Bias – A lurking danger that can convert algorithmic systems into discriminatory entities

A framework for bias identification and mitigation

Bachelor’s thesis

Degree Programme in Business Information Technology
Visamäki – Hämeenlinna University Center – Finland
Spring 2019

Thea Gasser
ABSTRACT

This thesis examines the existence of bias in algorithmic systems and presents them as the cause for unfair and discriminatory decisions that are made through the use of such systems. Extensive literature research was conducted to review current discussions on this issue. Content and thematic analysis was applied to over 100 journal articles, books and websites to bring together proposals on how bias can be identified and reduced. The results of the analysis were further developed to provide precise measures for project teams building Artificial Intelligence systems.

The findings demonstrate that humankind aims to map Human Intelligence to Artificial Intelligence, but no system has reached such intelligence yet due to the lack of machine sentience and self-awareness. Therefore, the human being maintains considerable influence on the design of the system. Cognitive bias is very likely to be reflected in algorithmic systems. Awareness of the topic of bias must increasingly be addressed in project teams and appropriate measures must be applied. It will also be illustrated that standardization work is in progress and that the areas of AI-responsibility, AI-safety and AI-fairness will have high priority in the future. The importance of the topic of bias in algorithmic systems has been recognized by researchers and managers and the demand for fair AI-systems is high. The outcome of this thesis is a framework that contributes to AI-safety. This framework could be considered as a guideline and proposes measures for identifying and mitigating bias in algorithmic systems. It can be adapted and extended to the specific project context. Future collaborations and regulations among business, institutions and society are required to successfully address this issue.

Keywords  algorithm, artificial intelligence, algorithmic system, bias, human intelligence

Pages  41 pages including appendices 46 pages
1 INTRODUCTION

Artificial Intelligence (AI) is present in almost every area of our society. Be it in medicine, finance, social media, education, human resource management and many more. The future will continue to reinforce this trend and take up a deeper part of people’s lives. The survey of Accenture (2017, p. 5) reveals, around 85% of the executives surveyed plan to invest widely in AI-related technologies over the next three years.

According to Accenture (2017, pp. 12, 15), AI will play a central role in how customers perceive a company and define to a large extent how interactions with their employees and customers take place. AI will become a core competency and reflects a large part of a company’s character. In five years more than 50% of the customers will no longer choose a service based on the brand but will focus on how much AI is offered for that particular service.

However, recently there has been growing concern surrounding unfair decisions made through the use of algorithmic systems that have led to discrimination against social groups or individuals (Koene, 2017, p. 31; Veale & Binns, 2017, p. 1). Methods for measuring algorithms, recognise and mitigate bias and provide fair AI-software are demanded (Hardt & Price, n.d., p. 1; see also Veale & Binns, 2017, pp. 1-2).

The goal of this thesis is to contribute to AI-safety by highlighting that bias in the area of AI is very likely, illustrating possible sources of bias and proposing a framework which supports the identification and mitigation of bias during the development, implementation and application phases of AI-systems. The following four research questions are answered throughout this thesis to achieve the goal mentioned: (1) What is expected of AI-systems in relation to how humans make decisions? (2) How is bias that affects human behaviour and decisions also present in algorithmic systems? (3) How can bias in algorithmic systems be identified? (4) What measurements can be taken to mitigate bias in algorithmic systems? The structure of this thesis which is applied to answer the research questions is described in the next chapter.
2 METHODOLOGY

This thesis consists of a research section and a results section. The methods used for the individual sections are described below.

2.1 Research

The research content of this thesis is based on literature collection and analysis. The literature collection, review and analysis was conducted according to selected methods illustrated in “Wissenschaftliches schreiben leicht gemacht für Bachelor, Master und Dissertation” (Kornmeier, 2018). Systematic analysis was applied by researching specific AI and bias related topics and content and thus identifying central sources. Based on their bibliography, further literature was identified with a backward search strategy. Over 100 journal articles, collected works, reference works, books and websites were researched.

The author of this thesis first came up with own thoughts about the topic of AI after reading several up to date articles about the relevance of security in AI. Plain web and database searches were used as a starting point. Sciencedirect.com, SAGE journals, SpringerLink (link.springer.com), Google Scholar and Journal storage (jstor.org) have been consulted. Among others, search terms like “expectations towards AI”, “human intelligence”, “algorithmic bias”, “bias in software development”, “mitigating algorithmic bias” have been applied. The author researched literature about what AI means and how it is used today. The question of whether a connection between Human Intelligence (HI) and AI existed had to be answered first. Since the comparison of HI and AI demonstrated that both influence human behaviour, the author considered it as relevant to examine the characteristics of this impact starting with HI. As people tend to strive for power, it can result in either good or bad influence. This in turn can arise intentionally or unintentionally which leads to the term of cognitive bias. HI is mapped to AI and therefore cognitive bias and other types of bias are mapped too. The author focused on the existence of bias in algorithmic systems and would like to suggest ways in how to identify and mitigate them.

2.2 Results

The findings of the research of this thesis were taken and developed further (own ideas, experiences, brainstorming) in the results section. From this, a framework which documents how to prevent bias from entering an algorithmic system as well as identify and mitigate existing bias during the development, implementation and application phases of a system was created.
2.3 Target group and benefit of the research

The target groups are project teams which create algorithmic systems and end users who make decisions based on these systems’ outputs. The benefit of this thesis is that project teams and end users are aware of possible bias that can enter algorithmic systems and have concrete suggestions at their disposal to prevent this bias from entering the system or how to identify and mitigate existing bias.

2.4 Scope

The topic of AI has multiple aspects such as technical, ethical, psychological, legal or business. In the scope of this thesis, ethical and psychological aspects together with business aspects are investigated. Technical aspects are mentioned to support ethical, psychological and business aspects but do not represent central points. The research of this thesis is not limited to a specific industry area and it is possible to apply (part of) the results as general findings in each industry as a starting point for the prevention, identification and mitigation of bias in algorithmic systems.

The author of this thesis anticipates the following definition of algorithmic systems as appropriate for the context of this thesis: “Algorithmic systems [...] refers to the combination of algorithms, data and the output deployment process that together determine the outcomes that affect end users” (Koene, Dowthwaite, & Seth, 2018, p. 39).

2.5 Relevance of the topic and initial situation

Manufacturers of AI-systems do not disclose their implemented algorithms due to various reasons. One of them being a possible loss of competitive advantage over competitors. Another reason is that by exposing the algorithms, an opportunity would be offered to malicious attackers. (Burrell, 2016, p. 3) The reader might ask whether these reasons are tenable and justified, however, this is not examined in this thesis. Nevertheless, the author would like to mention that concerns which claim that maybe this opacity is abused to escape regulation have been addressed (Pasquale, 2015, p. 2).

Without access to the code most of the software remains a so-called black box where the user can indeed still record which data acts as input and what the result (output) is. Nonetheless, the transformation from the first to the latter one remains a hidden process for the user. (Sentient Technologies Holdings Limited, 2018) However, this process is significant for a company to understand why the output is as it is since the black box proposes supposedly appropriate future business-decision-suggestions to management. In the long run, unquestioned and incomprehensible suggestions put into action can cause enormous damage not only to the
business itself but may also lead to discrimination of groups and individuals (Pasquale, 2015, p. 17). Because the algorithms of such black box systems are not disclosed, it is crucial to draw the attention of users that are authorized to make business-decisions based on the recommendations of such black boxes to the fact that obscure algorithms can contain distortion.

Examples of discriminating algorithmic systems that have recently attracted public attention include:

COMPAS, an algorithm in 2016 accused of predicting that “...black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk” (Larson, Mattu, Kirchner, & Angwin, 2016). Further in 2016, the algorithm PredPol targeted a criminal minority unfairly by leading the police to a particular neighbourhood when in fact the algorithm was contemplated to be implemented to reduce human bias in policing in the first place (Cossins, 2018). In addition, a case where race and gender bias were detected in 2018 for an algorithm that “...correctly identified the gender of white men 99 per cent of the time. However, the error rate rose for people with darker skin, reaching nearly 35 per cent for women” (Revell, 2018). In 2015, Google’s image search had been accused of bias indicating fewer women than actually true when searching for the term CEO. Additionally, Google’s advertising system displayed high-income jobs much less to women than to men (Cossins, 2018). Further, in 2015 a software engineer stated that his black friends are classified as gorillas by Google’s image recognition algorithm (Vincent, 2018). Another case concerns Facebook where in 2017 its automatic translation algorithm chose a wrong translation for a user post which resulted in the Israeli police interrogating the affected user before the error was discovered (Cossins, 2018). Moreover, different cases are known where self-driving cars failed to detect pedestrians or vehicles with the fatal consequences of people being killed (Levin & Wong, 2018).

These are just a few but shocking examples that demonstrate the presence of bias in very diverse ways within algorithmic systems and identify the need for discussion and action. In fact, bias might be lurking in algorithmic systems and only become visible when algorithms create new algorithms until the complexity is so high that it becomes increasingly difficult even for the creators of such AI-tools to entirely understand them (Sentient Technologies Holdings Limited, 2018; see also C. Smith, McGuire, Huang Ting, & Yang, 2006, p. 21). It is therefore of great importance to provide suggestions on how to identify and mitigate bias before discriminatory decisions are made.
3 FROM HUMAN INTELLIGENCE TO ARTIFICIAL INTELLIGENCE

This chapter introduces the area of AI through illustrating what intelligence means in the context of this discipline and comparing HI with AI. Along with this, expectations towards the field of AI and what has been achieved to this day are discussed.

3.1 Artificial Intelligence

In 1956 at the Dartmouth College Artificial Intelligence Conference, John McCarthy and a group of scientists brought to life the term Artificial Intelligence by using it within their proposal, therefore establishing it as a research discipline (Moor, 2006, p. 87). Today AI is considered a machine or computer system that is capable of simulating HI processes such as learning from past experience, reasoning (ending up at rough or exact conclusions by applying rules), discovering meaning, generalizing and improving itself by being autodidactic, meaning by correcting itself (Burns & Laskowski, 2018).

The Turing Archive for the History of Computing, defines AI as “…the science of making computers do things that require intelligence when done by humans” (Copeland, 2000). One of the first attempts in trying to model HI through AI was made by the computer pioneer and AI-theorist Alan Turing in 1950 where he asked the question ‘if a computer can think?’ (Turing, 1950, pp. 433-460).

