TU WIEN Informatics

Erweiterung von CRISP-DM zur systematischen Abdeckung regulatorischer Anforderungen zu Bias in AI Systemen

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Business Informatics

eingereicht von

Fabian Eberle, BSc
Matrikelnummer 01328887

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Andreas Rauber

Wien, 2. Oktober 2023

Fabian Eberle	Andreas Rauber

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Systematic extension of CRISP-DM by structured mapping of emerging regulatory requirements on bias in AI

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Fabian Eberle, BSc
Registration Number 01328887

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Andreas Rauber

Vienna, 2nd October, 2023

Fabian Eberle	Andreas Rauber

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Fabian Eberle, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 2. Oktober 2023

Fabian Eberle

Danksagung

Denn auch Diplomarbeiten brauchen ein Support-Team:

Diese Arbeit zu schreiben fühlte sich manchmal an, als würde man mit verbundenen Augen durch ein Labyrinth gehen und dabei mit flammenden Schwertern jonglieren. Aber wie bei jeder großen Zirkusnummer hätte ich das nicht allein schaffen können! Zunächst einmal möchte ich mich bei allen WissenschaftlerInnen, IngenieurInnen und EntwicklerInnen bedanken, die viele wunderbare Technologien, Tools, Plattformen und Libraries entwickelt haben, um die Arbeit und Forschung zu erleichtern! Meinem treuen Macbook: Danke, dass du während der nächtlichen Schreiborgien nie abgestürzt bist. Du hast mich in schlimmen Momenten erlebt, dich nie beschwert, selbst wenn ich dich angeschrien habe und mich dabei nie verurteilt. An die Kaffeebohnen und die Hersteller des koffeinhaltigen Elixiers: Ich schulde euch etwas. Meinem Betreuer, Andi: Du hast endlosen Fragen, Überarbeitungen und gelegentlche Verzögerungen hingenommen. Deine Geduld verdient einen Nobelpreis. An meine Familie: Dafür, dass ihr so tut, als würdet ihr mein Diplomarbeitsthema verstehen, und dafür, dass ihr nicht ständig gefragt habt, wann ich meine Diplomarbeit endlich fertigstellen werde. Eure endlose Unterstützung bedeutet mir sehr viel! An mein soziales Umfeld: Es tut mir leid, dass ich euch eine Zeit lang nicht gemeldet habe, aber ich verspreche, dass ich mich mit einer Feier, die in die Geschichte eingehen wird, revangieren werde.

Und schließlich an mich selbst: dafür, dass ich diese Achterbahnfahrt der Zweifel, Koffeinüberdosen und die nächtlichen Sessions überstanden habe, dafür, dass ich meinen inneren Perfektionisten herausgefordert habe, indem ich Absätze immer wieder überdacht und umgeschrieben habe und für meine gelegentlich fragwürdigen Zeitmanagementfähigkeiten zur Einhaltung von Deadlines. Zusammenfassend lässt sich sagen, dass diese Arbeit ohne diese Eigenschaften und ihre besonderen Beiträge zu meiner akademischen Reise nicht möglich gewesen wäre.

Abschließend möchte ich mich auch bei den glücklichen Zufällen in meinem Leben bedanken, die mich schließlich dazu gebracht und motiviert haben, diese Arbeit noch vor meiner Pensionierung abzuschließen.

Und nun viel Freude beim Lesen, aber beeilen Sie sich, denn dieses Thema kann schnell überholt sein!

Acknowledgements

Because even Diploma theses need a support team:

Writing this thesis was like navigating a maze blindfolded while juggling flaming swords. But, like any great circus act, I couldn't have done it alone! Firstly, to all the scientists, engineers and developers who did amazing work in creating wonderful technology, tools, platforms and libraries to make research easier! To my trusty Macbook: thanks for never crashing during those late-night writing binges. You've seen me at my worst, never complained about yelling at you and yet you never judged. To the coffee beans and the makers of the caffeinated elixir, I owe you my sanity. To my supervisor, Andi: You tolerated my endless questions, revisions and even delays. Your patience deserves a Nobel Prize. To my family: for pretending to understand my thesis topic and for not asking consistently when I finally will finish my thesis. Your endless support means the world to me! To my social life: I'm sorry I had to ghost you for a while, but I promise I'll make it up with a post-thesis celebration that will go down in history. And finally, to myself: for surviving this rollercoaster of doubt, caffeine overdoses, and late-night epiphanies, for challenging my inner perfectionist, overthinking and rewriting paragraphs over and over until they're perfect and for my occasionally questionable time management skills, exploring creative ways to meet deadlines. In conclusion, this thesis wouldn't have been possible without this cast of characters and their peculiar contributions to my academic journey.

In the end, I also want to thank the happy little coincidences that are occurring in my life which finally led and motivated me to finish this thesis before retirement.

And now, enjoy reading but hurry up, this topic can be outdated quite quickly!

