Professional MBA
Entrepreneurship & Innovation

Reliability of algorithms and AI-powered systems and trust in results generated

A Master's Thesis submitted for the degree of
“Master of Business Administration”

supervised by
Univ.Prof. Dr. Sabine Theresia KÖSZEGI

Elena Yalozova, BSc

11725399

Vienna, 30.06.2019

Affidavit

I, **ELENA YALOZOVA, BSC**, hereby declare

1. that I am the sole author of the present Master's Thesis, "RELIABILITY OF ALGORITHMS AND AI-POWERED SYSTEMS AND TRUST IN RESULTS GENERATED", 79 pages, bound, and that I have not used any source or tool other than those referenced or any other illicit aid or tool, and
2. that I have not prior to this date submitted the topic of this Master's Thesis or parts of it in any form for assessment as an examination paper, either in Austria or abroad.

Vienna, 30.06.2019

_____

Signature

Abstract

Recently, Artificial intelligence (AI) has attracted a lot of media coverage, much of which is dedicated to biased judgments, "AI-discrimination," "AI-racism," and other errors identified, which could harm humans and society. Given that, how can enterprises integrate AI into their underlying businesses, if people distrust it works flawlessly?

This thesis is aimed to demystify AI and propose guidance that facilitates the integration of AI-tools into the daily operations.

Research questions focus on the reliability of existing AI techniques and their implementation and the problem of users trust'. It covers the latest trends in the AI domain like responsible AI and explainable AI, their ability to open "black boxes."

This thesis contributes to the better comprehension of AI and computational reliability for non-specialists. It proposes a systematic approach to the integration AI-tools and draws companies' attention to the significant challenges for AI adoption - the lack of user's knowledge and necessity to build user's trust. These aspects prevent the exploitation of AI-tools and could lead to missing business opportunities.

Business leaders and entrepreneurs must take steps toward building AI-friendly culture, and include the latter problems into a digital transformation strategy which should be developed upfront tech initiatives and has to be seen interconnected with the business strategy.

Keywords: AI, XAI, FAT, ethical AI, AI reliability, trust in AI

Table of Contents

Abstract **1**

Table of Figures **4**

Table of Tables **4**

List of acronyms **5**

1 Introduction **6**

1.1 A concise history of AI *6*

1.2 Problem formulation *12*

1.3 Aims and objectives *13*

1.4 The course of the investigation *14*

1.5 Structure of the thesis *15*

2 AI and Machine Learning **16**

2.1 How Machine Learning works *21*

2.1.1 Data selection 21

2.1.2 Data modeling 21

2.1.3 Validation 22

2.1.4 Accuracy 23

2.2 Machine learning methods *24*

2.2.1 Heuristics 25

2.2.2 Classification algorithms 27

2.3 Five paradigms of Machine Learning *29*

2.3.1 Symbolists. Decision trees 29

2.3.2 Bayesians. Naïve Bayes and Markov 31

2.3.3 Connectionists. ANN 33

2.3.4 Evolutionaries. Genetic algorithms 34

2.3.5 Analogizers. SVM 34

3 Explainable AI **36**

3.1 Defining Interpretability and Trustworthy AI *38*

3.2 Taxonomy for ML-Interpretability *42*

3.2.1 A scale of complexity 43

3.2.2 Global and local interpretability 45
3.2.3 Model-agnostic and Model-Specific Interpretability 46
3.3 Common interpretability techniques 46
3.3.1 Data visualization techniques for interpretability 46
3.3.2 Techniques for creating White-Box models 47
3.3.3 Interpretability techniques in complex ML models 50
3.3.4 Reason codes 51
3.3.5 Fairness, Stability, and Trustworthiness 54

4 Empirical part 55

4.1 Selection of Industry and respondents 56
4.2 Data collection 57
4.3 Data Analysis and Pattern Recognition 60

5 Research Results 60

5.1 Status quo of the AI adoption by legal and auditing firms 61
5.2 Reflect and validate the robustness of the AI development process 64
5.3 Reflect and validate the user's trust problem 66
5.4 Interpretation and recommendations 69

6 Conclusion 71

7 Bibliography 73

Table of Figures

Figure 1. Source MIT Sloan School Survey 11
Figure 2 Three levels of AI 17
Figure 3 Source: (Mithchel-Guthrie, 2014) (Morrison & Rao, 2016) 19
Figure 4. Source: (Morrison & Rao, 2017) 28
Figure 5 Source: (Morrison & Rao, 2017) 29
Figure 6 Source: (Larose & Larose, 2015) (Morrison & Rao, 2017) 30
Figure 7 Source: (Morrison & Rao, 2017) (Pierce, 2019) 32
Figure 8 Source: (Morrison & Rao, 2017) 32
Figure 9.ANN. Source: (Diepen, et al., 2017) 34
Figure 10 Source: (Diepen, et al., 2017) 35
Figure 11 Source: (Diepen, et al., 2017) 35
Figure 12. XAI Concept. Source: (Gunning, 2018) 41
Figure 13. Interrelationship of the 7 requirements for trustworthy AI. Source: (AI HLEG, 2019) 42

Table of Tables

Table 1.Visualisation Techniques. Source: (Hall & Gill, 2018) ... 47
Table 2Visualisation Techniques. Source: (Hall & Gill, 2018) 47
Table 3. Techniques for creating White-box models. Source: (Hall & Gill, 2018) 49
Table 4. Techniques for seeing model mechanisms with model visualizations. Source: (Hall & Gill, 2018) 51
Table 5. Reason code generation. Source: (Hall & Gill, 2018) .. 54

List of acronyms

ACM FAT: Association for Computing Machinery for Fairness, Accountability and Transparency 38
ACO: Ant Colony Optimization 27
AGI: Artificial General Intelligent 17
AI: Artificial intelligence 1
ANI: Artificial Narrow Intelligence 17
ANN: Artificial Neural Networks 9
DAPRA: Defense Advanced Project Research Agency 38
DL: Deep learning 9
GBM: Gradient Boosted Models 36
LIME: Local Interpretable Model-Agnostic Explanations 52
ML: Machine Learning 15
NLP: Natural Language Processing 10
SVM: Support vector machines 35
XAI: Explainable artificial intelligence 37

1 Introduction

The belief that human desire to obtain knowledge contradict the laws of God or nature must eventually lead to the disaster takes roots in Myths of ancient Greece. It has persisted throughout history and deeply ingrained in human culture. Thus, it should not be unexpected that Artificial Intelligence encourages so much dissension within both academia and popular circles (Luger, 2009).

Moreover, it is natural for people to distrust what they do not fully understand, and there is a lot about AI what is not immediately apparent. Starting with the ambiguous term and continuous with puzzling structures and strategies for complex problem-solving.

This chapter sketches the history of AI, introduces basic concepts, the current state of AI development, and provides an outline of the thesis.

1.1 A concise history of AI

The history of AI takes its roots in Aristotle's logic and in attempts of Philosophers and Mathematicians to represent human thought and reasoning in symbols. One of the seasonable steps toward AI development was the introduction of formal logic by Gottfried Wilhelm von Leibniz in the 17th century, with his *Calculus Philosophicus* and the machine for Automating his tasks, this mechanical solution was represented as movements through the states of a tree or graph (Leibniz, 1887).

In the 18th century, the study of representations was introduced by Leonhard Euler (Euler, 1735), who showed the structure of relationships and their connections in the world as the distinct steps within a computation abstractly. This allowed the possibility

of the primary conceptual tool of AI - state space search. State space search is used to model a deeper structure of a problem, where the nodes represent possible stages of a problem solution. According to George F. Luger, it is "a powerful tool for measuring the structure and complexity of a problem and analyzing the efficiency, correctness and generality of solution strategies" (Luger, 2009, p. 10).

Another mathematician of the nineteenth century, who made computer science and AI possible, is George Boole. He designed an extraordinarily powerful but a simple system in his work *"An investigation of the Laws of thought*," on which the *Mathematical theories of logic and Probabilities* were based (Boole, 1854). Boole created three primary operations: "AND" denoted by * or ^, "OR" denoted by + or $\vee$, and "NOT" denoted by $\neg$, these operations remain the basis for all further developments in formal logic and modern computers. He also concluded that logic has its laws, where $X * X = X$ if something is true, repetition cannot increase that knowledge. Which led to only two numbers, that could satisfy Booleans equitation: 1 and 0. That provided the basis of the binary arithmetic, as well as demonstrated how such simple system can capture the full power of logic (Luger, 2009, p. 11).

The term Artificial Intelligence was coined in 1955 by John McCarthy, who was one of the AI pioneers (Ertel, 2008, p. 4).

The first modern workshop on AI took place at Dartmouth College in the summer of 1956 and brought together practitioners who were focused on the integration of computation and intelligence. This conference addressed topics such as complexity theory, neuron nets, machine learning, abstractions, language design, and others, which lately formed computer science. Many of the

computer science's characteristics have their roots in AI (Luger, 2009, pp. 35-37).

The expectations were promising, and by the mid of the 60s, AI research programs were heavily funded by the US Department of Defense (McCorduck, 2004, p. 131). However, the development was not so rapid as researches predicted, as the representation of human thoughts demands many computations. Thus, the performance of old AI systems suffered from the reliance on fragile and unjustified symbolic representations, insufficient data and severe limitations of memory capacity and processor speed led to the first "AI winter" (Bostrom, 2014, p. 8). In 1974, governments cut off exploratory researches in AI, and the progress had slowed down (Russel & Norvig, 2003, pp. 22-24).

In the early 80th Japan started working on its well-funded Fifth-Generation Computer Systems project, and developing a massively parallel computing architecture, that could serve as a platform for AI. Many other countries under the influence of Japanese's "post-war economic miracle" followed the trend, and started investing in AI again, which resulted in the rapid expansion of Expert systems. However, these projects failed to meet their objectives, because small systems had little benefits, and the large ones were too expensive to develop, keep updated, validate, and were complicated to use (Bostrom, 2014, p. 8).

In the 90s, AI started attracting not only governmental but more and more commercial funds due to increasing computational power and the introduction of new techniques, which included neural networks and genetic algorithms. New alternatives to the traditional logic paradigm, were offered by new technics, which boosted a more organic performance. For example, neural

networks displayed attributes of "graceful degradation," which means that the small damages to a network typically resulted in a slight deterioration of its performance, rather than in total collapse. Moreover, neural networks could generalize naturally from examples, detecting hidden statistical patterns, and learn from experience. Besides, a backpropagation algorithm was introduced, which allowed us to train multi-layered networks. (Bostrom, 2014, p. 9).

Despite having notable advances in Expert Systems[1] in the 80s and neural networks in the 90th (Russel & Norvig, 2003, p. 22), AI labs had been struggling roughly during four decades.

In 2011, when IBM Watson[2] defeated greatest Jeopardy![3] Champions, it started to gain popularity and drawing general public attention (IBM Journal of Research and Development, 2012). This success proved outstanding advances in Artificial Neural Networks (ANN) and Deep learning (DL) and attracted a new wave of investments to these particular areas, which, in turn, started boosting further the scientific development.

Another outstanding progress was in 2016 – AlphaGO Zero[4], this system was trained through pure self-play and quickly surpassed human level, by using a novel form of reinforcement learning, a neural network combined with a powerful search algorithm (Hassabis & Silver, 2017).

---

[1] – rule-based programs that made simple inferences from a knowledge base of facts, that emulates the decision-making ability of a human expert (Bostrom, 2014, p. 8)
[2] IBM Watson – Question answering system, based on in-depth content and Natural Language analyses, Information Retrieval
[3] Jeopardy – Quiz TV show in the US
[4] AlphaGo – the first computer program that defeated a world champion at the ancient Chinese game of GO, extremely complex game

The latest advances in: ANN, DL, reinforcement learning, Natural Language Processing (NLP), face recognition, constant improvement of algorithms, as well as the development of cloud computing, and constantly growing computational power – marked the completion of significant milestones in the AI development, as a result of almost 70 years of research in this field.

Thanks to the focus on solving specific problems, AI and ML began to be used in many fields. There are hearing aids with algorithms that filter out ambient noise; route-finders that show maps and assist with navigation; recommender systems that offer books and music based on a user's purchases and rating history; medical decision support systems that help doctors diagnose cancer, suggest treatment plans and assist in the electrocardiogram interpretation. Face recognition used at automated border crossing control systems in EU and Australia, the US Department of state use it for visa processing with over 75 million photographs. Automated stock trading systems operate in the global financial market (Bostrom, 2014, p. 18).

Also, it became noticeable that early adopters of AI like Google, Amazon, and Facebook went far ahead of competitors. Venture capitalists, private investors, and corporations started allocating money into AI-powered software. By 2017 massive hype around the industry was created.

However, the 2017 global survey of 3000 executives, managers and analysts conducted by MIT Sloan Management Review[5] revealed "the huge gap between ambition and execution at most

[5] *The 2017 Artificial Intelligence Global Executive Study and Research Project, in collaboration with The Boston Consulting group*

companies." Almost 85 % of respondents believe AI will allow them to sustain or obtain a competitive advantage; however, less than 39% have an AI strategy in place. In reality, majority of the companies, in particular, 54%, have not started adapting AI technologies yet, 23% have sort of pilot project in progress, and only less then quarter adapted it (See Figure 1).