3.1.1 The Alan Turing Test

Alan Turing designed the so-called Turing Test also referred to as the Imitation Game. This test was designed to investigate if a machine can think on its own and with the idea of representing a model for measuring intelligence. (C. Smith et al., 2006, pp. 4-5) The Turing test would be conducted in the following way: A computer (A) and a human (B) are put on one side and a human judge (C) on the other side. A conversation between the two sides is going on, at the end of which the human judge had to tell who the real person was and who was the computer. The machine would have passed the test if the human judge was not able to tell which entity was the human (B) and which the machine (A). (Turing, 1950, pp. 433-460)

However, it needs to be mentioned that among other critics Alan Turing himself wrote several objections towards his claims that a machine that passed the Turing Test would be considered intelligent and able to think. In his Argument of Consciousness, he states that it is not enough to just mirror a human as this would never cover the human being in its whole variety. For example, it cannot be figured out what or if a machine feels. A
machine trying to convince a human that they have feelings would reject his Argument of Consciousness because humans themselves can never know for sure what other human beings feel – if they do at all. Nevertheless, several other factors question the Turing Test’s veracity to consider a machine as human-like intelligent when passing the test. (Turing, 1950, pp. 433-460) C. Smith et al. (2006, p. 7) further point out that the Turing Test takes place under a determined situation which does not reflect other circumstances a human might be confronted with in everyday life. They also exemplify the view that a computer could still be able to think but not pass the Turing Test. As an analogy, they draw the example that young children who lack the necessary amount of experience and knowledge to pass the Turing Test would still be able to think.

Even today, the test is still being brought up and talked about whether the newest invention has passed it or not. Nevertheless, there are many critics objecting various arguments that lead to the question of what the Turing Test is actually measuring in the end. The test is accused of being more about tricking or as Pat Hayes puts it: representing a test of being a victorious liar than a real measurement for human-like intelligence. (Falk, 2014)

Ongoing discussions about the validity of the Turing Test reveal the disagreement and diverging views about if and how a system can genuinely be considered as intelligent. The Turing Test was established more than 60 years ago which is remarkable considering that Alan Turing was already at that time about the six AI-disciplines (Natural Language Processing, Knowledge Representation, Automated Reasoning, Machine Learning, Computer Vision and Robotics) which nowadays make up the majority of the AI-area. (Russell & Norvig, 2010, pp. 2-3) Therefore, it makes sense to have a glance at the expectations and perceptions towards AI that have built up over the years till this day.

3.1.2 Expectations towards AI

Marvin Minsky, a Massachusetts Institute of Technology (MIT) professor and researcher in the field of Computer Sciences said in 1970 in an interview for Life magazine “… from three to eight years we will have a machine with the general intelligence of an average human being. I mean a machine that will be able to read Shakespeare, grease a car, play office politics, tell a joke and have a fight.” Furthermore, he mentioned that the machine would be able to educate itself and learn to reach the level of a genius in a few months and will have achieved an inestimable dominance a few months later. (Computers will be playing office politics, 1970, p. 58D)

Nowadays, the expectations towards AI that Minsky had expressed are still maintained. Even though these expectations have not been met until today, AI is an area that extensively influences and advances many other fields of research. For example, Robotics and Multimedia Indexing, which developed out of AI research are now used in a variety of other research
areas helping to speed up the progress of these areas in turn. (Orio, n.d.) AI has become a universal discipline since it seems to be appropriate for any possible intellectual task and can therefore be found in every area of people’s lives (Russell & Norvig, 2010, p. 1).

From robotic vehicles or driverless cars, chess playing supercomputers beating world chess champion Garry Kasparov, speech and face recognition, autonomous planning and scheduling, game playing, virtual customer assistant, spam fighting, machine translation, finding patterns in very large combined data sets to suggest or even make organizational business decisions to deciding whom to offer a loan to and predicting future events (Russell & Norvig, 2010, pp. 28-29). AI systems are embedded in a vast amount of infrastructures of many sectors (Kurzweil, 2010, p. 264) and emerging AI technologies are expected to “…blur the lines between human and machines” (Gartner, 2018) in the next five to ten years.

AI is being promoted because the opportunity to make people’s life simpler is recognised. For example, instead of entering a password, the face is recognised to log on to the laptop. Or, in online shopping, where customers are automatically offered products they might like without having to search for them themselves actively (Kelleher, 2017).

3.1.3 AI at the core

The scope of AI is disputed. Scientists, authors, institutions and organisations like John McCarthy, Marvin Minsky, Demis Hassabis, Jim Sterne, IBM, Accenture, McKinsey, Salesforce, US Government have different views (Marsden, 2017). When reading about AI, terms like imitation, simulation or mimicking are repeatedly applied which by their definition imply doing something like something/someone. In this case acting, learning and reasoning like humans. (Holder, 2019) Therefore, if today’s AI-behaviour such as Apple’s Siri is considered, it could be claimed that the voice assistant is not intelligent by definition. Looking into details, Apple’s voice assistant is based on evaluated data and facts permitting to offer an appropriate answer (Goel, 2018). An independently thinking and reasoning machine is not yet present since, amongst other things, an input is still needed. Even though AI acquires intelligence and learns through an autonomous process it lacks sentience and self-awareness and is still only a simulation of HI and nothing more (Holder, 2019).

3.1.4 Types of AI

AI can be divided into narrow AI (also called weak AI) and general AI (also called strong AI). The first type stands for machines that are specialised in a narrowly defined area. They seem intelligent at first glance because they can beat a chess player, for example. However, in this type every possible next step was programmed manually beforehand. The latter type deals
with machines that are intelligent on their own and would act similar to a human being. Even though there are no machines that reach this category yet, scientists and researchers are striving to achieve this type of AI. (Kumar, 2018)

In 2019, there are some cases in which it might seem humanity has reached some HI with a machine however it is simply because the system has been trained for a specific task and situation but not for an overall problem approach. To sum up, AI today is still narrow and very focused. Supercomputers beating humans in game playing do so only because they are very fast in trying every possible move out and deciding the best course of action in a fraction of time which would be impossible for a human being (from a time point of view). However, these systems are still rule-based programmed without indicating intelligence. (C. Smith et al., 2006, pp. 4, 10) Since AI-systems are still in the narrow range at the moment, and computers lack consciousness and therefore do not represent intelligence in itself, the question arises what do all the AI-systems consist of and what makes them what they are as well as how did they get there.

3.1.5 Learning process of AI

When talking about AI, different branches form part of this discipline. One is Machine Learning which deals with the system learning autonomously out of large data sets and experience to come up with the optimal algorithm for a particular situation (Alpaydin, 2010, p. 3). According to Russel & Norving (2010, p. 2), Machine Learning systems can adapt themselves in changing environments. However, how are these learning processes applied? AI has its intelligent agents.

As Russel & Norving (2010, p. 34) describe: “An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators.” To illustrate this concept, eyes or ears would be examples of sensors in a human agent and hands, legs or vocal tracts actuators (Russell & Norvig, 2010, pp. 34-35). An intelligent agent would then be a computer system, that can act or operate independently in a specific or given environment to achieve the desired goal(s) by executing the best move in given circumstances. The intelligent agent in a software system would receive file content and/or keystrokes as sensory inputs and act in the environment by displaying content on the screen. (Russell & Norvig, 2010, pp. 30-35)
For this kind of methodology, it is often unknown how the underlying process to reach the defined goal actually works. For example, people cannot explain the procedure of recognising a person’s face regardless of the current haircut, if the person wears glasses or the pose. Therefore, one hopes to recognise patterns from vast amounts of data that illustrate this process and establish rules for it. People carry out an enormous number of such unconscious processes without being able to explain precisely how they do it. Transferring these processes to machines is not an easy task and a possible source of bias. (Alpaydin, 2010, p. 3) With this technology it is left to the algorithms to carry out analysis from tremendous amounts of data and to replace or widely support people in their decision-making. The algorithms change parameters and decision rules as they see fit (Burrell, 2016, p. 5). This means considerable uncertainty and opacity for the comprehensibility of this process and is hardly transparent for humans (Tutt, 2016, p. 102).

From the illustrated aspects it can be seen that software reacts and learns differently compared to a human being. A machine is dependent on inputs and these can only come from humans. The methodology processing these inputs are algorithms.

3.2 Algorithm

AI is based on pieces of software and hardware such as any other computer program. The core of these programs are algorithms which are present whenever a problem needs to be solved with the help of a computer. (Danks & London, 2017, p. 4691; Alpaydin, 2010, p. 1) An algorithm is responsible for ensuring that people can read from a system whether an account is overdrawn or not, or whether according to a government system
someone is still alive or not (Domingos, 2015). By definition, an algorithm is a list of instructions telling a computer how to solve a particular problem transforming an input to an output (Domingos, 2015; Alpaydin, 2010, p. 1).

It is important to realize that even in the case of self-learning machines mentioned above (Machine Learning), there must be program code at the very beginning telling the computer that it has to learn, and parameters are provided in advance (a so-called model) which it is hoped will be made more precise through Machine Learning (Alpaydin, 2010, p. 3). Therefore, it can be said that there is always a real human factor influencing the functionality of a particular system at the core. Even when algorithms start to write of their own kind, the human origin remains, and this is where the design of such program code needs to be prudentially and carefully crafted.

Since algorithms are programmed by humans, it implies that human intention is reflected in a system (Friedman & Nissenbaum, 1996, p. 334). This intention can be good or bad and wanted or unwanted and is often brought into relation with the term bias. (Danks & London, 2017, p. 4693).

3.3 Bias

According to the Cambridge Dictionary (Cambridge University Press, 2019), the noun bias can be defined as follows: “the action of supporting or opposing a particular person or a thing in an unfair way, because of allowing personal opinions to influence your judgement...” Another term for the noun bias is systematic error which is described as “any type of ... study that result[sic] in an incorrect estimate of the association between exposures and outcomes” (Boston University School of Public Health, n.d.). Compared to random error, bias has “...a net direction and magnitude so that averaging over a large number of observations does not eliminate its effect” (The Pennsylvania State University, 2018).

As AI, by definition, is about mapping the intelligence process or transferring the intelligence of a human being to a machine, it is important to briefly illuminate bias at the origin, the human psyche. From there, an important basic feature of HI can be explained and illustrated as to identify how bias emerges out of them and is connected with the occurrence of bias in algorithmic systems.

