pp. 1-16 (25.6.2018).

Parikh, Ravi B., Ziad Obermeyer, and Amol S. Navathe (2019) "Regulation of predictive analytics in medicine", Science, Vol. 363/6429, pp. 810-812 (22.2.2019).

Parnas, David L. (1994) “Software aging”, 16th International Conference on Software Engineering, 1994. Proceedings, ICSE-16, pp. 279–287.

Pearl, Judea (2000) "Causality", Cambridge University Press, 2000.

Pearl, Judea, Madelyn Glymour and Nicholas P. Jewell "Causal Inference in Statistics - A Primer", John Wiley & Sons (19.2.2016).

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

Rawls, John (1971) “A Theory of Justice”, Harvard University Press, 1971.

Rhonheimer, Martin (2015) “The True Meaning of ‘Social Justice’: A Catholic View of Hayek,” published in: Economic Affairs, Vol. 35/1, pp. 35–51 (February 2015).

Ribeiro, Marco Tulio, Sameer Singh and Carlos Guestrin (2016) “"Why Should I Trust You?' - Explaining the Predictions of Any Classifier", ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, 13-17.8.2016 (available at:

http://www.kdd.org/kdd2016/subtopic/view/why-should-i-trust-you-explaining-the-predictions-of-any-classifier; accessed 31.3.2019).

Rice, Lisa and Deidre Swesnik (2013) "Discriminatory Effects of Credit Scoring on Communities of Color", 46 Suffolk University Law Review, Vol. XLVI/935, pp. 952–957.

Rich, Alexander S. and Todd M. Gureckis (2019) "Lessons for artificial intelligence from the study of natural stupidity", Nature Machine Intelligence, Vol. 1, pp. 174–180.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986) "Learning representations by back-propagating errors", Nature, Vol. 323/6088, pp.533–536.

Seaver, Nick (2016) "Algorithms as culture: Some tactics for the ethnography of algorithmic systems", Big Data & Society, July-Dec. 2016, pp. 1-12.

Schröder, Tim (2019) "Auf Fairness programmiert", MaxPlanckForschung, Vol. 2-2019 (in German), pp. 68-73, (available at: https://www.mpg.de/13547648/W004_Material_Technik_068_073.pdf; accessed 29.6.2019).

Sharkey, Amanda (2017) "Can we program or train robots to be good?", Ethics and Information Technology, May 26, 2017, pp 1–13 (26.5.2017).

Silva, Selena and Martin Kenney (2018) "Algorithms, Platforms, and Ethnic Bias: An Integrative Essay", Phylon: The Clark Atlanta University Review of Race and Culture, Vol. 55/1-2, pp. 9-37.

Simon, Herbert (1957) "A Behavioral Model of Rational Choice", in Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting. New York: Wiley, 1957.

Simon, Judith (2019), "Brauchen wir eine neue Ethik für die digitale Welt?", Forschung & Lehre, 3-2019, Standpunkt, p. 217 (available at: https://www.forschung-und-lehre.de/brauchen-wir-eine-neue-ethik-fuer-die-digitale-welt-1574/; accessed: 27.3.2019).

Sloane, Mona (2019) "Inequality Is the Name of the Game: Thoughts on the Emerging Field of Technology, Ethics and Social Justice", in: Proceedings of the Weizenbaum Conference 2019 "Challenges of Digital Inequality - Digital Education, Digital Work, Digital Life", Berlin, 16.-17.5.2019, pp. 1-9.

Suchman, Lucy (2007) "Human-Machine Reconfigurations: Plans and Situated Actions - Learning in Doing: Social, Cognitive and Computational Perspectives, Cambridge University Press, 2007.

U.S. Census Bureau (not dated), www.census.gov/history/www/innovations/technology/the_hollerith_tabulator.html; assessed 6.3.2019.

Taddy, Matt (2019) "The Technological Elements of Artificial Intelligence", in: A. Agrawal et.al. (Eds.) "The Economics of Artificial Intelligence", University of Chicago Press/NBER, 2019 (drafts available at: https://www.nber.org/books/agra-1; acessed 22.3.2019).

Tencent Keen Security Lab (2019) "Experimental Security Research of Tesla Autopilot", 29.3.2019, available at: https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf (accessed 3.4.2019).

Thys, Simen, Wiebe Van Ranst and Toon Goedem (2019) "Fooling automated surveillance cameras: adversarial patches to attack person detection", 18.4.2019, accepted for CVPR Workshop: CV-COPS 2019, available at: https://arxiv.org/pdf/1904.08653 (assessed 24.4.2019).

Ullman, Shimon (2019) "Using neuroscience to develop artificial intelligence", Science, Vol. 363/6428, pp. 692-693.

Valera, Isabel, Adish Singlay and Manuel Gomez-Rodriguez (2019) "Enhancing the Accuracy and Fairness of Human Decision Making", Advances in Neural Information Processing Systems, Vol. 31, pp. 1774--1783.

Vepakomma, Praneeth, Tristan Swedish, Ramesh Raskar, Otkrist Gupta, and Abhimanyu Dubey (2019) "No Peek: A Survey of private distributed deep learning", arXiv:1812.03288, 8.12.2018 (accessed: 16.3.2019).

Wagnerová, Alena (2019) "Karel Čapeks Drama 'R.U.R.' : Träumen Roboter von der Liebe?" (in German), F.A.Z., 27.4.2019, available at: https://www.faz.net/aktuell/feuilleton/karel-apeks-kollektivdrama-r-u-r-16158982.html (accessed 1.5.2019).

Weber, Max (1922) „Der Sinn der 'Wertfreiheit' der soziologischen und ökonomischen Wissenschaften“, in: Gesammelte Aufsätze zur Wissenschaftslehre, Tübingen 1922, 7. Auflage, Mohr Siebeck, Tübingen 1977, 467ff.

Whitehouse, Harvey et al. (2019) "Complex societies precede moralizing gods throughout world history", Nature, 20.3.2019 (available at: www.nature.com/articles/s41586-019-1043-4; accessed 22.3.2019)

Williamson, Oliver E. (1981) "The economics of organization: the transaction cost approach", American Journal of Sociology, Vol 87/3, pp. 548–577.

Wolfram, Stephen (2002) "A New Kind of Science", Wolfram Media Inc, Champaign, IL, USA.