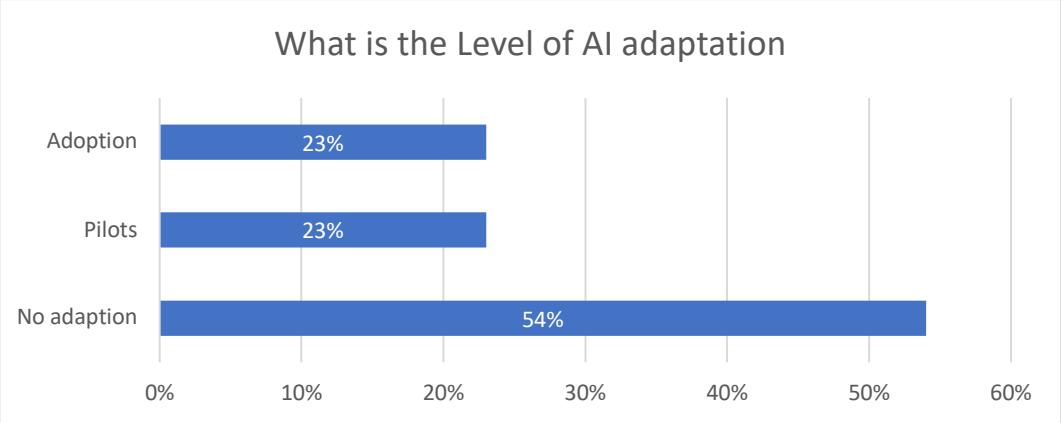

Figure 1. Source MIT Sloan School Survey

Several executives claimed that "the culture change required to implement AI will be daunting" (Ransbotham, et al., 2017).

This hype was partly pushed by those who already invested in AI technology on the wave of emerging Deep Learning, but without understanding that AI is a complicated and time-consuming R&D process. They wanted to boost the capitalization of their companies, return investments, or sell their products. Many products with inbuilt AI algorithms were released to the market, especially in the US. In many cases, these tools did not provide significant improvement for users, or they demanded knowledge in logic, data analysis, decision-making, and basics of programming. Moreover, many of them were not reliable because they were trained on uncleaned data.

The first signals of errors and biased results presented by algorithms were revealed back in 2010. As a result of giving a sensible assumption to the algorithm, which turned out to be incorrect, the financial market collapsed (2010 Flash Crash)(Bostrom, 2014, p. 20). Another example is the usage of historically biased data and deep learning model, for instance, computer vision and face recognition, which is struggling with recognition of dark-skinned people (Buranyi, 2017).

Such news became a trigger for further speculations about AI in mass media regarding adverse uses, fatal errors, jobs deprivation, and human extinction. Many decisions in this world are complex and controversial, can people rely on math, probability theory, and logic in such tasks? Is the danger manageable?

1.2 Problem formulation

Unfortunately, the arguments presented above, whether it is overestimated expectations, misuse or unjustified hoaxes, create an unrealistic and often negative perception of AI, undermine the reliability of technologies fundamentally and lead to skepticism and humane reluctance to its adaption.

Many machine-learning models that underlie AI applications are considered as "black boxes." People cannot always understand how exactly a given machine-learning algorithm makes decisions or generates results (Rao & Golbin, 2018). Such incomprehension of technology prevents the usage of the AI-powered systems or lead to misuse in various fields.

Therefore, enterprises, which intend to utilize AI-systems, should be able to rely on three main aspects: *explainability* – understanding reasoning behind; *transparency* – understanding of

AI model decision making; *provability* – mathematical certainty behind decisions (Rao & Golbin, 2018).

However, under dynamic market conditions, companies continuously face the challenge to innovate and be early adopters of prominent AI-tools or develop their own. Often, they do it without a profound risk assessment.

This thesis addresses the problems of reliability AI-systems with the intersection of people's distrust as a consequence of a limited understanding of technologies and their wrong implementations. These results in the following research questions:

1. Could be AI-powered systems reliable, transparent, and provable, if yes in which use cases?
2. Which critical preconditions need to be placed to enable thoughtful AI adaptation and build users trust in AI?

1.3 Aims and objectives

One of the goals of this master thesis is to obtain some clarity regarding whether public and private enterprises can safely harness AI technology or not. To answer this question will be needed:

1. Understand the main principles of work, what is behind reasoning and decision-making processes;
2. Evaluate which models could be considered as the most reliable;
3. In which areas it can be used harmless to business and society;
4. Find out if there are any technical solutions or the framework which could support responsible AI development and application.

5. As AI-powered systems have become the critical aspects of business development strategies in many large enterprises, many of them are already encountered problems of AI adaptation, such as insufficient knowledge, lack of talents, underestimation of complexity and user's distrust.

Therefore, the aims of the empirical part of this thesis are:

1. Identify obstacles on the way of building user's trust in AI;
2. Propose a methodological and applicable framework for thoughtful AI integration into daily operations.

1.4 The course of the investigation

To conduct a consistent analysis of the logic behind and structure of AI models in order to gain knowledge and a better understanding of their reliability and transparency; assess the scope of the problem of trust in AI and find possible solutions - a conceptual analysis method was chosen.

Literature related to the AI basics, existing researches, papers, and surveys on AI adaptation, transparent and responsible AI, and use cases was collected and analyzed.

The data collection process was organized by means of searching academic databases, libraries, and the World Wide Web to verify that sufficient articles and surveys are available to justify a literature review. The search revealed an acceptable amount of literature.

As sources academic databases, the libraries of Vienna University of Technology (http://www.ub.tuwien.ac.at/eng/), Vienna University of Economics and Business (http://www.wu.ac.at/en/library/), were used. Additionally, a

review of online resources led to websites maintained by organizations that provide material with a high degree of relevance to this thesis and helped to frame the research problem. For this study, resources made available through the websites of PwC US blog, Landing AI, O'REILLY, MIT Sloan Management Review, and Coursera were beneficial.

Also, the information retrieved from Journal of the Association for Information Systems, Journal of Information Technology, and several home pages and white papers of AI platforms or software providers like Google, SAS, CNC, Oracle were considered.

1.5 Structure of the thesis

Part 1. In this chapter, the concise history of AI, the current state of AI development, and an outline of the thesis was provided.

Part 2. Chapter two is dedicated to the comprehension and explanation to people from outside the field what AI can and cannot do. The role that Machine Learning (ML) plays in AI development; five "tribes" of ML; how ML works; the description of major ML methods; use cases and best practices. Problems and level of uncertainty related to the results generated and possible technical solutions for historically biased data will be addressed in this chapter.

Part 3. In this part described the ethical principle of AI development, Explainable AI, Machine Learning interpretability, and guidance of trustworthy AI.

Part 4. Empirical research on the problems of building user's trust in AI, and obstacles on the way of its mindful integration.

Part 5. List of recommendation to build users' trust in AI and incorporation of AI-friendly culture in enterprises.

Part 6. Conclusion

2 AI and Machine Learning

According to John McCarthy, who was a pioneer of AI and the first person who defined this term in 1955: "The aim of AI is to develop machines that behave as if they had intelligence" (Ertel, 2008). The problem with this definition is the lack of clarification of what intelligence is. Hence, AI, as an ambiguous term, has a lot of different interpretations.

Another definition suggests that it is "a study of how to make computers do things at which, at the moment people are better" (Rich, 1983), which reflects the real application of the study. However, it is better to refer to the definition of intelligence itself.

Professor Linda Gottfredson describes intelligence as "a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractedly, comprehend complex ideas, learn quickly, and learn from experience" (Gottfredson, 1997). Moreover, she clarifies that it is not related to simply book-learning or test-taking, (narrow academic skills), "rather it reflects a broader and deeper capability to comprehend our surroundings – "catching on," "making sense" and "figuring out" what to do "(Gottfredson, 1997).

According to this definition, not necessarily every human has intelligence. The process of acquiring knowledge (learning) represents only "narrow academic skills," but not intelligence. The similar classification is commonly applied to machines or artificial intelligence, which suggests three levels of AI development (See Figure 2) and help to answer the question of what AI currently can and cannot do. Nowadays, it is not very popular to claim this, but

despite having tremendous success and improvement, AI evolution is still at its initial stage. In parallel to human evolution, let say it is close to the Neanderthals.

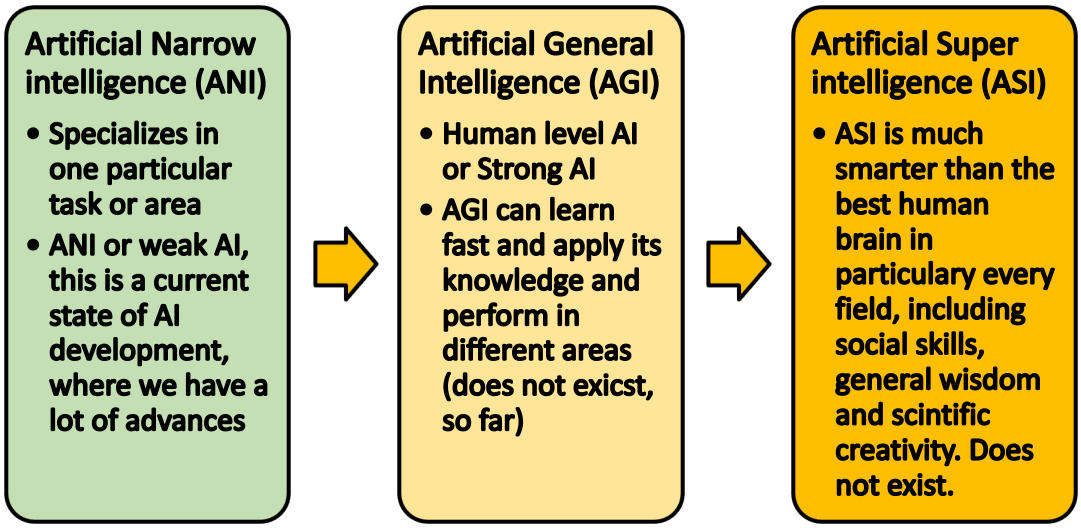

Figure 2 Three levels of AI

Essentially all the systems that exist now are Artificial Narrow Intelligence (ANI). ANI is single-tasking and quite far away from the definition of intelligence, but it contains components such as classifiers, search algorithms, planners, solvers, and representational frameworks – that might play a role in the future development of Artificial General Intelligent (AGI) (Domingos, 2015, p. 6).

Currently, the distinction between ANI and generic software is not sharp, or as McCarthy's dictum says, "when something works, it is no longer called AI" (McCarthy, 2007). It is almost impossible to find modern software without incorporated ML-technics; otherwise, this software can be considered as outdated. Therefore, it would be better to differentiate between systems that have "a

narrow range of cognitive capability[6]" and those that have generally suitable problem-solving potentiality. As Domingos argued, "all learned knowledge is uncertain, and learning itself is a form of uncertain inference" (Domingos, 2015, p. 52), each new feature can influence previously acquired knowledge and affect the result.

Given that, no matter whether it is learning humans or learning algorithms; the inference learned must be proved or validated. To become an architect, people have to validate their knowledge by passing exams and then by acquiring practical experience. Otherwise, it could lead to fatal mistakes. The same with algorithms, machine-learning systems, or AI, they must be validated, or at least properly defined boundaries for usage must be created; otherwise, they can harm.

To identify the boundaries of safe AI implementation, we must know as much as we can about the main attributes of all fundamental AI technics. These techniques initially represent Machine Learning, though some of them developed into separate research fields.

Machine Learning

Deep Learning or neural networks, knowledge extraction, supervised, unsupervised and reinforcement learning, recommender systems, pattern recognition, clustering and labeling, natural language processing and a word algorithm itself – are all related to ML. They often used without a proper explanation as marketing traps to attract customers, and

[6] Learning, information processing speed, the verbal and spatial ability

therefore are reasons for confusion or overestimated expectations.

To understand how various areas of computer science are interconnected, see Figure 3.