3.3.1 The psychological aspect of how a human being thinks

HI is defined by Robert J. Sternberg (2017) in the Encyclopaedia Britannica as the “… mental quality that consists of the abilities to learn from experience, adapt to new situations, understand and handle abstract concepts, and use knowledge to manipulate one’s environment.” Everything that is human is prone to error. Humans have their algorithms to judge and make
decisions on the basis of data and information (Kraemer, van Overveld, & Peterson, 2011, p. 251). Different factors such as culture, environment and circumstances influence the way people act and think. Especially when there are situations where risks arise, mental shortcuts in human brains take over instead of rational decision making. This hinders people from making right choices. In other words, the human brain is exposed to cognitive bias and since AI is supposed to imitate the human brain/intelligence so is AI. (Rosso, 2018)

When a computer system is not free from human and societal prejudices and biases due to the development of such systems through human beings, one speaks of algorithmic bias (insideBIGDATA, 2018). Bias is mostly perceived as something negative, however this does not automatically have to be the case (Danks & London, 2017, p. 4693). Cognitive bias is necessary because the human brain can never think of all likely possibilities in a particular situation and must limit the abundance of information to make a decision. This filtering process (heuristics) to prioritise and process information is effective but can lead to unwanted results in certain contexts. (Cherry, 2018) It makes sense to look at the different causes of cognitive bias that influence a human’s decision-making-process.

3.3.2 Cognitive bias

Cognitive bias can be a product of the following causes according to Benson (2016):

Table 1: Sources of cognitive bias (Benson, 2016)

<table>
<thead>
<tr>
<th>No</th>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Information overload</td>
<td>Information needs to be filtered</td>
</tr>
<tr>
<td>2</td>
<td>Lack of meaning</td>
<td>Information is connected and the gaps are filled with what is believed to be known out of experience</td>
</tr>
<tr>
<td>3</td>
<td>The need to act fast</td>
<td>People need to act even though they are constrained by time and information (otherwise they would be paralysed and not able to do anything)</td>
</tr>
<tr>
<td>4</td>
<td>Uncertainty of what needs to be remembered later</td>
<td>The human brain has a capacity limit of storing information. People often memorise a more generic information over more specific ones because the latter would need more space</td>
</tr>
</tbody>
</table>

Number one and two combined can have a self-reinforcing character. HI consists of many different aspects like reasoning, problem-solving and learning. These capabilities are general mental abilities but how they are executed by a human being varies (Colom, Karama, Jung, & Haier, 2010, p.
The author of this thesis bases this occurrence on the idea that every human being is an individual and thinks differently, has distinct experiences, comes from diverse cultures, has other values and therefore thinks, combines and acts dissimilarly. Thus, each person sees the environment somewhat differently and influences other people with their views, decisions and actions.

Cognitive bias in human decision-making and behaviour is inevitable. As a result, it is very likely to be found in AI-systems too. The following chapter highlights different sources of algorithmic bias that are introduced to a system directly or indirectly.
4 SOURCES OF BIAS IN ALGORITHMIC SYSTEMS

The automation of decision-making is subject to the original idea that human bias is not present in algorithmic systems and neutrality and objectivity are maintained. However, various examples indicate a different picture. Human judgement and choices are reflected in systems. (Burrell, 2016, p. 3; see also Baer & Kamalnath, 2017) Algorithms are still designed by humans even if they autonomously write other algorithms and adjust parameters. Interests and values are favoured over others by developers defining the parameters used by the algorithm together with user configurations on the system which directs the results in a particular direction (Kraemer et al., 2011, p. 251). Therefore, an algorithm, in the beginning, is entirely designed by human beings with their conscious and unconscious ideas, work-related and personal experiences, guidelines, subjective and objective perceptions, opinions, creative and legal freedom and/or limitations and technological availability and/or constraints.

Machine Learning is a case of black box systems. Here, the system adapts decision-making rules autonomously, without a person making any further adjustments. This process represents the independent learning of the algorithm by defining how new inputs are to be classified. This is problematic since, in many cases, it is no longer possible to understand how the algorithm determined the respective output. (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016, p. 3)

Aspects of the earlier mentioned HI and cognitive bias are very likely to be reflected in program code. Furthermore, misunderstandings may arise with the consequence that the required business interests are not represented correctly and the desired conversion process from input data to output data takes place with distortion. Such can be applied to a system unintentionally or on purpose (Diakopoulos, 2015, p. 404).

It is fundamental to readopt the view that there can be good and bad bias depending on the context. Danks & London (2017, p. 4694) explain that there are situations in which the presence or even deliberate use of algorithmic bias in one area of the system leads to bias being reduced or even prevented in a much more important area of the same system.

The question of whether recognised bias needs to be reduced at all should always be assessed in the individual system context since mitigating bias can be a major effort. On the one hand, several associations demonstrate differences in how and which values are put in the foreground and which seem less important. On the other hand, the situation can reach a level of complexity that no matter what perspective is adopted, some bias will always be identified from a certain point of view. In the end, technology cannot fully answer questions about social and individual values. It is therefore up to humans to make sure that the particular situation is always evaluated.
in a comprehensive context, meaning taking into account the whole ecosystem around the machine. (Danks & London, 2017, pp. 4695-4696)

4.1 Direct bias

Different authors identified various sources of bias in AI-systems. Barfield & Pagallo, (2018, p. 96) illustrate the three sources: (1) Input bias where the source data is biased due to absence of specific information, non-representativeness or reflecting historical biases; (2) Training bias arises when the baseline data is categorized, or the output is assessed; (3) Programming bias emerges in the design phase or when an algorithm modifies itself through a self-learning process. These three aspects are taken up subsequently and illuminated in more detail.

4.1.1 Bias in data (input)

Data mining involves processing large amounts of data with statistical methods to identify patterns that are useful for prediction, classifying data or optimising limited resources (Data Mining, n.d.). Errors like missing, incomplete and/or unclear data in data sets can be adopted by algorithms and are either visible in the models and results or hidden in them but affect the system one way or another (Barocas & Selbst, 2016, p. 674). Bias can further arise when there is not enough diversity present in the (training-) dataset (Buolamwini, 2016b).

Another source of bias in datasets can result out of missing knowledge about the quality and origin of the used data (Miller & Record, 2013, p. 117). In the end, the principle ‘garbage in, garbage out’ (Babbage, 1864, p. 67) can be applied here stating that “…conclusions can only be as reliable (but also as neutral) as the data they are based on” (Mittelstadt et al., 2016, p. 5).

Data sets can contain sensitive and protected attributes that directly lead to bias and discrimination in the predictive outcomes. Even when there are no such direct sensitive and/or protected attributes present in the data because they were eliminated, other attributes correlated to the particular sensitive or protected attribute may reflect bias in the forecasts. For example, a person’s choice of words may be enough to identify their ethnic background, so that it seems ineffective to remove the attribute describing ethnicity from the dataset. (Pekkarinen, 2018)

4.1.2 Bias in learning algorithms (training phase)

Algorithms can be affected by bias in different ways. They process data using inferential statistical methods to deduce conclusions. Due to the nature of inferential statistic methods uncertain knowledge is produced. (Mittelstadt et al., 2016, p. 4) A statistically biased estimator can be applied
on purpose to smoothen and balance bias in training data. This renders the
learning algorithm to statistically not neutral anymore but is not consid-
ered as bad bias. (Danks & London, 2017, p. 4693)

Another point of view looks at bias that arises from misinterpreting out-
puts of algorithms in a biased way. It is not guaranteed that the interpreter
will reflect the intended point of view of the algorithm creator(s) due to
“...unconscious motivations, particular emotions, deliberate choices, so-
cio-economic determinations, geographic or demographic influ-
4694).

In Machine Learning, the system learns based on training data. Patterns
are recognised and a model for prediction is established. This model is then
tested with test data to see how accurate the model predicts when fed
with new data. A common approach is to split the source data into training
and test data. (Minewiskan, 2018) However, the test data needs to be in-
dependent, meaning data the system has not seen before. (Alpaydin, 2010,
p. 40) Bias can arise when the model is trained on the test data or the test
data contains instances that are present in the training data. In these cases,
the model will not be adequately tested and will react differently than ex-
pected to completely independent data at a later stage. (Google
Developers, 2018)

4.1.3 Bias in programming

In software business it is common to reuse code by following the reusabil-
ity principles since programming from scratch is not only time consuming
but expensive. However, the danger is to bring bias into the own code con-
struct unintentionally. The reused code may not be fully understood, does
not fit into the context or is incomplete and therefore provokes bias.
(Buolamwini, 2016a)

Further, the link between the data set and the conclusions is often not
transparent or visible. This is because it is difficult to see how the individual
data and rules have interacted through the algorithm to arrive at the re-
sult. (Miller & Record, 2013, p. 117) An algorithm can therefore be influ-
enced by human complexity where the algorithm becomes too complex
for the human brain to understand due to intricateness and too many in-
teractions between several parts of the algorithm (Domingos, 2015).

4.2 Indirect bias

Friedman & Nissenbaum (1996, p. 332) take a slightly different approach
and group different types of bias that can emerge from computer design
into the following categories: Pre-existing bias, technical bias and
emergent bias. The author of this thesis titles them as indirect bias since these biases evolve around raw data and code.

4.2.1 Pre-existing bias

This type of bias often emerges through social institutions, practices and attitudes even before a system is designed. It is then reflected in the system either through individual(s) with a considerable contribution to the design of the system or through the larger society where organisational institutions or culture carry bias which is mirrored in for example industrial or legal systems. (Friedman & Nissenbaum, 1996, p. 334)

4.2.2 Technical bias

This category includes bias that emerges from technical constraints or technical considerations. Such as computer tools, in which computer technology limitations trigger bias by favouring data (combinations) due to the order or size of screens and visual results presentation. In addition, decontextualized algorithms are prone to spread bias by treating groups unfairly under conditions such as relying on an alphabetic listing as ranking criteria. Further, pseudorandom number generators can provoke bias if they are not working properly and favour one number-range over the other(s). Moreover, as a last technical consideration, the formalization of human constructs such as quantifying the qualitative, discretise the continuous, or formalise the non-formal in AI-systems can lead to bias as well. (Friedman & Nissenbaum, 1996, p. 334)

4.2.3 Emergent bias

This category only arises through the use of the system outside of its intended context of operation and results in unfair disadvantages for individuals or groups when confronted with the system (Friedman & Nissenbaum, 1996, p. 335; see also Danks & London, 2017, p. 4694). Such shifts in context can be due to changing societal knowledge, population or cultural values. A subcategory of emergent bias is New Societal Knowledge where new knowledge is not or cannot be included into a system. Further, a mismatch between users and system design can occur where the system is used differently than intended initially such as a difference in expertise or values. (Friedman & Nissenbaum, 1996, p. 335)

Danks & London (2017, p. 4693) amend the sources mentioned above with Algorithmic Focus Bias where specific categories are left out on purpose due to strong beliefs that they are statistically, morally, legally or for other reasons not vital when in fact these categories are relevant for other classifications. In addition, part of Algorithmic Focus Bias consists of input variables that are indeed statistically unbiased but legally not permitted for
particular forecasts and assessments. Applying such input variables might turn a neutral algorithmic system into a biased one.

From another perspective, bias can be transferred from the outside into the system through the behaviour of the user(s) that utilise(s) the system. The learning system takes the inputs of the users and adapts itself according to them. Discriminatory behaviour of society can thus enter the system and be continued in the further Machine Learning process. (MIT Technology Review, 2013)

A different point of view can be taken if one considers that the field of AI development is still dominated by men and thus the inputs of women are not present enough. This is prone to be reflected in these systems and continues to have an impact on society (Campolo, Sanfilippo, Whittaker, & Crawford, 2017).

Several stages make up the process of development-implementation-application for an AI-system. These are data sampling, measurement, algorithm design, algorithmic processing, application in the real world and interpretation by a human end user or some autonomous system. Algorithmic bias can arise in any of these stages and each entry point needs to be examined separately. (Danks & London, 2017, p. 4696) While the sources classified as direct bias can enter the system at the described stages input, training and programming, the author of this thesis believes that the indirect biases can penetrate a system through each of the stages mentioned by Danks & London (2017, p. 4696) at any given time.
5 LACK OF TRANSPARENCY IN ALGORITHMIC DECISION-MAKING PROCESS

This chapter deals with the challenges that arise when it comes to the usage of AI-systems. It then takes a glance at the topics AI-fairness, AI-responsibility and AI-safety.

5.1 AI challenges

Algorithms are penetrating more and more into people’s lives and will likely overtake even stronger parts of their daily routine so that they will depend heavily on how secure and efficient these algorithms are. Considering that algorithms are becoming increasingly complex (A. Smith, 2018), it is continuously challenging to predict how machines will behave when exposed to real world conditions and when and how they will fail. (Tutt, 2016, p. 86) In other words, machine behaviour becomes harder to forecast in the future.

Algorithms are not trained to solve a specific problem but are trained to learn to solve it and how they learn is usually not fully understandable. Moreover, applying this behaviour under the intense condition of reality brings with it considerable uncertainty. (Tutt, 2016, p. 87, 102)

According to the Future of Life Institute (n.d.), Max Tegmark (n.d.) specifies that while on the one hand AI can contribute undoubtedly to HI and can drive society forward, it must be ensured on the other hand that these technologies are kept healthy and beneficial to the population. Therefore, as AI-technology continues to improve, it is becoming increasingly important to ensure the security of such systems. It is about keeping control of the machines and knowing their impact on society and business.

The following thoughts are frequently brought to light when covering the aspect of fear towards AI:

The idea that AI has the potential to become smarter than a human being is a fear because it is uncertain and hard to predict what exactly could happen in such a scenario. Errors made by AI-systems might, in the future, have a far greater impact on society and individuals than nowadays. Thinking about attacks in air traffic control systems, errors in self-driving cars (Future of Life Institute, n.d.), there may be questions as to what would happen if the technology becomes even more complex and interconnected so that services such as search engines, maps, clouds and others, which are taken for granted nowadays, go down and leave users with situations they have never had to cope with before. It will resemble that a step forward might become several steps back considering the dependence and interconnectedness of these technologies.
Further, people worry about AI taking many jobs away leaving many people unemployed. It is unknown how the money generated through those AI-systems will be distributed in the society. In addition, the topic of singularity was pointed out where it is feared that humanity could become entirely dependent on AI. (CBT Nuggets, 2018)

Yet another input is provided by Elon Musk who raised his voice to address concerns about AI and its future development. He requests for standards and regulations be set out on the grounds that AI runs the risk of being weaponized in the wrong hands. Even best-intentioned AI projects could end up ignoring ethics and execute an operation at all cost since machines function as they were taught to. (OBT, n.d.)

Moreover, another concern is that systems are exposed to circumstances that a human has not yet been in and might never be confronted with. The system is not able to rely on experience and might make errors no human would ever have made in given situations. (Tutt, 2016 Calo, 2014, p. 7; Gorban, Grechuk, & Tyukin, 2018) This is another reason why it is becoming more difficult to predict how, if and when a system will make serious mistakes or fail which can bear fatal consequences. This in turn brings with it a very different issue. The difficulty of measuring, tracing and assigning algorithmic responsibility is a problem and will likely keep researchers busy for many years. Not to mention that situations where experts cannot find a way to teach the algorithm not to make the same error again as it was the case with Watson might arise. This makes it even more complicated to treat machines like humans since a human being would try not to repeat the same mistakes. (Tutt, 2016, pp. 89, 105, Hamm, (2011))

As Calo (2014, p. 14) states that unique challenges never presented before to law and legal institutions will arise through robotic technologies, the author of this thesis believes that this is true for AI-technologies as well. Ironically, precisely what makes an algorithm useful is what makes it so dangerous due to the inability of humans to understand and explain the algorithm’s behaviour (Tutt, 2016, p. 90).

AI is expected to “...do things which, at the moment, people do better” (Rich, Knight, & Nair, 2009). However, according to Cathy Bessant it will not do this unaided (Bank of America, 2018). Therefore, the focus must be on ensuring that determined rules are followed when creating the system to better ensure AI-safety. The establishment of agencies which only release the algorithms for the market if their safety and efficacy are guaranteed is suggested. (Tutt, 2016, p. 111; Calo, 2014, p. 11) Nevertheless, this is difficult to put into effect since questions arise addressing topics like what is to be considered as safe and effective and how reliable tests can be considered wanting to imitate the real world’s circumstances (Tutt, 2016, p. 103). The presence of bias in a system and the challenges mentioned earlier indicate that the AI-systems are likely to be remote from a desired ideal state. Consequently, the term AI-fairness is illuminated.
5.2 **AI-fairness**

The term fairness depends on the context it is used in. A system might be viewed as fair in some circumstances and in other situations it might be considered unfair. In addition, the existence of bias in an AI-system cannot be considered evidence in classifying a system as unfair, meaning neutral or even desirable bias can be present in AI-systems without leading to undesirable results. (Danks & London, 2017, pp. 4694-4695) Therefore, classifying an AI-system as fair is subjective and depends on the viewer.

Nevertheless, known personalities in science and technology like Stephen Hawking, Elon Musk, Steve Wozniak, Bill Gates and many others have addressed their concerns about AI-risks that could arise and demand regulations to keep AI safe and beneficial for humanity. (Future of Life Institute, n.d.) Different approaches are being pursued to meet the challenges.

5.3 **AI-safety and AI-responsibility**

Institutions are working on establishing standards under which developers and designer create less biased AI-systems. The non-profit organisation AI Now at New York University pursues AI-fairness with the rule “When it comes to services for people, if designers can’t explain an algorithm’s decision, you shouldn’t be able to use it” (Isele, 2018). Google also drives society, business and the world towards recognising the seriousness of the topic of bias in AI-systems. Projects that Google initiated include GlassBox, Active Questioning and Answering and PAIR initiative (People + AI Research) (Isele, 2018). The US Defence Advanced Research Projects Agency takes the approach of making AI-systems more accountable to their users. This project is called XAI (Explainable AI) (Isele, 2018).

On the other hand, efforts are made which do not intervene before the design of the system, but with black box systems trying to unpack them and explain the algorithms’ outcomes and solutions. This area deals with Understandable AI rather than Explainable AI. It is proposed that the human can intervene in ambiguous decisions of the algorithm and identify the nature of parameters that lead to the specific issue which in turn a UX/UI designer could convert into a descriptive AI output. (Chowdhury & Lake, 2018)

All these attempts are aimed at ensuring that system developers and designers have more responsibility in creating algorithmic systems. The demands for standards and regularities to be introduced may lead to more transparency in the system creation process and promote the possibility of being able to explain and understand the results of algorithmic systems. Another initiative relevant to this thesis is described in more detail subsequently.
5.3.1 The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

This initiative has the goal to provide “an incubation space for new standards and solutions, certifications and codes of conduct, and consensus building for ethical implementation of intelligent technologies” (IEEE, 2019b). The Ethical Aligned Design (EAD) is one of the main focuses of this Global Initiative. The EAD is a document that invites public discussions about standards and principles that help to prioritise human well-being. This is to be achieved by establishing “ethical and social implementations for intelligent and autonomous systems and technologies, aligning them to moral values and ethical principles”. (IEEE, 2019a)

The topics covered in the EAD are converted into actionable guidelines of frameworks with the aim to supply feasible industry standards. At the moment there are eleven IEEE P70xx standards under development. (Koene et al., 2018, p. 38) For this thesis the Standard IEEE P7003: Algorithmic Bias Considerations is of most relevance and therefore presented further.

This standard strives to offer creators of AI-systems the opportunity to ensure customers and regulatory authorities that up-to-date best practices have been applied during the different phases of the system creation process. As of today, the standard proposes the following procedures during different phases (Koene et al., 2018, p. 39):

In the early development cycle, it is suggested that the developer evaluates the possible customer groups to identify if there are special requirements necessary for certain subgroups which could be disadvantaged due to particular impairments. (Koene et al., 2018, p. 39) Afterwards, a testing dataset is to be created which represents all the identified customer groups on an equal level. This dataset is then used to test the current system and evaluate its validity and desired performance. The developer is asked to take into account legal and social aspects concerning the criteria that are defined as relevant and are applied in the recommendation process. The goal is to examine why exactly these criteria are applied and justify their usage. (Koene et al., 2018 p. 39) The documentation of these measures ensures that the business has dealt with the system’s context in detail and supports responding to future enquiries concerning the recommendations of the particular system. (Koene et al., 2018, p. 39)

To design systems that will be easier to audit by external parties (e.g. regulatory bodies) the standard describes several methodologies for creators of algorithmic systems. The methodologies mainly focus on designers and developers being able to “…assert how they worked to address and eliminate issues of unintended, unjustified and inappropriate bias in the creation of their algorithmic system” (Koene et al., 2018, p. 39).
As of May 2018, the methodologies contain the following elements among others yet to be determined (Koene et al., 2018, pp. 39-40):

- “A set of guidelines for what to do when designing or using such algorithmic systems following a principled methodology (process), engaging with stakeholders (people), determining and justifying the objectives of using the algorithm (purpose), and validating the principles that are actually embedded in the algorithmic system (product)
- A practical guideline for developers to identify when they should step back to evaluate possible bias issues in their systems, and pointing to methods they can use to do this
- Benchmarking procedures and criteria for the selection of validation data sets for bias quality control
- Methods for establishing and communicating the application boundaries for which the systems has been designed and validated, to guard against unintended consequences arising from out-of-bound application of algorithms
- Methods for user expectation management to mitigate bias due to incorrect interpretation of systems output by users (e.g. correlation vs. causation), such as specific action points/guides on what to do if in doubt about how to interpret the algorithm outputs
- A taxonomy of algorithmic bias”

5.4 Consequences discarding fairness

The author of this thesis concludes that if bias in algorithmic systems is not identified and not reduced, the system may provide recommendations or results that have discriminatory effects on specific population groups and/or individuals. Further, if the topic of AI-fairness is neglected, this can lead to bias in systems being transferred to society and remaining there (Ding & Triolo, 2018). Originally, the creators of AI-systems largely determine people’s lives. (Finn, 2017)

The consequences resulting from the execution of discriminatory results affect not only the victims but also the business that has allowed these decisions to go through. Image damage is very likely and can be difficult to correct. It is therefore in the interest of system manufacturers and end users to create systems that are as fair as possible.