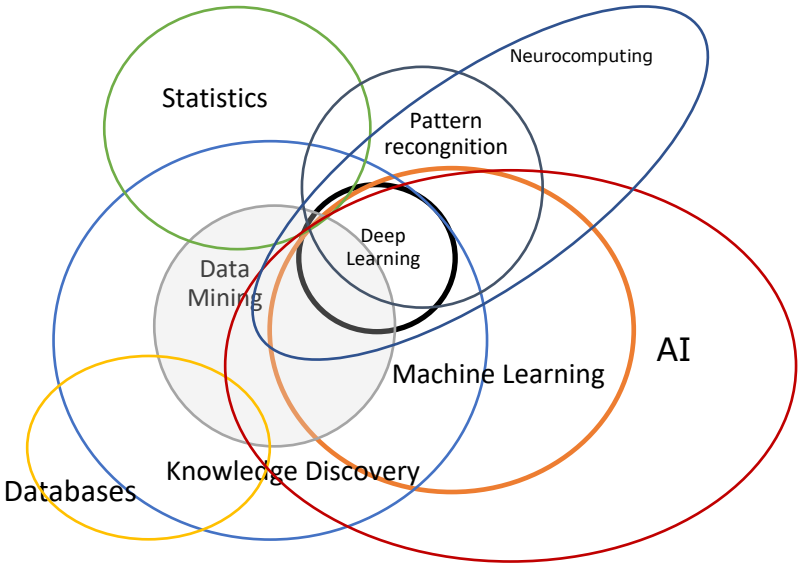

Figure 3 Source: (Mithchel-Guthrie, 2014) (Morrison & Rao, 2016)

In computer science, an algorithm defined as a sequence of instructions telling a computer what to do (Domingos, 2015, p. 1). Computers are hardware made of billions of tiny switches called transistors, which can execute instructions by turning these switches on and off billions of times per second. The state of one transistor is one bit of information, where 1 is ON and 0 is OFF. For instance, one bit can say whether your bank account is overdrawn or not. When a transistor A turns ON only if transistors B and C are ON, it is doing a tiny piece of logical reasoning. By combining millions of such operations, we can carry out very intricate chains of logical reasoning (Domingos, 2015, pp. 4-6).

The process of designing algorithms abound with pitfalls – any wrong assumption or a computing mistake make the correct output impossible. Debugging7 is the time-consuming but initial part of this process. Over time computer scientists invent algorithms for new thinks, built them on previous ones, combine with others, which in turns producing results for still more algorithms. The more sophisticated algorithms are, the more difficult it is to validate them. When they are too complex for poor human brains to understand, we cannot find the mistakes (Domingos, 2015, p. 5).

Traditionally, the workflow of algorithms is represented as follow – the data as input goes into the computer >> the algorithm processes the data >> and the results come out. ML turns this around – given the data and desired results, goes out the algorithm that creates the other algorithms (because it learns), that is, the complexity can grow infinitely (Domingos, 2015, p. 6).

Therefore, the tension regarding algorithmic opacity and distrust in society is justified, but even non-transparent algorithms could be acceptable to use, given that it constrained properly (Robbins, 2019).

All algorithms that underlie various systems should be considered as many other technologies, as when they utilized with violation of user's manual and safety precautions, it can because of harm. The only difference is that for algorithms, we do not have such regulations so far. To proceed further with thoughtful integration

[7] To debug – find every error, and fix it until the computer runs your program or commands without screwing up (Domingos, 2015, p. 4)

and address the question of the algorithm's reliability, people have to acquire at least basic knowledge about ML technics.

2.1 How Machine Learning works

People acquiring knowledge by means of reading, watching, listening, and experiencing, that is, processing the information or data – identifying the patterns, generalizing, and making inferences. Machines do the same, and in order to start learning, machines need data.

2.1.1 Data selection

The first step in ML is data selection. Chosen data must be divided into three groups: training data, validation data, test data. The training data is crucial regarding how that algorithm will work. Two identical algorithms that share the same code could present absolutely different results because they were trained using different datasets. A facial recognition system trained on pictures of young white men would not work correctly for old black women (Robbins, 2019).

The reliability of algorithms and their robustness depend directly on the quality and diversity of the training data set. Therefore, all users have to understand the importance of this, before such algorithms will be integrated into the daily routine. The information about training data set provided by developers (e.g., number of faces, breakdown of age, sex, ethnicity, etc.), could play an essential role in building trust in AI and its mindful implementation (Robbins, 2019).

2.1.2 Data modeling

The next step in ML would be data modeling. To build an abstract model, we need a training data set to identify different features of

data related to the problem, standardize their relations to one another and the attributes of the real-world entities. For instance, the data set representing a house price consist of different features that affect the price, such as size, number of rooms, location, and transport hub. The learning algorithm extracts the rules on how prices correlated to these features.

When developers do not control the learning process, and the algorithm continues to identify more and more features, at some point it can demonstrate ridiculous inferences (e.g., if the owner is women >> the price can be lower because women earn less than men). In order to avoid such conclusions, the number of features must be limited, or in supervised learning, it can be equated to 0, consequently considered as a not important attribute.

2.1.3 Validation

To validate the model – means asses the model with the validation data set, check how good the level of accuracy, and improve the model by training it. That means we do not believe anything what learning algorithms tell us until we have verified it on data that learner did not see. If hypothesized patterns hold true on new data, we can be confident in results, if not the model overfit.

This is just a standard scientific method applied to ML. When scientists come up with a new theory it is not enough to explain past evidence, the theory must also make new predictions, and only if it was experimentally verified, it can be accepted (and even then only temporarily, because future evidence could still falsify it) (Domingos, 2015, p. 75).

Therefore, models must be trained through an iterative process of providing positive and negative feedback on the results they give (Auerbach, 2015).

Test model. After the validation, the model's performance must be checked with test data and the level of accuracy must be assessed.

Use the model – deploy the fully trained model to make predictions on new data.

Tune the model – improve performance of the algorithm with more data, different features, or adjusted parameters and try to increase the level of accuracy (Morrison & Rao, 2016).

2.1.4 Accuracy

When we think about 100 % of accuracy, it is more likely about traditional programming than about ML. Traditional programming has always been about thinking deterministically, whereas ML is about thinking statistically. For instance, if the spam classifier is 99% correct that does not mean it has bugs, it means it is the best it can do, and it is good enough to benefit from it. Such difference in thinking was one of the biggest Microsoft problems, why it has struggled with catching up Google than it did with Netscape[8]. Web browser itself just a piece of standard software, but a search engine required a different mindset, which for many old school programmers was inconceivable (Domingos, 2015, p. 9).

The same paradox of mindset problem we can see now from the users' side, who think that machine should provide 100% result,

[8] Netscape – was the dominant web browser in term of usage in the 1990s, but by 2002 its market share shrank to 0.6% (Jay, 2008).

and if it is not, it is useless. At the same time, human accuracy in many decision-making tasks is far away from 100 %, and decisions are very subjective, and it is a lot worse than demonstrate AI. Such limited "black and white" way of thinking in correlation to acceptance of probabilistic approach as a useful tool, must be addressed in culture changing aspect for AI adaptation, and benefits must be explained.

The probabilistic approach and generalization are the main advantages of ML, as the opposite to memorization and exhaustive search, which on a big scale, are impossible. At the same time, the main ML problem is generalizing to cases that we have not seen before (Domingos, 2015, p. 62).

2.2 Machine learning methods

Through mathematics and logic, humans represent their thought abstractly. Every day people make thousands of various decisions, simple and complex, which predetermine future events. All algorithms represent the way how naturally humans or other representatives of nature make decisions; they designed by humans with a purpose to find the right solution in limited time.

Can we rely on the human's decision? Of course, it depends on many variables that we will consider. However, since humanity still exists, history says – yes, we can. Despite knowing, that "human judgment is subject to cognitive limitations…and human strive to make rational choice" (Cherry, 2019), we still can rely on these decisions, even when they lead to fatal aftermaths, because by means of "trial and error" people can learn from mistakes and avoid them in the future.

Probably the same with algorithms, first they learn our mistakes and biases, but it is possible to correct them. The outcome could be not only an algorithm's impartial and sound decisions but a society based on integrity without any room for hypocrisy.

In order to avoid or at least to minimize a negative impact during algorithms evolution, it is necessary to understand how they learn.

2.2.1 Heuristics

The methods that demonstrated successes in early systems proved difficult to apply to more complicated problem examples or a greater variety of problems. "Combinatorial explosion" of possibilities, that is their rapid growth, that must be explored by methods which exploit exhaustive search[9], is one of the reasons for that (Bostrom, 2014, p. 7).

In order to prove a five-line long proof theorem in deduction system with 1 inference rule and 5 axioms – 3125 possible combinations must be checked to find out if it has the intended conclusion. For 5-, 6-, 7-line proofs exhaustive search could work, but with more complicated problems it will run into trouble. To prove a 50-line proof, the theorem does not require 10 times longer than a 5-line proof, only if we do not use exhaustive search. Because by using exhausting search we have to check $5^{50} \approx 8.9 \times 10^{34}$ possible sequences – this task computationally impossible even with the fastest supercomputers. Algorithms that utilize structure and take advantage of prior knowledge by using heuristic search, planning, and flexible abstract representations,

[9] Exhaustive search (brute force search) – based on generating and systematically enumerating all possible solutions and testing each of them, whether it satisfies the problem's statement or not (Fox, 2019).

were needed to overcome the combinatorial explosion (Bostrom, 2014).

Heuristic approach sometimes called cognitive laziness. Due to the limited amount of time and limited human's ability to process the volume of information available, people use mental shortcuts to make decisions — this approach based upon how easy to bring something in mind (Cherry, 2019). For example, recently acquired information that popped up in our mind, or the information that is not representative, but impressive could influence the decision; or temptation to generalize – if "a yoga teacher," then "vegetarian" – which is not necessarily true, but the probability is high. That is how the human brain works and why human decisions are biased and not ideal in many cases. If people use the same approach when expressing the decision-making process in algorithms, algorithms reflect these flaws as well.

As the heuristic approach the easiest method of decision-making, it used a lot in AI and ML. There are two types of Heuristic technique in ML or AI: tailored heuristic and generic heuristic.

Tailored heuristic is used for specific tasks, for example, to determine the minimum number of coins for a particular amount of money. We have four denominations of coins 5,4,3 and 1, and algorithm have to find the best solution for the amount of 7 cents. As heuristics is a "lazy" technique, the algorithm will use a greedy heuristic, which means the most significant number will be chosen first. In this example, 5 is the biggest number, so the algorithm will repeat the process of adding up coins to 5 until it reaches 7. The only possible solution here is to add two 1 cent coins, and the answer is 3 coins needed to form the desired amount. However, with 3+4, we will need only 2 coins. In this case, the algorithm

would not find the best solution, but the first sufficient solution (Diepen, et al., 2017).

Heuristic method utilized a lot, for instance in Genetic Algorithms or Ant Colony Optimization[10] (ACO). Real ants lay down pheromones pointing each other to resources while investigating territory, artificial ants or agents act similarly, they record their position and solution quality, to allow other agents in later iterations locate better solution (Dorigo, 2007). ACO algorithms serve good for logistic purposes. For instance, vehicle routing, with the purpose to find the routes for one or more vehicles visiting a number of locations (Diepen, et al., 2017).

Heuristic approach is not very elaborated or sophisticated, but for many cases, it is good enough. If problems get too complex, and there are no exact methods to find a guaranteed best possible solution, this technique serves to find maybe not optimal, but satisfactory solution for immediate goals.

2.2.2 Classification algorithms

In ML, statistics, and data analysis, classification is a supervised learning approach for learning classifiers from examples – "decision trees", "logistic regression models", "support vector machines", "naïve Bayes", "k–nearest–neighbors regression", "boosted trees", "random forest" and neural networks – are among others, that represent this large class of algorithms.

[10] ACO – a population-based high-level procedure that can be used to find an approximate solution to a difficult optimization problem, where a set of software agents called artificial ants search for a good solution for a given optimization problem (Dorigo, 2007).

2.2.2.1 Regression

Regression maps the behavior of a dependent variable relative to one or more dependent variables. For example, logistic regression separates spam from the non-spam text (See Figure 4).

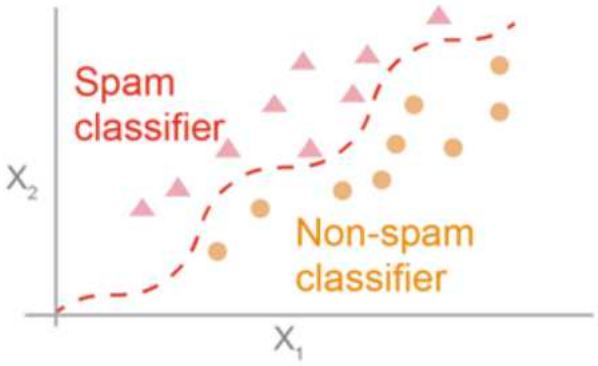

Figure 4. Source: (Morrison & Rao, 2017)

Advantages. Regression is useful for identifying continuous (not necessarily distinct) relationships between variables.

Use cases. Traffic flow analysis, email filtering

2.2.2.2 Random forest

Random forest algorithms improve the accuracy of the decision trees by using multiple trees with randomly selected subsets of data. This example reviews the expression levels of various genes associated with breast cancer relapse and computes the relapse risk.

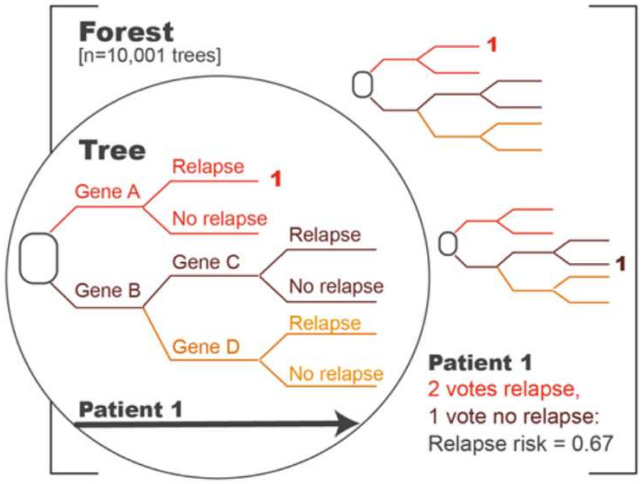

Figure 5 Source: (Morrison & Rao, 2017)

Advantages. Prove useful with large data sets and items that have numerous and sometimes irrelevant features.