The mentioned difficulty of measuring, tracing and assigning responsibility for the algorithmic system (Tutt, 2016, p. 105) is in the eyes of the author of this thesis, a warning signal that the subject must be taken seriously. For if it is already considered complicated to deal with these responsibility aspects nowadays, it will not become easier in the future with the progress and increasing complexity of technology.
6 IDENTIFYING AND MITIGATING BIAS IN ALGORITHMIC SYSTEMS

This chapter suggests possibilities on how to identify bias in algorithmic systems. First, the requirements for being able to use bias identification methodologies within a company are presented. Afterwards, suggestions based on the results of the literature research and analysis are presented.

6.1 Mindset

Awareness of the topic is the first step towards addressing bias in algorithmic systems. Reports from SAS, Accenture Applied Intelligence, Intel and Forbes Insights document that 92% of AI-leaders make sure their technologists receive ethics training and 74% of the leaders assess AI-outcomes every week. (SAS Institute Inc., 2018) However, Rumman Chowdhury, Responsible AI Lead at Accenture Applied Intelligence, reports that even though these are steps moving forward in the right direction, it is not enough for a company to just dispose of directional AI ethics codes preventing harm. According to the SAS Institute Inc. (2018), Ray Eitel-Porter (n.d.), head of Accenture Applied Intelligence UK suggests establishing usage and technical guidelines which form part of a governance process that guarantees AI-transparency, AI-explainability and AI-accountability.

6.2 Possibilities to identify bias

In the literature analysed, the topic of ethics is often given priority when it comes to algorithmic distortion. According to Mittelstadt et al. (2016, p. 2), it is very laborious to recognise subjectivity in the design and configuration of algorithms. It is argued that subjectivity would only come to light through the rise of a specific use case and would maybe continue to exist even when some examination towards subjectivity has taken place in advance. Nevertheless, efforts are being made in this area to identify existing bias in algorithmic systems.

An approach to dealing with bias is monitoring and auditing the AI-system’s creation process comprehensively (Raymond, n.d., p. 37). This includes having a cross-disciplinary team on behalf of the public as well as private sectors which can question the whole process from various perspectives. Humans with different backgrounds such as ethnicity, gender, culture, education, age and socioeconomic status are necessary to ask the right questions and comment on the results the system generates. (Isele, 2018)

Friedman & Nissenbaum (1996, p. 334) suggest applying rapid prototyping, formative evaluation and field testing continually during the design process of a system to identify bias that is unintentional. Furthermore, it is
beneficial that the view is raised beyond the internal system’s features and evaluate it in the context of future use. In this way, both technical and other bias can be identified. It may be something as simple as asking whether a list should be displayed in alphabetical order or not. Such a question is important, since it deals with significant aspects like the simplicity of access or distributive justice of access. (Friedman & Nissenbaum, 1996, p. 334)

Another possibility consists of manipulating the test data purposefully to determine whether the results are an indication of existing bias in the system. For this, cases can be simulated where similar data is used, and the algorithm tested for whether one case is discarded or not. If the former occurs, it could be further tested to see if overlooking a specific feature would change the outcome. If this is the case, it can be said with great certainty that the system is biased concerning the specific attribute. (Pekkarinen, 2018)

In addition, audits should be executed in a way that promotes critical thinking and challenges assumptions through asking and answering questions (Socratic method). In doing so, scrutiny and reformulation play a central role in the identification and reduction of bias. (Isele, 2018) A. Smith (2018) suggests addressing questions like “What is the worst thing that could happen in this algorithm once it starts interacting with others?” or what is “the least optimal case, and how do we identify and measure it?”

Furthermore, language is historically biased and therefore associations and relationships between words and images can easily cause bias in systems when overlooked (Isele, 2018). It is therefore necessary to identify and examine the field of the system usage and evaluate which vocabulary will most likely be used in the circumstances the system is designed for to provide as much equity in language as possible. This helps planning how the algorithm needs to be designed in case it is a learning algorithm that learns based on the inputs the users provide it (especially in the field of Natural Language Processing) and thus taking measures that mitigate the risk of ending up having a system that turns into a biased machine. An example where this has happened is the robot Tay created by Microsoft. As Hunt (2016) explains, Tay turned into a robot which held racist and inflammatory conversations with Twitter users which contained a great number of political statements. It learned from the user inputs and integrated most of them in its answers.

Requesting information about how the design process of the AI-system took place is recommended. It is crucial to evaluate if the people involved have received bias training and are aware of the subject and how to monitor and address it (Isele, 2018).

From a technical perspective, there are several open source tools available that claim detecting and mitigating bias in algorithmic systems to a certain
level. For example, IBM has launched a tool called AI Fairness 360 (AIF 360). IBM states that it is “...a comprehensive open-source toolkit of metrics to check for unwanted bias in datasets and machine learning models, and state-of-the-art algorithms to mitigate such bias” (Varshney, 2018). Google also launched a tool with the name What-If with which it is possible to visualise inference results, edit data points and see how the model performs, exploring the effect of a single feature and test algorithmic fairness constraints among other functionalities. (What-If Tool, n.d.)
7 FRAMEWORK FOR BIAS IDENTIFICATION AND MITIGATION

This chapter exhibits the results of this thesis in the form of a framework. Recommendations are made based on the research and analysis of this thesis. First it describes how the results section of this thesis is performed and provides necessary background information. Then the framework is presented and the individual aspects examined in more detail.

7.1 Justification for the results section

Based on the research it is known that bias is present in the human decision-making process and can be reflected in people’s activities. Thus, bias can enter algorithmic systems and influence decision-making accordingly. As a result, bias can ultimately lead to discrimination and disadvantage to social groups and individuals. The research has also proved that AI-systems are still in the narrow range at the moment and therefore cannot yet be considered as independent, intelligent entities. The author concludes therefore that human influence on AI-systems can still be controlled and therefore bias to some extent be prevented from being reflected in the system.

Based on these findings, the author has created a framework to support all members involved in the process of creating an algorithmic system. The author suggests an approach on how bias can be prevented, monitored and mitigated during the development, implementation and application processes of an algorithmic system.

7.2 Framework creation

The designed framework covers all aspects the author of this thesis considers most important. It is a metamodel which is refined by a checklist for the areas depicted in the model. The elements of the checklist consist of concrete statements and questions which need to be addressed by the project team. These are justified by presenting the considerations made by the author of this thesis and referring to the research section. The checklist is presented in the appendix ‘APPENDIX 1 CHECKLIST’.

Factors that are regarded as relevant to avoid bias as far as possible are supported by standards suggested by IEEE and other projects and statements that aim to prevent and mitigate bias, as explained earlier in this thesis. Ideal situations are illustrated which could be used as suggestions to improve the project team’s circumstances.

The framework should be understood as a guideline. This guideline aims to present steps towards ensuring that basic properties, which should
prevent bias from entering an algorithmic system to a certain extent, have been taken into account. The author of this thesis refers to the development, implementation and application processes from here on as ‘algorithmic system lifecycle’.

7.3 **Goal of the framework**

The framework should be seen as a ground-breaking tool through the entire algorithmic system lifecycle, using the framework claims to lay the foundation to minimize or even prevent bias in the algorithmic system lifecycle. It helps to ensure that the most relevant aspects of possible bias sources have been considered and investigated.

7.4 **Prerequisites for framework application**

The author reached the conclusion that an overarching governance must be in place in the company that determines and enforces AI-responsibility. Under it the developed framework can be used in project teams. The project members should be committed to the framework and therefore it needs to be considered as a binding standard in the company. There has to be a working atmosphere where the opinions of each team member can be expressed and weighed up the same, independently of what position the team member holds in the project and what their background is. A proper training and sensitization for this topic are fundamental to apply the framework successfully, not only to uncover unconscious but also self-evident aspects that are important for preventing bias.

7.5 **Scope of the framework**

Based on the study, the author has become aware that the area of bias in algorithmic systems is extensively large so that it is not possible to consider all factors related to this topic in this thesis. The framework is therefore mainly based on the research section and presents recommendations based on the findings.

7.6 **Metamodel**

The metamodel expresses the areas that must be considered during the algorithmic system lifecycle to recognise and minimise bias. Each aspect illustrates where potential bias can arise and enter a system. The individual areas of the metamodel are described hereafter.
Figure 2: Metamodel. Identifying and mitigating bias in algorithmic System

7.6.1 Project Team

Due to the insight that pre-existing bias is very likely (as a consequence of different goals and distinct backgrounds of each individual) every team member can transfer their knowledge into the project to create the algorithmic system (Friedman & Nissenbaum, 1996, p. 334). Thus, measures must be taken to ensure the neutrality of the system as far as possible. Knowledge, views and attitudes of individual team members cannot be deleted or hidden, as these are usually unconscious factors, due to each individual’s different background and experiences. Therefore, it is necessary that there is an exchange among project members where everyone shares their views, ideas and concerns openly, fully and transparently before creating the system. This serves to avoid (data) exclusions.

Speaking about exclusion, it is indispensable to make it possible to see whether there are misunderstandings, ideas of conflict, too much euphoria
and whether the culture of each member would influence system behaviour through unconscious assumptions that would otherwise not be expressed. An open exchange also reveals whether some aspects would have been lost if expressing opinions had not taken place. Pre-existing bias at project team member level can thus be identified ensuring that the following elements are met:

- **All project members**
  - Have had ethical training
  - Are aware of the topic of bias that exists in the human decision-making process
  - Know about the fact that human bias can be reflected in an algorithmic system
  - Consider the same attributes and factors as most relevant in the system context

- **The project team**
  - Represent stakeholders of all possible end user groups
  - Is a cross-functional team including diversity in ethnicity, gender, culture, education, age and socioeconomic status
  - Has representatives from the public as well as the private sector

- **Independent consultants**
  - Are included for comparison with competing products

For the uncovering of pre-existing bias in the context of cross-company views and attitudes, independent consultants should be included and comparisons with competing products made. Thus, it can be recognised whether unconscious attitudes of the corporate culture would lead to unwanted bias. It must be ascertained whether the company strategy, goodwill, principles, attitudes and practices have an influence on the creation and final state of the algorithmic system and limit its implementation in such a way that adjustments and decisions must be made that could lead to bias.

It is also advantageous to examine the experiences and backgrounds of the individual project members since bias already begins with the human being itself. Furthermore, as stated in the research, the field of AI is dominated by men and inputs from women are likely to be missing. (Campolo et al., 2017) It is therefore important to ensure that from a gender point of view the composition of the project team is examined and undesirable influence on the functioning of the system is detected.

### 7.6.2 Environment and Context

The realisation was reached that emergent bias can originate. Namely, when the system is in use in a context that no longer corresponds to what the system was originally created for. Or, if the end users depict a different
behaviour or different cultural values and attitudes than initially assumed as Friedman & Nissenbaum (1996, p. 335) state. The framework addresses these challenges with the following elements:

- All possible end user groups are included in the testing phase
- All possible end user groups have been evaluated
  - End user groups’ behaviour is monitored and evaluated from different perspectives (surveys, interviews, recording behaviour, letting them explain what they do and think while testing)
- Consequences and intentions have been considered
  - For what and with what intentions was the system created for?
  - “What is the worst thing that can happen in this algorithm if it starts interacting with others?” (A. Smith, 2018).
- Context is faithful to the original source
  - Does the current context represent the one, for which the system was originally created?

7.6.3 Constraints

A system often has to be developed within regulations and restrictions. These can be industry regulations, standards, quality restrictions, laws, boundaries, cost considerations or other mandatory restrictive measures. These limitations can affect the entire system creation process. For example, not enough personnel resources may be available, technical tools and aids may not be available, data may not be obtained in its entirety or patents may require a fundamental change in the system to name a few examples. Such circumstances give particular rise to exclusions, respectively all the direct bias types and technical bias. Since for instance, few personnel resources might lead to unattractive and ill-conceived data mining, code writing and/or testing. Quality requirements might lead to specific usage of methodologies that have not been applied in the company before. This lack of experience is more susceptible to bias. Laws can lead to the fact that certain software and hardware and procedures may not be used, which has the consequence that expertise is not sufficiently available, and bias can easily arise. The technological bias can occur in hardware if, for example, the screen size influences the representation of results, as illustrated earlier in this thesis (Friedman & Nissenbaum, 1996, p. 334).

It is therefore essential to determine, on the one hand, what impact regulations have on the system and, on the other hand, to look beyond the system context in the specific development environment and to evaluate what the technological characteristics and compositions look like in the reality where the system will ultimately be used. The following points can help tackle constraint challenges:

- Business aspects are reviewed
  - Under what circumstances will the system be developed?
- Scope is reviewed
- Technical aspects are reviewed
  - Do technical constraints affect the way the system is designed?
- Legal aspects are reviewed
  - Do regulatory/law constraints affect the way the system is designed?

7.6.4 Input (datasets)

Input bias was identified as a type of direct bias. The findings demonstrated that various factors could cause this bias. The data sets can be incomplete, contain wrong data, represent the wrong context or contain bias themselves (Barocas & Selbst, 2016, p. 674). Further, the correctness or quality of the data may not be ensured or verified (Miller & Record, 2013, p. 117), or law and regulations prohibit the use of certain attributes for algorithmic systems with recommendation character. (Danks & London, 2017, p. 4693)

Because data mining consists of statistical methods, inferences are drawn on the input data to create models that are then used to make predictions (Data Mining, n.d.). The use of statistical methods must therefore be well considered and one must be aware of the implications they will bear. It is also necessary to make demands on the data set (quality, scope, diversity) so that the application of generalisation does not lead to bias or at least remains within a manageable scope. This is mentioned because it has been recognised that bias in one place in the system is not necessarily considered to be bad bias and can prevent or minimize bias in another, much more important place in the system (Danks & London, 2017, p. 4694).

Another factor which is not to be underestimated is how well the data in the data set is understood by the project team. A lack of understanding of data entirely can lead to misunderstandings and incorrect interpretations. A finding of this thesis research concludes that causes of cognitive bias are when gaps are filled with information that is believed to be known from experience “Lack of meaning” (Benson, 2016). The existence of cognitive bias in that particular case can lead to the system behaving differently than previously thought based on the faulty interpretation of the data. Therefore, the subsequent aspects need to be fulfilled:

- The data set is fully understood
  - The meaning of each attribute in the context is understood and its purpose in the system context is clear
- Data is transparent
- It is ensured that the data set represents the correct scope (enough data to represent a population or target group)
- The source of the data is known and verified
- The quality of the data is ensured
- It is clarified which attributes can legally be used
7.6.5 Training data

Because an algorithm in machine learning learns based on training data, this data must first be determined, and it is derived from the original data set (Minewiskan, 2018). In some cases, it may be necessary to categorise and clean up the data so that the system can process it as usable training data. Adapting data sets can result in technical bias where human constructs such as qualitative aspects are expressed in numbers, the continuous is made discrete or non-formal aspects are turned into non-formalized ones (Friedman & Nissenbaum, 1996, p. 334).

The elements mentioned in the category Input(datasets) of this framework also apply to the training data. A review of the original data is not sufficient if it is subsequently changed so that the test data is no longer representative or of the same quality, or even biased in itself by omitting or adding attributes. Training data must therefore be created and provided very carefully and deliberately.

- The training data set is still as representative as the original data set
- Added or omitted attributes are carefully chosen and justified

7.6.6 Test data

The research has implied that accurate testing of the system depends on the quality of the test data. How accurately the model predicts can only be tested if test data is independent, i.e. if it represents new data that the system has not yet seen. (Alpaydin, 2010, p. 40) When the system is in production; it is fed with data it has never seen before. Testing with data that is not independent would result in false assumptions and conclusions being made about the precision of the system or model.

The author of this thesis proposes to define test scenarios with similar data, which have a distinction in a critical attribute. This way, it can be detected whether the system treats both instances equally or discriminates against one instance.

If bias is identified, then it should be checked whether omission or addition of (sensitive) attributes reduces or even eliminates the bias. Omission and addition of attributes must be justified and achieve the desired effect. As seen in the research, sometimes it is not enough to remove only sensitive attributes due to the fact that other attributes may have correlations or links to them, and the system continues to be biased exactly the same based on this data. For illustration, recall the example where removing the gender attribute is useless because there is another attribute that contains words that can be used to identify gender. (Pekkarinen, 2018) The framework addresses the following elements:
- Test data is independent
- Test data is defined
- Test data is reviewed

7.6.7 Project Management

It is suggested that the complete algorithmic system lifecycle should be accompanied and controlled through all phases with a project management approach. The classical element risk analysis must be expanded with a focus on risk factors that could favour bias and the effects recognised bias could have.

The stakeholder analysis should also be broadened so that groups identified as potentially disadvantaged are more thoroughly investigated and involved. This involves a change of perspective where a worst-case scenario with regards to discriminatory effects represents the starting point.

An attitude and a mindset to continuously question the system creation process should be established. As found in the research, the right questions should be asked, critical thinking adopted, assumptions challenged, and the results of the system evaluated. (Isele, 2018) This can only happen if an actively lived sceptical attitude is steadily applied. Monitoring and auditing methods can be used for this purpose.

The author of this thesis considers it necessary that doubts and fears that arise among team members should be given space. An open and understanding communication culture without judgement should not be underestimated. In the end, data reflects society and thus a fear or idea may reflect only a fraction of humankind however linked together new views and insights can be gained that would otherwise remain hidden. An input from one team member can lead to a chain of other ideas and thoughts from other team members. As mentioned in the research, certain system subjectivity can only come to light when a specific use case occurs. (Mittelstadt et al., 2016, p. 2) That is why it is important to go through possible scenarios in advance, and this requires the diverse views, opinions and interpretations of as many participants with different backgrounds as possible. The following elements are to be focused on:

- Project management process includes methods that focus on bias issues
  - Stakeholder analysis is adjusted for disadvantaged group identification in worst case
- Risks concerning bias are assessed and known to each team member
- Critical thinking is promoted and demanded at every stage of the system creation process
- Perspectives are changed continuously to challenge assumptions
- Monitoring measures are defined, communicated and applied
- Auditing measures are defined, communicated and applied
- Workshops/meetings are set which address upcoming doubts of team members
- Scenario thinking is fostered
- Freedom of expression is guaranteed and desired

Google (2018) suggests addressing the following questions to foster critical thinking: (1) “How would changes to a data point affect my model’s prediction?” (2) “Does it perform differently for various groups—for example, historically marginalized people?” (3) “How diverse is the dataset I am testing my model on?” The developed What-If Tool helps addressing these questions to make it easier answering these questions.

7.6.8 Hardware

The research of this thesis has revealed that hardware aspects can be a cause of bias in algorithmic systems. Purely visual limitations can influence the result in such a way that the end user interprets it differently than intended. For example, the screen size was mentioned, which can determine the arrangement of results of a system. (Friedman & Nissenbaum, 1996, p. 334) It is possible that equally important results are only displayed on a second page. To reach this page requires additional human action (scrolling or clicking). As a result, the end user is influenced because they think either out of comfort or conviction that the results on the first page must be more important than the others. The focus is then laid more on the results on the first page and other equally important information may not be considered at all.

In addition, the author of this thesis expects that hardware limitations such as the speed at which the system performs, or the availability and quantity of hardware during system creation but also in the production environment, influence the system architecture which can lead to measures where restrictions are taken which then cause bias in the system. For example, if, due to performance problems fewer results can be displayed than are available. Or, when too few servers are at disposal, and the amount of data has to be reduced. Excluding data can mean that the data set is no longer representative.

In such cases, hardware can lead the interpretation of the end user to a certain direction. It is therefore necessary to consider the following questions:

- Do hardware limitations exist?
- Do these limitations influence the system creation process?
- Do these limitations influence the system’s functionality in the production environment?
7.6.9 User interface

Speaking about human actuators such as eyes or ears (Russell & Norvig, 2010, p. 34), it is necessary that the user interface (UI) is designed carefully. The user interface can influence how the end user interacts with the system and how they interpret results. The representation and placement of text and forms in the user interface are of central importance as it influences the user's perception and experience. For example, the size and colour of text and shapes determine where the user's focus is first directed to. Graphics and images lead to associations in the user's brain. They associate them with experiences or ideas that affect the neutral perception of the system results. The design of the visual representation should therefore be questioned, and different characteristics should be tested with users. As suggested in the research, rapid prototyping can be used to quickly receive feedback from users and obtain results that are truncated to the user faster. Furthermore, different forms of evaluations like formative, process, outcome, impact, goal-based or summative (SocialCops, 2017) can be brought in to find out how users react to certain system elements and what they perceive as important and unimportant. For example, it can be found out whether displaying results in alphabetical order has an influence on the results, their interpretation and finally the actions of the end users.