Use cases. Customer churn analysis, risk assessment

2.3 Five paradigms of Machine Learning

Pedro Domingos in his book "The Master Algorithm" outlines five contemporary Machine Learning paradigms – symbolism, connectionism, evolutionary algorithms, Bayesian networks, analogical reasoning – which he supposed could be combined in one future "master algorithm" capable of learning everything, in other words, an AGI or even ASI algorithm (Domingos, 2015).

2.3.1 Symbolists. Decision trees

Symbolists tribe – use symbols, rules, and logic to represent knowledge and draw a logical inference. This paradigm predominated in the 1980s with rules and decision trees algorithms; worked on server or mainframe architecture; applicable in expert systems, knowledge databases, inference engines, decision support systems with limited utility.

2.3.1.1 Decision Trees

A decision tree is a predictive model, that utilizes a hierarchy of variables or decision nodes and can be represented in the form of if-then rules (Morrison & Rao, 2017).

For example, let us take credit scoring, it is important for a bank to predict the risk associated with a loan. Given the amount of loan and the information about the client, a bank can classify a given customer as a high- or low-risk group representative. To do so, we should select the data; it must be data related to the customer's financial capacity – profession, age, income, savings, collateral, credit history. (Alpaydin, 2016, pp. 45-50).

Then we should do data modeling. In order to model our data, we have to look at the past transactions and label the customers who paid back their loans as creditworthy, and who defaulted as not. Then analyzing the data, we will learn the class[11] of high-risk customers and particular features that assigned to them. We assume that these characteristics are not shared by low-risk customers, and that we can determine the class by the presence of these characteristics, called discriminant (Alpaydin, 2016).

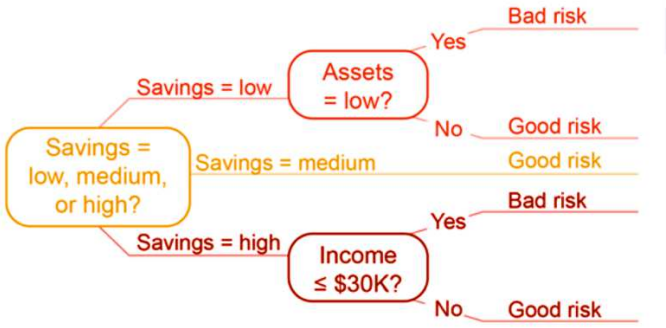

Figure 6 Source: (Larose & Larose, 2015) (Morrison & Rao, 2017)

[11] A class – a set of instances that share a common property (Alpaydin, 2016, p. 46)

It is impossible, at least now, to have access to complete knowledge about all the factors that could influence the decision, and deterministically calculate 100% impartial result. However, we can calculate the level of certainty, for instance – 97%.

If one of the discriminants in a given algorithm is a place of residence, and the majority of people from the given district have low income and classified as high-risk customers, even if you have a good income, your loan application probably will be rejected.

Decision trees are useful when evaluating lists of distinct features and qualities of people, places or things. However, in a case with people, discrimination may definitely occur.

2.3.2 Bayesians. Naïve Bayes and Markov

Bayesians – assess the likelihood of occurrence for probabilistic inference. Bayesians tribe predominated in the 1990s and 2000s; favored algorithms Naive Bayes or Markov based on probability theory; scalable comparison and contrast that good enough for many purposes, for example, spam classifier.

2.3.2.1 Naïve Bayes classification

Naïve Bayes classifiers compute probabilities, given tree branches of possible conditions. Each individual feature is "naïve" or conditionally independent of, and therefore does not influence, the others. For example, what the probability you would draw two yellow marbles in a raw, given a jar of five yellow and red marbles total? The probability, following the topmost branch of two yellow in a row, is one in ten. Naïve Bayes classifiers compute the combined, conditional probabilities of multiple attributes.

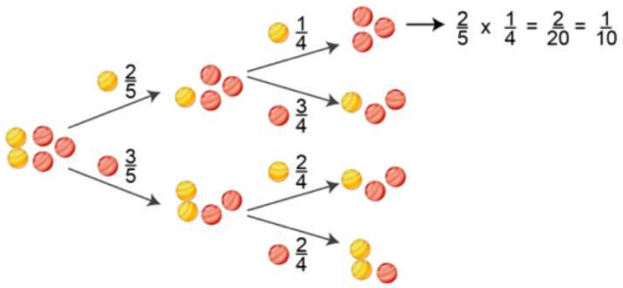

Figure 7 Source: (Morrison & Rao, 2017) (Pierce, 2019)

Advantages. Naïve Bayes methods allow the quick classification of relevant items in small data sets that have distinct features.

Use cases. Sentiment analysis, consumer segmentation

2.3.2.2 The Markov decision process (Hidden Markov model)

Observable Markov processes are purely deterministic – one given state always follows another given state. Traffic light patterns are an example.

Hidden Markov models, by contrast, compute the probability of hidden states occurring by analyzing observable data and then estimating the likely pattern of future observation with the help of the hidden state analysis. In this case, the probability of high or low pressure (the hidden state) is used to predict the likelihood of sunny, rainy, or cloudy weather.

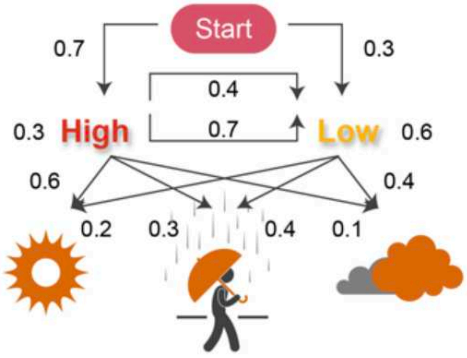

Figure 8 Source: (Morrison & Rao, 2017)

Advantages. Tolerates data variability and effective for recognition and prediction.

Use cases: facial expression analysis, weather prediction.

2.3.3 Connectionists. ANN

Connectionists – recognize and generalize patterns dynamically with matrices of probabilistic, weighted neurons; allowed more precise image and voice recognition, translation, sentiment analysis, etc. This paradigm was widely distributed in the early and mid-2010s, thanks to cloud architecture and higher computational power; favored algorithms neural networks; based on neuroscience and probability theory.

2.3.3.1 Artificial Neural Networks

The human brain was the primary source of inspiration for neural networks. ANN are some of the most powerful learning algorithms ever developed, but they are also some of the most complex. The hierarchical and nonlinear transformations that applied to data can be sometimes incomprehensible. Very often, small changes in network hyperparameters can dramatically affect the ability of the network to learn.

Weights are used to indicate the strength of the connection between neurons, ANN identifies which neurons must be in the next level. ANN learning involves changing the weight between neurons. By providing a broad set of training data with known characteristics to ANN, it is possible to calculate the best weight of artificial neurons to ensure that ANN recognizes these characteristics best (Diepen, et al., 2017).

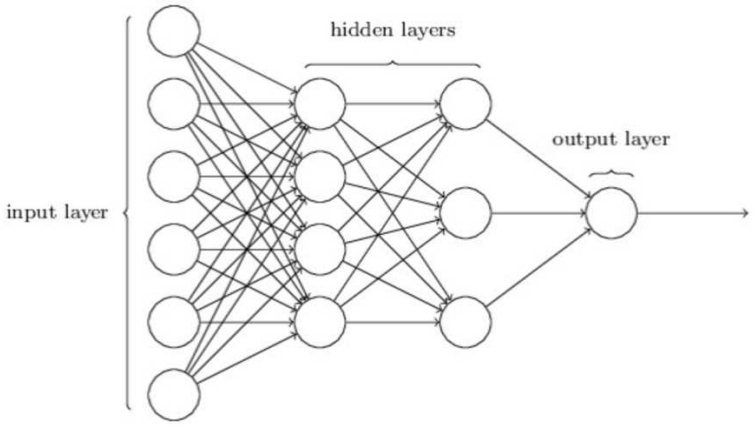

Figure 9.ANN. Source: (Diepen, et al., 2017)

The lack of model interpretability and a large amount of data to learn are the two most significant flaws of ANN.

2.3.4 Evolutionaries. Genetic algorithms

Evolutionaries – natural selection is the basis of this method. Generate variations and then assess the suitability of each for a given purpose in order to evolve, each generation is fitter than the previous one; in contrast to others genetic algorithms are full of random choices, and better able to come up with something truly new. One of the examples of genetic algorithm is ACO were described earlier (p.27).

2.3.5 Analogizers. SVM

Analogizers – recognize similarities between situations and thus inferring other similarities between them; optimize a function in light of constraints. When two patients have similar symptoms, probably they have the same disease. The main problem here is to determine the level of similarity. Support vector machine algorithm is the main representative of this tribe.

2.3.5.1 *Support Vector Machines*

"Support vector machines (SVM) classify groups of data with the help of hyperplanes" (Morrison & Rao, 2017). The main idea is to find a boundary line, that maximizes the separation between the two classes. If we take a simple data set of green dots and red squares, which could show two different segments of customers, represented by all kinds of properties for each of them. Any line that can keep these two groups separated is contemplated as a valid boundary line. Therefore, an infinite number of lines could be drawn (See Figure 10) (Diepen, et al., 2017).

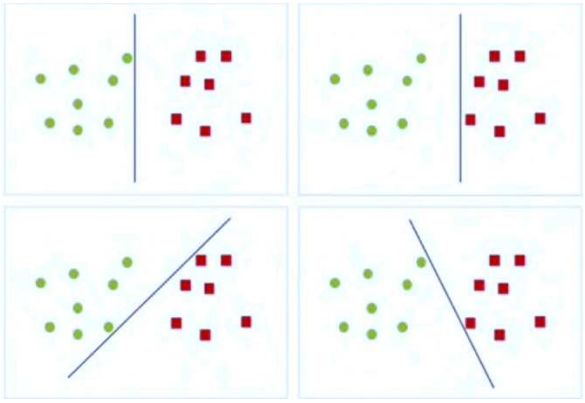

Figure 10 Source: (Diepen, et al., 2017)

However, we need only one with the largest separation space between them (See Figure 11).

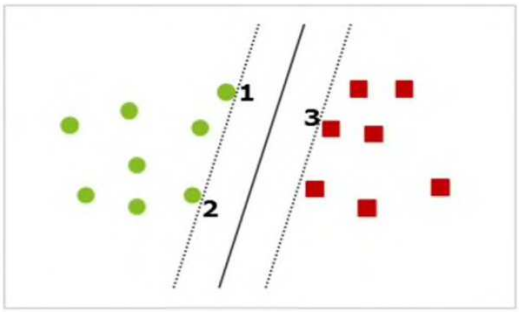

Figure 11 Source: (Diepen, et al., 2017)

In the example above, there are three support vectors, which determine a margin's boundaries, so the margin or hyperplane can act as a linear classifier.

SVP are good for the binary classification of X versus other variables and are useful whether or not the relationship between variables is linear (Morrison & Rao, 2017). SVM models can be used in news categorization, image recognition (handwriting converted to text, face recognition).

3 Explainable AI

In previous chapters were described how different scientific fields, including mathematics, logic, electrical engineering, AI and ML, as well as many other subfields of computer science – are all interconnected. The immense heritage of human knowledge and their complexity demand explicit techniques to validate human theories, assumptions, and observations and convert them into science.

A benchmark of good science in AI-field is comprehension and veracity of models and their results. Nevertheless, the pressure of innovation and competition are forcing data scientists to strive for ever-more sophisticated predictive modeling and ML-algorithms (Hall & Gill, 2018).

The problem of complexity partly was described in chapter 2 of this work, a fundamental trade-off between accuracy and interpretability concerns data scientists greatly. Very complex algorithms, including gradient boosted-ensembles (GBM), ANN and random forest, among many others, are more accurate for predicting nonlinear, faint or rare phenomena. Unfortunately, complexity is what makes these models accurate and make their

predictions challenging to understand. However, due to the critical importance of explainability for business adoption, user's acceptance and trust, AI-systems must not only demonstrate excellent performance, but convince people that it makes a decision for the right reasons, and it is worthy of our trust (Hall & Gill, 2018).

Although Explainable artificial intelligence (XAI) has become popular in the last few years, it is not a new field.

At least from the very beginning of C.S. Pierce's abductive reasoning utilization in expert systems of the 1980s, there has been an architecture of reasoning that supports an explanatory function for complex artificial intelligence systems, including applications in medical diagnosis, complex multi-component design, and real-world reasoning. (Goebel, et al., 2018). However, back then AI-systems did not have such a strong impact on our day-to-day lives. Therefore the development of XAI techniques was not of the utmost importance.

Nowadays, the commercial use and widespread of AI-systems, as well as recently discovered mathematical and sociological flaws in AI-systems, raised human concerns. Demands for interpretability, fairness, accountability, and transparency are of great importance for further AI adoption. How to match up the explanatory requirements of the different application – to be lawful, ethical and robust (AI HLEG, 2019) – with the capabilities of underlying ML techniques is the main XAI task.

3.1 Defining Interpretability and Trustworthy AI

In the context of machine learning, interpretability was defined as "the ability to explain or to present in understandable terms to a human" how results were generated (Doshi-Velez & Kim, 2017).

However, there are several communities with various advanced concepts of what interpretability or explainability is today and should be in the future (Hall & Gill, 2018). There are a few well-known groups of academics, that work on explainability research: the Institute for Ethical AI & ML, Association for Computing Machinery for Fairness, Accountability and Transparency (ACM FAT), and Defense Advanced Project Research Agency (DAPRA), and the High-Level Expert Group on AI organized by European Union Commission.

The Institute for Ethical AI & ML

The Institute for ethical AI & ML is a research center based in the UK, that conducts high-tech research in the field of responsible machine learning systems. It is an analytical center that brought together technology leaders, policymakers, and scientists to develop industry standards for data governance and ML. They identified 8 principles of responsible AI development and came up with a practical framework for technologists how to design, develop, and maintain systems responsibly. These principles include:

1. Human augmentation. Assess the consequences of wrong predictions and, if reasonable, design systems with human-in-the-loop review processes (e.g. justice, health, transport, etc.);

2. Bias evaluation. Develop processes that can help to understand, document and monitor computational and societal bias that is inherent in data, features, and inferences, which is impossible to avoid, but possible to document, diminish and exclude;
3. Explainability by justification. Develop tools and procedures to continually improve transparency and explainability of ML models, that allow to explain results;
4. Reproducible operations. Develop the infrastructure and set of rules, that allow ML systems diagnose and respond effectively when something bad happens with a model (e.g., reverting model to a previous version, or reproduce an input to debug a specific functionality, etc.);
5. Displacement strategy. Document relevant information to mitigate the impact towards workers replacement (e.g., a change-management strategy when rolling out the technology);
6. Practical accuracy. Ensure that accuracy and cost metric functions comply with the domain specific applications, which means to have a thorough understanding on underlying means to asses accuracy (what may be correct for a machine, may be wrong for a human);
7. Trust by privacy. Build and communicate processes that protect and handle personal data that may interact with a system directly or indirectly, properly;
8. Data risk awareness. Both types of security risks, human errors, and adversarial attacks must be taken into consideration during the development of ML systems;

In addition to this framework, the Institute for Ethical AI & ML created and maintain the open-source XAI library, where all ML-

tools and techniques are based on 8 principles for responsible AI and were designed with AI explainability in its core (The Institute for Ethical AI & ML, 2019).

ACM FAT

ACM FAT is an interdisciplinary community of researches and practitioners, who are mainly focused on promoting interpretability and fairness in algorithmic systems. However, their research challenges are not limited to technological solutions regarding potential bias, but also include ethical and philosophical aspects, as well as AI systems' social and commercial impact (FAT, 2019).

DAPRA

Civilian and military researches funded by DAPRA are primarily interested in improving interpretability in complex pattern recognition models required for security applications. The main objective of their XAI program is to create a set of new or modified ML-methods, that allow producing more explainable models. They believe that new ML systems will be able to explain their rationale, identify strength and weakness, and provide an understanding of their future behavior. These models will be combined with ultra-modern human-computer interfaces with the ability to convert models into understandable explanations for the end user (See Figure 12).

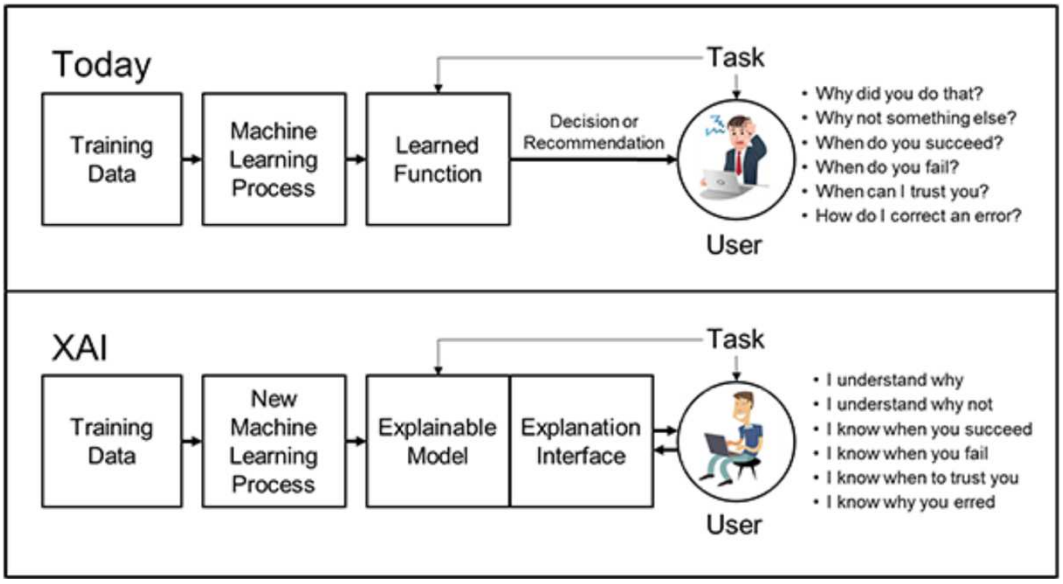

Figure 12. XAI Concept. Source: (Gunning, 2018)

XAI is one of the current DARPA programs expected to allow "third-wave AI systems," that can understand the context and environment in which they operate. At the end of the program, they expect to deliver a toolkit library, that will provide developers with techniques covering the "performance-versus-explainability" trade space, and human-computer interface blocks. This portfolio can be used for developing next-generation AI systems, that enable human users to understand, trust, and effectively manage AI partners (Gunning, 2018).

EU Commission. Ethics Guidelines for Trustworthy AI

EU commission published Ethics Guidelines for Trustworthy AI in April of 2019, this guide was written by the High-Level Expert Group on AI. This document is recommendatory, it aims to promote Trustworthy AI, and it does not constitute any legal or regulatory framework. The Guidelines lists 7 key requirements applicable to developers, deployers and end users, who are taking

a part in AI system' life cycle (AI HLEG, 2019)[12]. These requirements include human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination, and fairness; societal and environmental well-being; accountability.

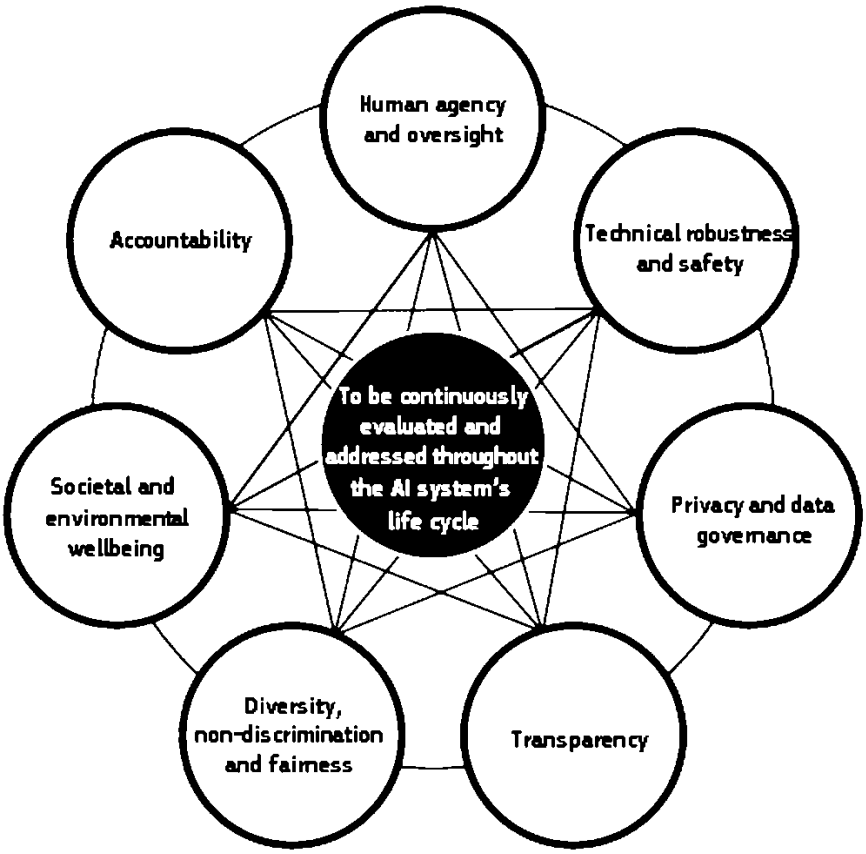

Figure 13. Interrelationship of the 7 requirements for trustworthy AI. Source: (AI HLEG, 2019)

3.2 Taxonomy for ML-Interpretability

An ML-interpretability taxonomy is a very subjective and complicated subject due to the diversity of applied practitioners. For some of them, technical description of algorithms, and techniques such as cross-validation, error measures, and

[12] AI HLEG – is an independent High-level Expert Group on Artificial Intelligence that was set up by the European Commission in June 2018 (AI HLEG, 2019)

assessment plots, provide enough insight to trust a model. For others, it says nothing and very simplistic, but a realistic explanation, perhaps with visualization must be provided (Hall & Gill, 2018).

Nevertheless, there is some standard classification of interpretability techniques that represent models: in terms of complexity; classify them by the global or local scope of explanation they generate; and differentiate them as model agnostic or model specific (Hall & Gill, 2018).

3.2.1 A scale of complexity

The problem of complexity is well-known already – the more complex the model is, the more difficult it is to explain. Vapnik-Chervonenkis dimension that represents the number of weights or rules in a model is a good way to quantify a model's complexity (Hall & Gill, 2018).

High interpretability. Linear, monotonic functions

Functions created by linear regression algorithms are possibly the most understandable class of models. These models called "linear and monotonic[13]", which means that when any independent variable (or sometimes a combination or function of an independent variable) changes, the response function changes at a certain rate, only in one direction and in a value represented by an easily accessible coefficient (Hall, et al., 2017).

Monotonicity also allows intuitive and automatic reasoning about predictions. For example, if a credit card application was rejected,

[13] A monotonic function is a function which is either entirely nonincreasing or nondecreasing. It is monotonic, if its first derivatives (which does not have to be continuous) does not change sign (Stover, 2019)

a bank can easily explain the reason, because its probability-of-default model often considers your credit rating, your account balance, and the length of your credit history, that is monotonically related to your ability to pay your credit card bills. When these explanations are created automatically, they are usually called reason codes. Linear and monotonic response functions also allow you to calculate the relative importance of the measures. These functions play an important role in ML-interpretability because they have several applications in explanatory techniques. (Hall & Gill, 2018)

Medium interpretability. Nonlinear, monotonic functions

Usually, machine-learned response functions are nonlinear, but they can be constrained to be monotonic in relation to any given independent variable. Although there is no single coefficient representing a change in the response function caused by a change in one independent variable, nonlinear and monotonic response functions always change in the same direction when one input variable changes. These functions enable the generation of reason codes and relative variable importance measures; hence, they are interpretable and most likely suitable in regulated applications (Hall, et al., 2017).

Low interpretability. Nonlinear, nonmonotonic functions

Unfortunately, the majority of ML-algorithms create nonlinear, nonmonotonic response functions. For any alteration on an input variable, these functions can change in a positive and negative direction and at a varying rate. Mainly, these functions can provide only one of the standard interpretability measures – relative variable importance measures. Therefore, to explain these complex models, a combination of interpretable techniques must

be used (Hall & Gill, 2018). Random forests and ANN are examples of models that exhibit highly nonlinear and non-monotonic response functions.

3.2.2 Global and local interpretability

It is essential to understand the behavior of the complete model on a global scale (the inputs and their entire modeled relations with a prediction target), but sometimes global interpretations can be very approximate. Therefore, it is crucial to derive local explanation by zooming into small regions of the data set and understand model predictions for a single row of data, or to a group of similar rows. Because small parts of ML-response functions are usually linear and monotonic, local explanations are more accurate (Hall & Gill, 2018).

Global interpretability – some ML interpretability techniques make possible only global explanation of ML-algorithms, their results, and the machine-learned relationship between prediction and input variables. Therefore, they can be labeled as "Global".

Local interpretability – facilitate understanding of small regions of the machine-learned relationship between predictions and inputs, such as clusters of inputs and their corresponding prediction, or very small parts of predictions and corresponding input rows, or even single row of data.

Sometime explainability techniques can be labeled as “Global” and “Local”, which means they can explain both, entire dataset relations and can provide granular views of local parts of the dataset.

3.2.3 Model-agnostic and Model-Specific Interpretability

We also can classify ML-interpretability techniques as model agnostic or model specific.

Model agnostic – means these techniques are universal and can be applied to various types of ML-algorithms. The problem is that these techniques often rely on approximations that can degrade the accuracy of the explanation they provide.