Another aspect the author of this thesis considers as essential is the usage of languages in the system context. It must be taken into account which language the data on which an analysis happens is available and which language the system must have in the end. In case the source data language differs from the end system language, a translation is likely to be applied. This process may change the meaning of the information and falsify the results. On the one hand, the natural application of languages through translation can mean that information and terms have to be written in another way because they exist differently in one language or do not mean the same thing. A wrong translation changes the context and might lead to the end user having a different interpretation. On the other hand, it is possible that the translation does not make much difference between two languages from a terminological point of view, but the context is interpreted differently by the user group. For example, in different regions the same word or expression means something dissimilar or is used more or less frequently. The choice of words affects the interpretation, the perception and the stream of thoughts and thus the decision-making process of the end user. The various meaning of words and information in different contexts can be attributed to the historical distortion of language as mentioned in the research section (Isele, 2018). Considering these aspects, the system's default language should be the one the target user group is fluent in. The following elements are suggested to be evaluated:
- Visual aspects are determined appropriately
  - Text: the font-style, font-size, font-colour and placement are justified and reflect the intention of the system’s functionality
  - Forms/elements (e.g. boxes containing text or graphics): colour, size and placement are justified and reflect the intention of the system’s functionality
- Does visual result representation (alphabetically or random) make any difference (user always choses the results displayed first?)
- Does a change in navigation representation lead the user to favour different results?
- Is graphical UI limiting/favouring data over other data?
- Is a translation of data/information necessary?
  - How does the chosen language influence the user’s perception and interpretation in different contexts and circumstances?
- Do the information and results become distorted through the application of translation?
  - How is the translation interpreted by the end users?
- The system features mentioned above are changed and end users are monitored on the above elements once more to see how their behaviour changes

7.6.10 Programming

The direct adoption or reuse of all or part of the code has disclosed in the research that unintended bias can be transferred to the own code. The reason for this can be the lack of understanding of the adopted code. However, also incomplete and not for the context intended code can be the cause. (Buolamwini, 2016a) Therefore, criteria and guidelines (code conventions) should be defined for how to deal with the adopted code. These can be measures such as examining the code in more detail in an extra review or having it analysed by different developers and seeing whether they all arrive at the same result as for how the code affects system functionality.

Not only taken over code can be a problem, but also the self-programmed constructs. As can be concluded from the research, a programmer transfers his/her personal attitude, experiences, opinions and ideas into the program code. The factors that trigger this kind of bias source are innumerable since any kind of bias that can affect people in everyday life can affect the programmer as well, since it is still the same person. Moreover, perhaps experiences outside of work contribute to how a programmer interprets demands. Misunderstandings about business requirements can also arise if the programmer implements the specifications in a different way than the customer has imagined. This discrepancy does not have to be immediately apparent but can only appear through the interaction of several data sources and different algorithmic systems or even remain unnoticed. As mentioned in the research section, the interaction of algorithms, especially self-learning algorithms, can become so complex that
even the creators of the systems do not understand them fully (Sentient Technologies Holdings Limited, 2018; see also C. Smith, McGuire, Huang Ting, & Yang, 2006, p. 21).

To reduce the subjectivity in the program code, code audits can be performed where real test scenarios are applied. For example, the input data is clear to all participants and then it is tested whether the output meets the expectations of the participants. This way it can be determined whether the specifications have been implemented correctly or whether there are misunderstandings.

In the case of systems where the algorithm learns based on user input, it must be ensured to think outside the box. It must be thought in advance, which information and behaviour users will provide to the system because it will adapt to them. Measures must be taken to prevent the system from turning into a system that discriminates against certain groups. For example, not letting the system accept and train on certain critical keywords.

External audits may also be considered to address the above-mentioned issues. With the effect that independent and more neutral views can be gained. The subsequent elements are suggested to be considered:

- Code reviews take place
- Independent code audits are conducted
- Possible user behaviour is analysed beforehand to keep a learning system from adopting discriminatory behaviour

7.6.11 Deliberate Bias

Bias can be deliberately applied in some cases to prevent bias from arising in another, more important area of the system. As an example, a statistically biased estimator was mentioned in the research of this thesis (Danks & London, 2017, p. 4693). However, it is also possible that bias is not intentionally actively produced and inserted into the system but exists and is not regarded as harmful bias since it does not influence the end result or cause any discriminatory consequences. The research has exposed that not every bias is bad (Danks & London, 2017, pp. 4694-4695). Therefore, it is important not solely to eliminate every bias, but first to investigate whether this is necessary at all. If it is difficult to predict whether the bias is good or bad, then it is important to keep an eye on it and test its effects by, for example, adjusting parameters that would favour the specific bias if it were present. Through the framework the following elements should be taken into account:

- Bias is identified and categorized
  - Are the identified biases considered as good, neutral or bad ones?
- Is there any bias which was implemented on purpose to mitigate other?
- It is ensured that all the identified biases are monitored during the whole system creation process

7.6.12 Documentation

Documentation is necessary to help the involved people to understand how to use, repair and maintain the system and is an aspect that should be carried out as accurately and comprehensively as possible. This ensures traceability, justification and thus business continuity. The author of this thesis suggests that business should define and impose binding standards with clear requirements for the documentation of AI-systems.

- Are the relevant information present?
- Is the documentation comprehensible?
- Has the documentation been reviewed and approved?

7.7 Application of the framework

This framework is intended to be understood as a basic framework. This means that it can be extended with specific aspects based on the context. The research has revealed that the area of bias in human behaviour and in algorithmic systems is very large and complex. Therefore, the framework has a guideline character and needs to be finalized individually by the particular project team.

The project team adjusts the framework in a way that its content meets the project team’s ideas of system neutrality. The elaborated framework can then be used as a standard within the corresponding project. This standard ensures that the system achieves the neutrality defined by the project team. Testing whether the standard (framework) has been applied and the requirements have been adhered to will help to find out how neutral the system is and where needs for action exist.
8 CONCLUSION

With AI, humankind strives to leave decisions to the machine and use it to fill gaps that humans alone cannot fill. The machine should perform the tasks faster and more accurately than humans to simplify processes and open up business opportunities that were never there before. The research of this thesis has pointed out that HI is expected to be mapped to AI however, at this moment AI-systems are only semi-intelligent because they still do not have their own awareness and are dependent on a programmed sequence of instructions (algorithms).

Because these systems are still partly intelligent, they are shaped by the influence of humankind and with it by bias which by nature is present in humans and then reflected through society and individuals in systems. Often, too little attention is paid to the fact that the goal pursued by the machine was influenced and predetermined by humans and the consequences as a result. However, it is in the hands of the human being to decide how a machine behaves and consequently whether the bias should be allowed to flow into the system or not. The challenge is to create a system as neutral as possible without bias, since according to research, machines were originally used because they were believed not to contain subjectivity. However, here too, the basic question arises as to what is neutral; because also in this case each person has a different perception and a different background, which make a definition of the term ‘neutral’, difficult.

Increasing concern has arisen as a result of the identification of some cases where bias in systems has led to discriminatory consequences. The literature research has proven that efforts are currently underway to develop topics such as AI-fairness, AI-responsibility and AI-safety and to establish standards to control bias in algorithmic systems.

The fact that bias is very likely to be transferred to algorithmic systems has led the author of this thesis to create a framework to illustrate how bias can be identified, mitigated or even partially eliminated in the algorithmic system lifecycle. The framework is to be regarded as a basic framework which should be used as guideline along with recommendations to head towards a fundamental system neutrality defined by the author of this thesis. The specification, addition or elimination of aspects and inputs/questions by the project team is necessary, since the context largely determines the form of the framework. Each metamodel element would require a separate investigation to assess the priority of the question/requirement and to ensure useful answers. Such studies would go beyond the scope of this thesis. Future research should include the application of the framework in realistic software project situations such that its added value could be observed, evaluated, validated and subsequently adapted based on the project experiences.
9 OUTLOOK

The author of this thesis believes that the testing for correct application of the framework and compliance with standards should be automated. The reason to automate the evaluation of the framework is that through Human Evaluation, subjective opinions and views continue to be expressed by the testers.

The automation would require the project team to enter the answers to the framework’s questions to a software. A test mechanism would then be implemented to reveal which areas of the framework have been sufficiently or not sufficiently considered and could possibly accommodate bias. In other words, the machine should be intelligent enough to say that due to the framework and the data received the standard is adhered to or not. For instance, let’s assume in the standard it was defined that the training data must have at least 10,000 entries and there need to be at least 300 women as part of test users. This way it would be checked if according to the data provided by the project team, the system meets the requirements to be able to say that the system can offer conclusions and recommendations to the end user.

This approach could be used for the project team that creates and maintains AI-systems to perform tests, but also on another level for the end user to display how reliable the system recommendations are. The end user would then know where the standard was met or could not be met and would thus be better able to decide how reliable the system’s recommendations are and whether they might need to take further measures and clarify matters before execution.

However, to reach this point, there are several aspects that need to be considered. Elements from the framework would have to be worked out on micro-level to define for example what a stakeholder is or how to identify that the test user was indeed of the particular gender. Or, when the question if a team member has received training in ethics can be answered with yes? Is reading about it enough or a course needed or even a certificate? In addition, the fact that people can always make false statements to beautify the result needs to be taken into account. Therefore, mechanisms must be integrated that take into account the truthfulness of the answers.

The author is aware that the idea of automating the framework is a complex undertaking. The validation by application in concrete projects will reveal whether the framework satisfies its purpose and can be used reasonably.
10 REFLECTION

The elaboration of this bachelor thesis has given the author the opportunity to deal, in greater depth, with the topic of bias in algorithmic systems. The matter is highly relevant and will become more and more important in the future. The examination of the subject by the four research questions from different perspectives has proven to the author, the complexity and importance of this topic in today’s society.

Through an intensive examination of the literature, it was possible to gain insights that answered the research questions well. It has to be said, however, that the understanding of AI and methods to identify and mitigate bias is very broad and as a consequence not all aspects could be covered in this thesis that could emerge anywhere at any time.

The result of this thesis serves as an example guideline on how bias can be combated in a project. The author is aware that the developed framework has to be extended and adapted according to the context of the project. However, the framework serves as a good starting point to raise awareness about the topic of bias in project teams.

The author has learned that this topic encompasses many aspects and that combating bias in algorithmic systems requires great cooperation and clear guidelines. Research has demonstrated that such efforts are underway. However, it will be some time before concrete measures are established and widely applied. Until then, it is important to make a contribution step by step and not to close one’s eyes to this subject, but to draw more attention to it.