Model specific – related to the technics that are applicable only to the particular type or class of algorithm. These techniques tend to use the model for direct interpretation, which leads to potentially higher accuracy (Hall, et al., 2017).

3.3 Common interpretability techniques

In this section the most common interpretable techniques are listed without a detailed explanation, just to point out that there are quite some of them that can be used for proper model validation and explanation. All XAI techniques are open-source and can be used by all responsible developers. These techniques allow to open black boxes and increase the transparency of model and consequently increase user's trust.

3.3.1 Data visualization techniques for interpretability

Have a strong understanding of a dataset is the first step toward transparency. Therefore, data visualization is essential. The following techniques are chosen, because they can help illustrate many important aspects in just two dimensions (Hall, et al., 2017).

Technique 1	**2-D projections**
Description	Projecting dataset's rows from high-dimensional space into low dimensional, by using Principal

	Component Analysis (PCA), Multidimensional scaling (MDS), t-Distributed Stochastic Neighbor Embedding (t-SNE), Autoencoder Networks
Usage	A high-quality projection visualized in a scatter plot depicts key structural elements of a dataset (clusters, hierarchy, sparsity, outliers). 2-D projection is often used in fraud detection.
Scope	Global and Local. Advanced visualization toolkits allow users to pan, zoom, and drill-down easily.
Type	Model-agnostic. Visualizing complex datasets with many variables.
Complexity	Can help to understand very complex relationships in a dataset

Table 1.Visualisation Techniques. Source: (Hall & Gill, 2018)

Technique 2	**Correlation graphs**
Description	2-D representation of the relationships (correlation) in a dataset, where the nodes of the graph are variables, and the edge weights (thickness) between the nodes are defined by absolute values of their pairwise Pearson correlation. It allows us to see groups of correlated variables, identify irrelevant one, and discover the important relationship that ML model should incorporate
Usage	Very useful for text mining or topic modeling to see relations between entities and ideas. Also popular for finding relationships between customers or products in transactional data, and for fraud detection to find unusual interactions
Scope	Global and Local
Type	Model-agnostic
Complexity	Can help understand the complex relationship, but it is becoming difficult with more than several thousand variables

Table 2Visualisation Techniques. Source: (Hall & Gill, 2018)

3.3.2 Techniques for creating White-Box models

It is very important to identify first what the level of interpretability is required for success. If it is as of paramount importance, it is better to use interpretable modeling techniques described below from the very beginning (Hall & Gill, 2018).

Technique	**Decision trees**
Description	Can be explained with if-then rules
Scope	Global
Type	Model-specific
Complexity	For best interpretability, restrict to a shallow depth and binary splits; prediction can be restricted to be monotonic with respect to input variables
Technique	**explainable Neural Networks (XNN)**
Description	Straightforward calculation of derivatives of the trained ANN response function with regard to input variables made possible by the proliferation of deep learning toolkits such as TensorFlow, LRP[14]; these derivatives allow for the breakdown of the trained ANN response function into input variables contributions for any observation
Scope	Typically, Local, but can be both
Type	Model agnostic; used as surrogate models
Complexity	Used to directedly model extremely nonlinear, nonmonotonic phenomena; and used as surrogate models to explain other extremely complex nonlinear, nonmonotonic models
Technique	**Monotonic gradient-boosted machines (GBMs)**
Description	Can turn difficult-to -interpret nonlinear, nonmonotonic models into highly interpretable ones; it can be achieved with monotonicity constraints in GBMs by enforcing a uniform splitting strategy in component decision trees; appropriate for consistent reason code generation[15]
Scope	Global; monotonicity constraints create globally interpretable response function
Type	Model-specific, because implementations of monotonicity constraints vary for different types of models in practice
Complexity	Create nonlinear, monotonic response functions

[14] Layerwise Relevance Propagation (LRP) a technique for determining which features in a particular input vector contribute most strongly to a neural network's output (Shiebler, 2017)

[15] Consistent reasoning code generation is generally considered as a gold standard of model interpretability (Hall & Gill, 2018)

Technique	**Logistic, elastic net, GAM[16], and quantile regression**
Description	Use contemporary methods to augment traditional, linear modeling methods; these techniques are highly sophisticated; the inferential features and capabilities of linear models are rarely found in other classes; suit for regulated industries
Scope	Global
Type	Model-specific
Complexity	Alternative regression functions are generally linear and monotonic; GAM can create quite complex nonlinear functions
Technique	**Rule-based models**
Description	Composed of many simple Boolean statements, can be built by using expert knowledge or learning from real data
Scope	Global and Local
Type	Model-specific; can be highly interpretable if rules are restricted to simple combinations of input variable values
Complexity	have good explainability for users, because they obey Boolean logic (if-then), but also can model extremely complex nonlinear, nonmonotonic phenomena
Technique	**Supersparse linear integer models (SLIMs)**
Description	These predictive models require users to add, subtract, or to multiply values related to a handful of input variables to generate accurate predictions; perfect for serious situations, where interpretability and simplicity are critical, e.g. diagnosing newborn infant health against infant mortality using the well-known Apgar scale
Scope	Globally interpretable
Type	Model specific
Complexity	SLIMs are simple, linear models

Table 3. Techniques for creating White-box models. Source: (Hall & Gill, 2018)

[16] Generalized Additive Models

3.3.3 Interpretability techniques in complex ML models

There are different demands to various ML projects, sometimes accuracy is more important than transparency, but a certain level of explainability is always desirable. In some cases, unstructured or dirty input data can zero out the use of highly interpretable classical regression models. To handle such problems the following techniques must be used to extract explanations from complex, nonlinear, black box models. These techniques can also be implemented for white-box models to improve their interpretability further (Hall, et al., 2017).

Technique	**Decision-tree surrogates**
Description	Decision surrogate models are usually created by training a decision tree on the original inputs and predictions of a complex model. Variable importance, trends, and interactions displayed are supposed to be indicative of the internal mechanisms of the complex model
Scope	Generally Global, but can fit to more local regions
Type	Model agnostic
Complexity	Suits to models from medium to high complexity
Technique	**Partial-dependence plots (PDP)**
Description	PDP are small graphical images of the prediction function, which simplifies the understanding of the relationship between the results and the predictors they are interested in; pairs nicely with ICE, which can expose when PDP becomes inaccurate in the presence of strong interactions
Scope	Global in terms of the rows of a dataset, but local in terms of input variables
Type	Model agnostic
Complexity	Can describe nearly any functions, including nonlinear, nonmonotonic
Technique	**Individual conditional expectation (ICE) plots[17]**

[17] Plots – graphical means (bar charts, mosaic plots, box plots, histograms etc.) of communicating results in statistics and data analysis

Description	ICE is a newer adaptation of PDP, it depicts how a model behaves for a single row of data, and can be used to validate monotonicity constraints; some practitioners feel that ICE can be misleading in the presence of strong correlations between input variables
Scope	Local, because they apply to one observation at a time
Type	Model agnostic
Complexity	Can describe nearly any functions, including nonlinear, nonmonotonic
Technique	**Residual plots**
Description	Refers to the difference between the actual value of a target variable and the predicted value of it for every row in a dataset; it is a proven model assessment technique, good way to find outliers, and see all of your modeling results in 2D
Scope	Global and Local
Type	Model agnostic
Complexity	Can assess ML models of varying complexity

Table 4. Techniques for seeing model mechanisms with model visualizations. Source: (Hall & Gill, 2018)

3.3.4 Reason codes

Reason codes are text explanations of a model prediction with regard to input variables, it is also called turn-down codes. For example, creditors and insurance companies in the US according to regulations must provide grounds for refusal a credit application, or for insurance rate calculated for a particular person. They must explain why they reject applications, or why they charge a certain insurance cost, based on the values of input variables, such as credit rating and account balance, or age and health condition for their risk assessment models (Hall & Gill, 2018).

The importance of generating reason codes is fundamental for ML interpretability. Thanks to reasons code practitioners can understand whether the high weight was given to potentially bias

sensitive variables such as gender, age, marital status or ethnicity. To generate reason codes for simple linear models is not a big deal, but to do that for more accurate ML-models, consequentially more complex models, is a relatively new task.

Deriving reason codes in relation to input variables implies local scope, because we have to check the weights of particular attributes. Local surrogate models similarly to global surrogate models represent simple models of complex models. However, they are trained only for specific rows of data.

For example, Local Interpretable Model-Agnostic Explanations (LIME) is a designated method for designing local linear surrogate models for particular observations. Decision trees, as well as other rule-based models, also can be applied in local regions. All of these models can define how decisions on specific observations are made, and reason codes can be obtained by sorting the input variable contributions in these local models (Hall & Gill, 2018).

The characteristics of promising reason code generating techniques are describe in a Table 5 below.

Technique	**Anchors**
Description	Generate high-precision sets of plain-language rules to describe a machine learning model prediction in regard to model's input variables
Scope	Local
Type	Model agnostic
Complexity	Can create explanations for very complex functions, but the rule set needed to describe the prediction can become very large
Technique	**Leave-One-Covariate-Out (LOCO) variable importance**
Description	Creates local interpretation for each row, by scoring the row of data once and then again for

	each input variable (e.g. covariate[18]) in the row; in each additional scoring run, one input variable is set to missing, zero, mean, or another appropriate value to leave it out of the prediction; the variable with the largest absolute impact is the most important for that row's prediction
Scope	Typically, local, but can be global, by estimating the mean change in accuracy for each variable over an entire dataset
Type	Model agnostic
Complexity	More useful for nonlinear, nonmonotonic response functions, but can be applied for many others
Technique	**LIME**
Description	Uses local linear surrogate models to explain regions in a complex machine-learned response function around an observation of interest
Scope	Local
Type	Model agnostic
Complexity	Suited for functions of high complexity, but can fail in regions of extreme nonlinearity or high-degree interactions
Technique	**Treeinterpreter**
Description	Decomposes decision tree, random forest, and gradient-boosting machine predictions into bias and component terms for each variable used in a model; it simply outputs a list of bias and individual variable contributions globally and for each record
Scope	Global, when it represents average contributions of inputs to overall model predictions; local when used to explain a single prediction
Type	Model specific
Complexity	Explain nonlinear, nonmonotonic response functions created by a decision tree, random forest, and GBM algorithms
Technique	**Shapley explanations**
Description	Techniques with credible theoretical support that unifies approaches such as LIME, LOCO and

[18] Covariate -can be an independent, unwanted or confounding variable. By adding a covariate to a model, the accuracy of the results can be increased. (Statistics How To, 2019)

	Treeinterpreter for developing consistent local variable contributions to black-box model predictions
Scope	Local, but can be aggregated to create global explanations
Type	Can be both – model agnostic and model specific
Complexity	Applied to any ML models

Table 5. Reason code generation. Source: (Hall & Gill, 2018)

3.3.5 Fairness, Stability, and Trustworthiness

The importance of fairness and objective result generation was described earlier in this master thesis. Any techniques that can detect bias, can correct bias in model predictions, also can learn to make fair predictions, hence can be classified as fairness techniques (Hall & Gill, 2018). Consequently, LOCO, LIME, Treeinterpreter and Shapley (See Table 5) among others can also be considered as such techniques.

Forecasts of the ML model can change dramatically due to insignificant changes in input values. Therefore, it is important to check the model on stability. Sensitivity analysis examines whether the model's behavior and results remain stable when intentionally distorting data or other changes in data are simulated. In addition to traditional assessment methods, this is perhaps the most important validation method for ML models (Hall & Gill, 2018).

Given the approximate and probabilistic nature of ML or AI models, ML explanations may call into question the credibility of the interpretable models themselves, and rightfully so. However, according to Patrick Hall there is a technique to test interpretability and explanations for accuracy as well (Hall, et al., 2017).

It seems as infinite process to validate the model, then interpret the model, then validate the validation and interpretation models

and forever and ever, and still there will be a tiny piece of uncertainty. Thus, it is very important to think first whether it is smart solution to use AI-driven systems in some domains, and if yes then to what extent.

To conclude, XAI, FAT /ML, ML interpretability are rapidly changing and developing fields. Techniques and models described in this chapter represent only partially the variety of opportunities to check and validate models of various complexities and open black boxes. Given that, accuracy vs explainability it is not necessarily a trade-off, the combination of symbolic AI and ANN can improve understanding, therefore contribute in building user's trust.

4 Empirical part

In previous chapters of this master thesis, the main principles of ML- and AI-system's functionality were presented (chapter 2). Given the probabilistic, generalizing, heuristic, random, and statistical nature of many algorithms, the critical question of users' distrust is justifiable. The new trends of socially responsible and technically robust development of AI were presented, steps, and guidance towards explainable and ethical AI were introduced (chapter 3). Based on these recommendations, the subsequent case study delves into realities of development and application of AI-powered tools in tax and legal services. It seeks to firstly, examine the status quo of AI tools developed for and used in this domain. Secondly, validate the responsible and explainable approach in the development stage. Finally, assess users' trust problem and underlying reasons, as an obstacle toward ethical and legitimate AI adoption in this domain. Subsequently, the findings are presented and included in the recommendations to build users'

trust in AI as a prerequisite for successful deployment in daily operations.

4.1 Selection of Industry and respondents

AI and ML tools used for data analytics, fraud, and risk detection by banks and authorities challenge business to work flawlessly and comply with different legislation. As well as they challenge tax and legal advisors to work in constantly changing conditions, and audit, control and analyze ever-growing data.

Tax and legal services were selected out of a personal interest of the author and due to the high level of social and governmental importance and numerous legal and ethical issues related to the use of technology. Also, it was chosen due to the availability of contacts to managers and employees who deal with AI software development or exploit these tools for accounting, auditing, tax and legal services.

The respondents for this case study were selected to represent the variety of professional and business profiles within 2 main groups of stakeholders: users (legal and tax advisors); and AI/ML-tools developers focused on this domain. Among 9 respondents, 3 are currently in executive or top- management positions of AI software development companies (that develop products for legal compliance and due diligence, corporate finance and audit), 3 are senior-level experts in tax and legal services, and 3 are data scientists and tech initiatives experts. They were selected based on their diverse functional background, their expertise within data analysis and ML systems, and insights regarding AI systems utilization in this domain. In terms of functional expertise, 3 respondents have a background in legal and tax advisory, 2 has experience in AI and ML development, 1 works as a full stake

developer, 1 is a chief accounting officer, 1 is a chief financial officer, and 1 work as a tax crime counsel. In terms of work experience, most participants have at least 5 years of experience, the most senior experts, who represent users, have more than 15 years of relevant experience; AI-tools developers and data analytics have from 4 to 7 years of relevant experience.

All users represent Austrian legal and tax advisory firms. Developers are represented by 1 Austrian company, 1 based in Germany and 1 based in the UK.

4.2 Data collection

Semi-structured interviews with data scientists, AI developers, and industry experts were conducted in order to collect data for this case study. Semi-structured interviews were selected as a method of data collection because they allow respondents to elaborate on their views and experiences and add new facets to the problem under discussion. Moreover, the interviewer can ask for additional information in case of ambiguous answers. Each interview lasted about 35 to 60 minutes, depending on the depth of answers that respondents provided.

The questionnaire for the semi-structured interviews reflects the main requirements of trustworthy AI introduced by the EU Commission, applicable to different stakeholders partaking in AI-systems life cycle (developers and end-users). The questionnaire is divided into three main parts: (1) the reflection of the current state of the AI adoption, (2) the analysis of the current AI development process in regard to AI explainability and accuracy and (3) the reflection and validation of the user's trust problem. The following section details the key questions that guide the interviews:

Part 1: Analysis of the status quo of AI adoption

The goal of the first part of the interview is to gain insights into the current process of AI adoption in tax and legal services, from both users' and developers' point of view.

a) Are you aware for which purposes mainly AI-tools used in your industry / sector and how advanced they are?
b) Does your organization utilize any AI- or ML – powered tools? If yes, for which tasks? If not why?
c) Do you have a strategy/framework for AI implementation? If yes, is there a risk assessment procedure (assessment of the possible negative impact on brand/reputation, customers loyalty – in case of inaccurate, prejudiced and subjective decisions made as a result of AI tools exploitation). If not, why?

Part 2: Reflect and validate the robustness of the AI development process

The goal of the second part is to gain insights into whether and how AI-software development practices support the key aspects of ethical and responsible AI development. Also, to evaluate the end user's perception and understanding of the same aspects. The requirements for trustworthy AI developed by EU commission, or by the Institute for Ethical AI & ML were not presented to the interviewees, but the questions aim to check the level of understanding of the most critical aspects and their usage in practice.

a) Human agency and oversight. Do you consider the appropriate level of human control over the system while

developing tools; these tools are aimed only to enhance or augment human capabilities, and not to replace them?

b) Technical robustness and safety. Do you use any measure or systems to ensure the integrity and resilience of the AI system against malicious attacks? Do you put in place verification and validation of systems reliability and reproducibility?

c) Transparency. Do you assess and ensure traceability and explainability of your systems; by which means? Do you inform end-users on the mechanisms, reasons, and criteria behind the AI system?

d) Should AI development be regulated by governments or it should be up to organizations, but with full liability for that?

Part 3: Reflect and validate the user's trust problem

The goal of the third part of the interview is to gain insights into user's trust problem in tax and legal services. Particular focus is placed on the problem of trust in relation to responsibility and liability of stakeholders.

a) Who, in your opinion, should be responsible /liable for legal /tax malpractice caused by the use of algorithms

b) Who is usually liable for these errors now?

c) Do you understand how exactly AI-tools work or generate results? Do you have any suggestion how can practitioners improve the clarity and understanding of such tools?

d) How can you characterize your perception or altitude toward AI-systems in regard to jobs replacement?

Some questions in the survey refer to practical experience in AI development and some – to user's trust. It is expected that those

respondents, who represent practitioners with real experience in a specific sphere, will be able to provide more valuable insights. To assess the difference between experienced and non-experienced respondents' opinions, the subsequent analysis considers both sides.

4.3 Data Analysis and Pattern Recognition

All the interviews, which were used to collect data for the case study, were recorded and transcribed. The analysis and comparison of individual responses have been conducted to label similarities and identify patterns.

The findings were extracted by comparing the consolidated answers of users and developers to identify underlying reasons of users' distrust problem. Similarities and divergences are described in the findings.

Those findings that represent the divergences in opinion between developers and users, were identified to describe additional steps towards building user's trust in tax and legal services.

5 Research Results

The semi-structured interviews were conducted to verify the problem of reliability of AI systems and users' distrust from representatives of legal and auditing companies. What exactly drives distrust to AI systems and which steps must be taken by companies when introducing these technologies to integrate it smoothly into daily operations. The findings described in this section, validate the importance of building users' trust for beneficial and ethical AI adoption. This section describes the research results relating to (1) status quo of the AI adoption in

legal and auditing firms, (2) technical robustness and transparency in the development process and (3) users' trust.

5.1 Status quo of the AI adoption by legal and auditing firms

a) Are you aware of which purposes mainly AI-tools used in your industry and how advanced they are?

The majority of respondents have heard or know about different AI-driven tools used in the industry, especially for legal, due diligence, compliance, document generation, anomaly detection, controlling, e-investigation. However, only four out of nine understand how advanced such tools and what kind of help or assistance can be expected from them. Four respondents mention that tax authorities thanks to such technologies, can analyze all transactions between companies and can much faster identify mistakes or non-compliance. Therefore, they have to use such tools as well to be competitive and foresee problems. One respondent mentioned, that tax authorities use ANN to identify typical patterns and anomalies, and based on this information, choose entities for unscheduled tax inspections.

Finding 1: unrealistic perception of technology by users. Some respondents that represent applied practitioners believe that there some tools on the market that can read or understand text document; and that technology are much advancer than they are.

Finding 2: those who work closer to accounting and taxes (tax advisors, auditors), have an understanding that AI-tools can be very useful in error detection and controlling.

Finding 3: respondents who represent developers, data scientists have better awareness and deep understanding of the functionality of available products on the market

b) Does your organization utilize any AI- or ML – powered tools? If yes, for which tasks? If not why?

The majority of applied practitioners does not utilize such tools. However, two of the respondents mentioned that their company has due diligence tools and doc generators, but they do not use them. Overall, there is broad agreement among respondents that AI-tools are the future, and they can bring valuable competitive advantages to the companies. However, many respondents mentioned that in their domain, it is not so easy to use them and that lawyers are in general very conservative, and the only must-have tool that they use is MS Word.

Finding 4: company can acquire an AI-driven tool and does not integrate it into operations. As the reasons for that, the following issues were named:

1. the necessity to check first that the results achieved with the tool are the same as without (no time for that);
2. too difficult to use, demand some knowledge of logic or basic programmers' skills (regarding doc generator, based on decision tree model) – no time for that;
3. personal legal liability motivates to check all documents manually
4. if the mistake would be made because of the machine, professional insurance would not cover it, and the reputation of the firm would be destroyed;
5. the clause about the usage of such tools must be included in an engagement letter, otherwise, lawyers cannot use it.

6. Data security. Most tools are platforms solution, whenever we upload data there, the risk of leakages is growing.

c) Do you have a strategy/framework for AI implementation? If yes, is there a risk assessment procedure (assessment of the possible negative impact on brand/reputation, customers loyalty – in case of inaccurate, prejudiced and subjective decisions made as a result of AI tools exploitation). If not, why?

Some respondents from international firms believe that global AI strategy exists in their companies, but they are not aware of any strategies or guidelines about AI integration locally. Four respondents argue that the usage of AI tools is carrying an only experimental character, and they do not have such strategies so far; the field is quite new, and there is a lack of understanding. However, all agreed on the importance of such tools, because they can bring competitive advantages.

Finding 5: there are no precise instructions on AI integration in daily operations or risk assessment of AI exploitation in place, or employees are not aware of them.

Finding 6: some developers are ready to provide kind of guidelines on AI integration, but right now they do not have it. Regarding risk assessment, they provide the only tool. Therefore all risks depend on how users utilize the tool and the quality of data.

Finding 7: the other part of the developers admit that it depends mainly on management and people's attitude toward technology. Also, they mentioned that companies that use a top-down approach mostly succeed with integration.

5.2 Reflect and validate the robustness of the AI development process

a) Human agency and oversight. Do you consider the appropriate level of human control over the system while developing tools; these tools are aimed only to enhance or augment human capabilities, and not to replace them?

First, all respondents who represent developers and data scientists claimed that there is no autonomous tools or systems for this industry, that can work without human interaction. Second, the demands to robustness and accuracy is too high in this domain due to the legal liability of auditors, lawyers, and tax advisors. Therefore, they use only highly explainable models developed for humans. Overall, developers say that at this stage of the AI development, they are developing only tools to delegate simple iterative tasks to the machine.

Finding 7: currently, in developing tools for tax and legal services, the presence of applied professionals is not questioned

Finding 8: a few respondents from users' side mentioned, that they are using AI systems, they are teaching them, and with the rapid speed of technology development, it can finally affect their own or their colleagues' jobs.

b) Technical robustness and safety. Do you use any measure or systems to ensure the integrity and resilience of the AI system against malicious attacks? Do you put in place verification and validation of systems reliability and reproducibility?

Finding 9: developers are confident in methods and techniques they use; all their models are trained, tested, and validated on

different datasets; they have procedures on how to protect data in case of attacks. They regularly test the platforms on resilience to attacks, have security rules for data protection in case of attacks.

Findings 10: users cannot evaluate the robustness and security aspects, and can only rely on their IT departments.

c) Transparency. Do you assess and ensure traceability and explainability of your systems; by which means? Do you inform end-users on the mechanisms, reasons, and criteria behind the AI system?

In the service sector, it is essential to use only transparent or explainable and validated algorithms, because no matter which tools or methods were used, practitioners have personal responsibility for the results of their work.

Finding 11: developers use the explainable and rule-based models for this domain; usually, they use supervised learning algorithms. For anomaly detection and pattern recognition can be used ANNs, but they are not responsible for decisions and just indicate the anomaly or similarities, which must be checked by practitioners; moreover, anomalies do not necessarily mean mistakes or errors, they show only that out of 1000 contracts /transactions with the same label, 100 have some discrepancies. Also, their algorithms domain specific and trained only for a very narrow range of tasks.

Finding 12: all respondents who represent developers argued that they are providing the information about systems and word explanation of functionality, as well as user guidelines, and believe

that it should be sufficient for the user to understand and trust AI systems.

Finding 13: only one developer marks their product as XAI and used rules visualization models. However, all they use visualizations for anomalies representations.

Findings 14: users' side partly thinks that better description can help to improve understanding.

d) Should AI development be better regulated by governments or it should be up to organizations, but with full liability for the consequences?

Finding 15: more regulations do little to help consumers or users, but they will slow down the development and use of AI in Europe by holding developers to a standard that is often unnecessary and infeasible.

Finding 16: there are enough laws and regulations; it is more about contracts between suppliers and customers, and agreements between them.

Concerning the usage of AI systems for professional services, it is better to evaluate risks, by involving legal, IT and data specialists, before buying any products.

5.3 Reflect and validate the user's trust problem

a) Who, in your opinion, should be responsible /liable for legal /tax malpractice caused by the use of algorithms: developers; companies that utilize these algorithms; individuals who designed the algorithm?

Finding 17: all users are convinced that developers must be responsible for results, data protection, and security issues.

Finding 18: all developers argue that the products they offer now are aimed only to improve practitioner's productivity by enhancing their ability to search, compare, or identify anomalies faster than they would do it manually, and accuracy in such tasks is 100%. They cannot be liable for the quality of users' work, because their software does not generate any results or decisions.

b) Who is liable now for these errors?

In all standard contracts, developers have a clause, that they are not liable for any processes in which their products were involved, including doc generation, and certified practitioners must validate all results. Legal counsels, auditors, and tax advisors, consequently leave it up to clients, whether they want or not to use such a tool, even if they can reduce the time significantly.