The author has also become aware that the minimization of bias can only be effective if the perspective is opened as far as possible and areas are considered that would not be immediately recognized as sources of bias at first. In the end, the whole topic is based on human behaviour, which is why it is necessary to start at the very beginning, the human psyche. This means that each human being should be evaluated individually in the respective project context and no universally general conclusions can be drawn about the behaviour of individuals in detail. This involves some effort, but it is inconceivable for the author to reduce bias in any other way.

The topic of bias in algorithmic systems has been recognized by researchers and managers, but in the opinion of the author and based on literature research, it is still not widespread enough. The author assumes that the efforts currently underway will contribute to this topic gaining even more momentum. This is a prerequisite for society to address and successfully reduce this issue.
REFERENCES


MIT Technology Review. (2013). Racism is Poisoning Online Ad Delivery,


LIST OF TABLES

Table 1: Sources of cognitive bias (Benson, 2016) .......................................................... 11
LIST OF FIGURES

Figure 1: Agents interact with environments through sensors and actuators (Russell & Norvig, 2010, p. 35)........................................................................................................................................9
Figure 2: Metamodel. Identifying and mitigating bias in algorithmic System...............28
<table>
<thead>
<tr>
<th><strong>Element</strong></th>
<th><strong>Description/Comments</strong></th>
<th><strong>Yes</strong></th>
<th><strong>No</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project Team</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All project members have had ethical training</td>
<td>Members have a confirmation that they have completed courses or workshops or similar. The minimum requirements to consider this element as fulfilled must be defined in the company.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All project members are aware of the topic of bias that exists in the human decision-making process</td>
<td>Members took part in courses or workshops or similar. The minimum requirements to consider this element as fulfilled must be defined in the company.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All project members know about the fact that human bias can be reflected in an algorithmic system</td>
<td>Members took part in courses or workshops or similar. The minimum requirements to consider this element as fulfilled must be defined in the company.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All project members consider the same attributes and factors as most relevant in the system context.</td>
<td>A workshop is held where each member shares their views. Discrepancies are pointed out and a common understanding is developed. The workshops aim to share views, ideas and openly to reveal conflicts and misunderstandings. Due to cultural and background dissimilarities members might (unconsciously) weight attributes differently.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The project team represents stakeholders of all possible end user groups</td>
<td>Stakeholder analysis comprehensively identifies end user groups with a focus on identifying users who might be disadvantaged through the system outcomes. The stakeholder analysis should also be carried out with a change of perspective, where the worst scenario, i.e. if the system behaves discriminatory, identifies the groups that would be disadvantaged. (See area Project Management)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The project team is a cross-functional team including diversity in ethnicity, gender, culture, education, age and socioeconomic status</td>
<td>The inputs of the same number of men and women, of young and old etc. are included.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The project team has representatives from the public as well as the private sector</td>
<td>Exclusions need to be avoided</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent consultants are included for comparison with competing products</td>
<td>Pre-existing bias in the context of the company’s culture, attitude and values can be revealed. Independent consultants are needed because they are not biased by the companies’ views</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Element</td>
<td>Description/Comments</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>-----</td>
<td>----</td>
</tr>
<tr>
<td><strong>Environment and Context</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All possible end user groups are included in the testing phase</td>
<td>The behaviour of end users can only be reliably recorded if they test directly on the system. Hidden behaviour can thus be detected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All possible end user groups have been evaluated</td>
<td>End user groups’ behaviour is monitored and evaluated from different perspectives (surveys, interviews, recording behaviour, letting them explain what they do and think while testing)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consequences and intentions have been considered</td>
<td>For what and with what intentions was the system created for?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>What is the worst thing that can happen in this algorithm if it starts interacting with others?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Context is faithful to the original source</td>
<td>Does the current context represent the one, for which the system was originally created?</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Constraints</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Business aspect reviewed</td>
<td>Under what circumstances will the system be developed?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scope reviewed</td>
<td>The requirements for the scope of the data set and the diversity are to be determined in the respective project.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technical aspect reviewed</td>
<td>Do technical constraints affect the way the system is designed?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Legal aspect reviewed</td>
<td>Do regulatory/law constraints affect the way the system is designed?</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Input (Datasets)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The data set is fully understood</td>
<td>The meaning of each attribute is understood and its purpose in the system context is clear</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data is transparent</td>
<td>Data must be reliable, accurate and kept up to date</td>
<td></td>
<td></td>
</tr>
<tr>
<td>It is ensured that the data set represents the correct scope (enough data to represent a population or target group)</td>
<td>Enough data and diversity are available</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The requirements for the scope of the data set and the diversity are to be determined in the respective project.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The source of the data is known and verified</td>
<td>Unknown source of the data might lead to that the data is used in a context it was originally not intended to</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The quality of the data is ensured</td>
<td>Data with low quality will cause even worse outputs since AI-systems might reinforce errors in data sets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>It is clarified which attributes can legally be used</td>
<td>Use of illegal attribute leads to a system becoming biased even though the attribute itself is not cause for bias</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Element</td>
<td>Description/Comments</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>-----</td>
<td>----</td>
</tr>
<tr>
<td><strong>Training Data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The training data set is still as representative as the original data set</td>
<td>Adjusting source data to training data can bear exclusion which needs to be prevented</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Added or omitted attributes are carefully chosen and justified</td>
<td>One attribute can influence different areas in a system. Interconnectedness needs to be considered</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Test Data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test data is independent</td>
<td>The system uses test data it has never seen before</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test data is defined</td>
<td>Test scenarios are defined which are designed to detect bias which could be caused by a certain attribute</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test data is reviewed</td>
<td>Tests include omission and addition of attributes to test how system output changes</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Project Management</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project management process includes methods that focus on bias issues</td>
<td>Stakeholder analysis is adjusted for disadvantaged group identification in worst case</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risks concerning bias are assessed and known to each team member</td>
<td>Risk analysis is adjusted for additional focus on bias and worst-case scenarios provoking to bias</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical thinking is promoted and demanded at every stage of the system creation process</td>
<td><strong>Following questions needs to be asked:</strong> How would changes to a data point affect the model’s prediction? Does it perform differently for various groups? For example, historically marginalised people? How diverse is the dataset I am testing my model on? Is the system context the one the system was intended to? Can the outcome/result/system recommendation be justified? How diverse is the dataset I am testing my model on? Does it perform differently for various groups—for example, historically marginalized people? How would changes to a data point affect my model’s prediction?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perspectives are changed continuously to challenge assumptions</td>
<td>Different point of views ensure identification of hidden assumptions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monitoring measures are defined, communicated and applied</td>
<td>End user groups’ behaviour is monitored and evaluated from different perspectives (surveys, interviews, recording behaviour, letting them explain what they do and think while testing)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Auditing measures are defined, communicated and applied</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Project Management

<table>
<thead>
<tr>
<th>Element</th>
<th>Description/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workshops / meetings are set frequently which address upcoming doubts of team members</td>
<td>Critical thinking is continuously fostered in workshops and outside</td>
</tr>
<tr>
<td>Scenario thinking is fostered</td>
<td>-</td>
</tr>
<tr>
<td>Freedom of expression is guaranteed and desired</td>
<td>Every input of any team member can reveal hidden bias</td>
</tr>
</tbody>
</table>

## Hardware

<table>
<thead>
<tr>
<th>Element</th>
<th>Description/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do hardware limitations exist?</td>
<td>-</td>
</tr>
<tr>
<td>Do these limitations influence the system creation process?</td>
<td>-</td>
</tr>
<tr>
<td>Do these limitations influence the system’s functionality in the production environment?</td>
<td>-</td>
</tr>
</tbody>
</table>

## User Interface

<table>
<thead>
<tr>
<th>Element</th>
<th>Description/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visuals aspects are determined appropriately</td>
<td><strong>Text:</strong> the font-style, font-size, font-colour and placement are justified and reflect the intention of the system’s functionality</td>
</tr>
<tr>
<td></td>
<td><strong>Forms/elements</strong> (e.g. boxes containing text or graphics): colour, size and placement are justified and reflect the intention of the system’s functionality</td>
</tr>
<tr>
<td>Does visual result representation (alphabetically or random) make any difference (user always choses the results displayed first?)</td>
<td>-</td>
</tr>
<tr>
<td>Does a change in navigation representation lead the user to favour different results?</td>
<td>-</td>
</tr>
<tr>
<td>Is graphical UI limiting/favouring data over other data?</td>
<td>-</td>
</tr>
<tr>
<td>Is a translation of data/information necessary?</td>
<td>How does the chosen language influence the user’s perception and interpretation in different contexts and circumstances?</td>
</tr>
<tr>
<td>Do the information and results become distorted through the application of translation?</td>
<td>How is the translation interpreted by the end users?</td>
</tr>
<tr>
<td>The system features mentioned above are changed and end users are monitored on the above elements once more to see how their behaviour changes</td>
<td>Several features may need to be changed various times to reveal hidden assumptions of end users</td>
</tr>
<tr>
<td>Element</td>
<td>Description/Comments</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------</td>
</tr>
<tr>
<td><strong>Programming</strong></td>
<td></td>
</tr>
<tr>
<td>Code reviews take place</td>
<td>Measures that aim to understand adapted or reused code fully</td>
</tr>
<tr>
<td>Independent code audits are conducted</td>
<td>Independent audits foster considering the code from a different point of view and reveal unconscious assumptions</td>
</tr>
<tr>
<td>Possible user behaviour is analysed beforehand to keep a learning system from adopting discriminatory behaviour</td>
<td>Thinking outside the box is fostered especially considering word and language usage in the system context. The system can handle discriminatory user behaviour</td>
</tr>
<tr>
<td><strong>Deliberate Bias</strong></td>
<td></td>
</tr>
<tr>
<td>Bias is identified and categorized</td>
<td>Are the identified biases considered as good, neutral or bad ones? Is there any bias which was implemented on purpose to mitigate other?</td>
</tr>
<tr>
<td>It is ensured that all the identified biases are monitored during the whole system creation process</td>
<td>Bias needs to be tracked and changes identified as well as recorded throughout every stage of the project</td>
</tr>
<tr>
<td><strong>Documentation</strong></td>
<td></td>
</tr>
<tr>
<td>Are the relevant information present?</td>
<td>Traceability, justification and business continuity are ensured</td>
</tr>
<tr>
<td>Is the documentation comprehensible?</td>
<td>The language may only contain such a high degree of complexity and technical language that every project member understands it. Prevention of misunderstandings is ensured</td>
</tr>
<tr>
<td>Has the documentation been reviewed and approved?</td>
<td>The documentation needs to be reviewed by several project members and stakeholders</td>
</tr>
</tbody>
</table>