Finding 19: developers transfer the liability to users, applied practitioners put it on clients.

c) Could you name the main obstacles toward AI adoption in your domain?

Respondents proposed a wide range of obstacles that could prevent usage of AI systems in legal and tax services firms, some of them were mentioned more than once or related to the same type of problems, they were grouped in corresponding findings, others represent the unique opinion, but reflect the deep understanding of the problem.

Finding 20: in legal and tax services sector, the complexity and difficulty of the work measured by billing-ours, which professionals spend working on the client's problem. AI-tools aim to increase productivity and reduce time spent on a particular task – it means fewer billing-hours and less revenue.

Finding 21: all applied practitioners are concerned about the liability issue.

Finding 22: some applied practitioners see a threat to their profession.

Finding 23: developers noticed that computer literacy and the lack of savvy PC users play a role as well.

Finding 24: the technology is not good enough yet, and do not provide tremendous improvement

Finding 25: the cost of a junior specialist in comparison to the investments in AI-tools is much cheaper; it does not demand time on integration and staff training.

Finding 26: technology is entirely new; it is too risky to delegate any legal tasks (trust).

Finding 27: respondents who work more with taxes and auditing see fewer obstacles for implementation, but they would not use any tools for the legal part of work as well, because they do not trust AI with word processing.

Finding 28: too many tools on the market, challenging to find the right one; it would be better to have one system and use it for the whole firm.

d) How can you characterize your perception or altitude toward AI-systems regarding jobs replacement?

The respondent's opinions are quite the opposite; however, the majority believe that there is a high risk of replacement, especially for junior specialists.

Finding 29: developers tend to think that, there is no thread from the tools that they develop now; they can only enhance practitioner's efficiency.

Finding 30: users believe that these systems enhance the productivity of one practitioner at the expense of others, who will not get this job.

Finding 31: technology will reshape the way how services are provided. However, there is no threat to the profession fundamentally, because all new technologies usually create new legal issues and cases.

One of the respondents gave a good example from the past regarding reshaping the traditional way of work in legal firms. Thirty years ago, all big law firms had huge departments of secretaries, now they have 3 or 4 team assistants for the whole firm, but it does not seem that the number of lawyers or advisors were reduced. Nowadays, lawyers do this job by themselves on their PCs. Therefore, it is more likely that applied practitioners will have to acquire new technical and computer skills.

5.4 Interpretation and recommendations

To build users' trust in companies that provide tax and legal services is the task of high complexity. First of all, because of the traditionally conservative culture in this domain. Second, lawyers and tax advisors are generally rationalists, but because of the professional need to prove all their logical inferences, they need empirical evidence to support all their ideas. Consequently, their demands and standards regarding reliability of AI-tool are high.

However, since all respondents are agreed that AI tools can contribute to efficiency and create competitive advantages in their

domain, sooner or letter, all companies will have to deploy these tools. Therefore, building trust in AI should be one of their strategic steps.

Based on findings listed above the following list of recommendation can help build user's trust:

1. Build AI-friendly culture in a company first, and then invest in AI tools (findings 1,4,7).

In order to build AI-friendly culture, AI integration guidance must be included in tech initiative as a part of business strategy. Management have to communicate the strategy, the importance of AI integration and steps toward achieving these goals to employees.

2. Include AI, logic and data analysis in the knowledge management program.

Include these topics in knowledge management programs will allow users to understand the underlying mechanisms of AI systems. These topics must be restricted to domain-specific examples only.

Provide a helicopter about the usage of such tools in their domain globally and increase awareness about the subject, advantages, and drawbacks of using such tools.

3. Before integrating such systems in operations, companies have to identify financial and business objectives.

It is essential for users to understand that a company will benefit from technology. To achieve that, companies have to define whether they invest in technology to acquire competitive advantages and deliver their services much faster than competitors approximately for the same price, and will in a scale

(bigger market share, number of clients); or they want to sell the usage of AI-systems to their clients.

The last model raises the biggest concern because in this case, employees perceive AI systems as their competitor, who takes away their billing-hours. With this perception to build AI-friendly culture would be impossible.

4. Developers have to communicate the capabilities of the systems better.

The Liability problem is the primary users' concern in this domain, mainly because they are far from AI, ML, and computer science in general, and do not understand reality. Probably, because developers or their marketers present the achievements of their systems as outstanding or extraordinary, and users think that machine is able already to read, understand the content.

Developers should also demystify AI from their side, especially in such conservative domains.

5. Characteristics such as explainability and interpretability of AI-system must be communicated better.

Findings 11,13 show that AI systems for this domain are very robust. However, users do not understand it. Consequently, developers should communicate it better or clearly to users.

6 Conclusion

Technological advancements and rapid evolutions in business models are disrupting traditional modes of operations, processes, and global value chains and pushing the existing legal and taxation boundaries. AI, ML, and data analytics increase transparency of transactions between legal and private entities; open up

opportunities to transform the way how authorities operate, control, and interact with taxpayers. Therefore, the importance to integrate AI in a daily operation is critical in all business spheres.

Digital payments are growing in scale and significance – data are becoming a precious source for transparency and traceability (Barnay, et al., 2018). These and other changes are raising concerns about internal and international compliance of all business operations and challenging the conventional role of authorities and their agents: legal counsels, auditors, tax advisers.

This research contributes to a better understanding of AI reliability and accuracy problem and its correlation to users' trust in general. As well as provide insights into AI integration problems in tax and legal services firms.

The empirical case study identified the massive gap between suppliers and AI-tools users in tax and legal services area. Developers should communicate better XAI and FAT approaches used in the development, to ensure users in the reliability of the products. AI and ML products labeled as XAI or FAT, must be recognized by users as label BIO in the grocery store, then users will look for these labels, and developers will have to correspond to transparency and explainability standards.

Fear of job deprivation also one of the main reasons to distrust technology. However, technology will likely open up a new rage of new jobs.

7 Bibliography

AI HLEG, 2019. *Ethics Guidelines for Trustworthy AI.* [Online]
Available at: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1#Robustness
[Accessed 15 June 2019].

Alpaydin, E., 2016. *Machine Learnig: The new AI.* Cambridge(MA): MIT Press.

Auerbach, D., 2015. *The programms that become a programmers.* [Online]
Available at: https://slate.com/technology/2015/09/pedro-domingos-master-algorithm-how-machine-learning-is-reshaping-how-we-live.html
[Accessed 10 May 2019].

Barnay, A. et al., 2018. *Four innovations reshaping tax administration.* [Online]
Available at: https://www.mckinsey.com/industries/public-sector/our-insights/four-innovations-reshaping-tax-administration
[Accessed 05 June 2019].

Bostrom, N., 2014. *Superintelligence: Paths, Dangers, Strategies.* Reprinted with correction 2017 ed. Oxford: Oxford University Press.

Buranyi, S., 2017. *Rise of the racist robots - how AI is learning all our worst impulses.* [Online]
Available at: https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses
[Accessed 3 April 2019].

Cherry, K., 2019. *Heuristic and cognitive Biases.* [Online]
Available at: https://www.verywellmind.com/what-is-a-heuristic-2795235
[Accessed 23 May 2019].

Diepen, G., tot Everlo, T. S. & el Bouazzaoui, H., 2017. *Part 2. Attificial intelligence techniques explaind.* [Online]
Available at: https://www2.deloitte.com/nl/nl/pages/data-analytics/articles/part-2-artificial-intelligence-techniques-explained.html
[Accessed 28 may 2109].

Domingos, P., 2015. *The Master Algorithm.* London: Allen Lane.

Dorigo, M., 2007. *Ant colony optimization.* [Online]
Available at: http://www.scholarpedia.org/article/Ant_colony_optimization
[Accessed 29 May 2019].

Doshi-Velez, F. & Kim, B., 2017. *Towards a rigorous science of interpretable maccine learning.* [Online]
Available at: https://arxiv.org/abs/1702.08608
[Accessed 10 June 2019].

Ertel, W., 2008. *Grundkurs Künstliche Intelligenz.* Wiesbaden: Friedr. Vieweg & Sohn Verlag.

FAT, 2019. *ACM FAT conference.* [Online]
Available at: https://fatconference.org/index.html
[Accessed 5 June 2019].

Fox, E., 2019. *Machine Learning: Clustering & Retrival.* [Online]
Available at: https://www.coursera.org/lecture/ml-clustering-and-retrieval/complexity-of-brute-force-search-5R6q3
[Accessed 30 May 2019].

Goebel, R. et al., 2018. *Explainable AI: the new 42?.* Hamburg, Germany, hal-01934928, pp. 295-303.

Gottfredson, L. S., 1997. Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence,* 24(1), pp. 13-23.

Gunning, D., 2018. *Explainable Artificial Intelligence (XAI).* [Online]
Available at: https://www.darpa.mil/program/explainable-artificial-intelligence
[Accessed 10 June 2019].

Hall, P. & Gill, N., 2018. *An introduction to machine learning interpretability : an applied perspective on fairness, accountability, transparency, and explainable AI.* 1st ed. Sebastopol, CA: O'Reilly Media.

Hall, P., Phan, W. & Ambati, S., 2017. *Ideas on interpreting machine learning.* [Online]
Available at: https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning
[Accessed 20 June 2019].

Hassabis, D. & Silver, D., 2017. *AlfaGo Zero: Learning from scratch.* [Online]
Available at: http://deepmind.com/blog/alphago-zero-learning-scratch/
[Accessed 3 April 2019].

IBM Juornal of Research and Development, 2012. *This is Watson, Volume 56 Issue 3.4.* [Online]
Available at: https://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=6177717
[Accessed May 2019].

Jay, P., 2008. *Curtains for Netscape.* [Online]
Available at: https://www.cbc.ca/technology/technology-blog/2008/02/curtains-for-netscape.html
[Accessed 2 June 2019].

Jeckson, P., 1998. *Introduction to Expert Systems.* 3 ed. Boston: Addison-Wesley.

Larose, D. T. & Larose, C. D., 2015. *Data Mining and Predictive Analytics.* 2nd ed. Hoboken, New Jersey: John Wiley & Sons.

Leibniz, G. W., 1887. *Philosophische Schriften.* Berlin: s.n.

Luger, G. F., 2009. *Artificial Intelligence.* 6 ed. Boston: Pearson Education.

McCarthy, J., 2007. From here to human level AI. *Artificial Intelligence,* 171(18), p. 1174.

McCorduck, P., 2004. *Machines who think.* 2 ed. Natick, MA: A.K.Peters Ltd..

Mithchel-Guthrie, P., 2014. *The SAS Data science blog.* [Online]
Available at: https://blogs.sas.com/content/subconsciousmusings/2014/08/22/looking-backwards-looking-forwards-sas-data-mining-and-machine-learning/
[Accessed 25 March 2019].

Morrison, A. & Rao, A., 2016. *Machine Learnig overview.* [Online]
Available at: http://usblogs.pwc.com/emerging-technology/a-look-at-machine-learning-infographic/
[Accessed 20 March 2019].

Morrison, A. & Rao, A., 2017. *Machine learning methods.* [Online]
Available at: http://usblogs.pwc.com/emerging-technology/machine-learning-

methods-infographic/
[Accessed 20 May 2019].

Pierce, R., 2019. *Mathisfun.* [Online]
Available at: https://www.mathsisfun.com/data/probability-false-negatives-positives.html
[Accessed 30 May 2019].

Ransbotham, S., Kiron , D., Gerbert, P. & Reeves, M., 2017. *the 2017 Artificial Intelligence Global Executive Study and Research Project.* [Online]
Available at: https://sloanreview.mit.edu/projects/reshaping-business-with-artificial-intelligence/
[Accessed 30 April 2019].

Rao, A. & Golbin, I., 2018. *Next in Tech: PwC US blog.* [Online]
Available at: http://usblogs.pwc.com/emerging-technology/to-open-ai-black-box/
[Accessed 5 May 2019].

Rich, E., 1983. *Artificial Intelligence.* s.l.:McGraw-Hill.

Robbins, S., 2019. *AI and the path to envelopment: knowledge as the first step towards the responsible regulation and use od AI-powered machines.* [Online]
Available at: https://link.springer.com/article/10.1007/s00146-019-00891-1
[Accessed 30 May 2019].

Russel, S. J. & Norvig, P., 2003. *Artificial Intelligence: A Modern Approach.* 2 ed. Upper Seddle River, New Jersey: Prentice Hall.

Shiebler, D., 2017. *Understanding of Neral Networks with Layerwise Relavance Propagation.* [Online]
Available at: http://danshiebler.com/2017-04-16-deep-taylor-lrp/
[Accessed 10 June 2019].

Statistics How To, 2019. *Covariate Definition in Statistics.* [Online]
Available at: https://www.statisticshowto.datasciencecentral.com/covariate/
[Accessed 20 June 2019].

Stover, C., 2019. *"Monotonic Function".* [Online]
Available at: http://mathworld.wolfram.com/MonotonicFunction.html
[Accessed 20 June 2019].

The Institute for Ethical AI & ML, 2019. *The Institute for Ethical AI & ML.* [Online]
Available at: https://ethical.institute/index.html
[Accessed 10 June 2019].