Kurzfassung

Durch die breite Anwendung von Künstlicher Intelligenz (KI) ziehen Beispiele, die in Bezug auf Voreingenommenheit, Fairness und Diskriminierung schief gelaufen sind, die öffentliche Aufmerksamkeit auf sich und werfen damit ein schlechtes Licht auf KI-Systeme. Ein bekanntes Beispiel dafür ist COMPAS, welches die Wahrscheinlichkeit der Rückfälligkeit von Straftätern vorhersagen sollte und dabei (unwissentlich) gegen die Hautfarbe diskriminierte. Obwohl es in den USA angewandt wurde, ziehen ähnliche Vorfälle in Europa die Aufmerksamkeit der EU-Gesetzgeber auf sich, die den Markt für KI-Systeme regulieren und im April 2021 einen entsprechenden Vorschlag vorgelegt haben. Die Regulierung zielt vor allem darauf ab, potenziellen Schaden an und gegen Menschen, wie etwa Diskriminierung oder Bedrohungen gegen Leib und Leben zu verhindern und steht damit in einem starken Zusammenhang mit Voreingenommenheit bzw. Bias. Bei der Untersuchung bestehender Prozessmodelle wie Cross Industry Standard Process Model for Data Mining (CRISP-DM) im Hinblick auf die Eignung für aufkommende Regulierungen haben wir ein Mangel an Leitlinien und Handlungsanweisungen für die Aufdeckung und Behandlung von Voreingenommenheit bzw. Bias während des Entwicklungsprozesses als Lücke festgestellt.

In dieser Arbeit wird eine Vielzahl von Verzerrungsarten bzw. Bias-Typen in der Literatur identifiziert, indem eine breite Literaturrecherche durchgeführt wird, klassifiziert nach der Entstehungsursache bzw. wie sie aufgedeckt werden können. Diese identifizierten Bias-Typen bilden die Grundlage für das Kernstück dieser Arbeit, die Erstellung eines systematischen Mappings der identifizierten Bias-Typen entsprechend der zugehörigen Aufgaben in CRISP-DM. Darüber hinaus wird dieses Mapping mit den obligatorischen Anforderungen aus dem European Union Artificial Intelligence Act (AI-Act) verglichen und mit dedizierten Schritten erweitert, um einen Ansatz zur Konformität mit der Regulierung vorzuschlagen.

Abstract

Through the broad utilization of Artificial Intelligence (AI), examples that have gone wrong according to bias, fairness and discrimination, have attracted public attention and shed a bad light on AI systems. One famous example is COMPAS, which aimed to predict the probability of criminals reoffending and it turned out, that COMPAS (unwittingly) discriminated against skin colour. Although it has been applied in the U.S., similar incidents in Europe drew the attention of European Union (EU) legislators, motivating them to regulate the market for AI systems and presented a proposal to do so in April 2021. The regulation especially targets preventing harmful outcomes to humans, such as discrimination, which is stated in a strong context to bias. Examining existing process models such as the Cross Industry Standard Process Model for Data Mining (CRISP-DM) according to fitness for emerging regulatory obligations unveiled a lack of guidance for unveiling and treating bias during the development process which was identified as a gap.

In this work, we identify a broad variety of bias types in the literature by performing a mapping study providing explanations on different bias types, where they emerge and how they can be unveiled. The identified bias types form the basis for the core of this work, the provision of a mapping of the identified biases according to the associated tasks in CRISP-DM. Moreover, the mapping is compared with requirements from the European Union Artificial Intelligence Act (AI-Act) and enriched with dedicated steps to propose an approach to reaching compliance with the regulation.

Contents

CHAPTER 1

Introduction

When speaking about bias in Artificial Intelligence (AI) systems, several definitions of bias (lit.[1] prejudice, tendency, preference) can be observed, heavily depending on the context, the (academic) background and viewpoint:

- „*... systematic difference in treatment of certain objects, people, or groups in comparison to others, where "treatment" is any kind of action, including perception, observation, representation, prediction or decision*“ [58]
- „*... any basis for choosing one generalization over another, other than strict consistency with the instances*“ [49]
- „*... one or a set of extraneous protected variables that distort the relationship between the input (independent) and output (dependent) variables and hence lead to erroneous conclusions*“ [2]
- „*... the difference in the underlying distribution of the model learning outcome with respect to a certain group(s) influenced by their affiliation to the specific group. The group could be gender, race, age or any other protected attribute*“ [3]
- „*... deviation from a standard*“ [32]

Besides these different definitions of the term "bias" itself, it is also important to be aware of different types of biases, their origin (overall and within the development process) and steps to deal with certain types as well. Although the scientific community identifies and points out different types of bias as Mehrabi et al. did in their survey paper [76] by identifying 19 different types, there is little (systematic) guidance on how biases can be unveiled, treated and observed. Bias in data has different facets and effects which can be

[1] https://www.dict.cc, accessed 25/04/23

one reason for unfairness in the resulting system [76]. Bias can be an indicator or lead to unfair and in the worst case even discriminatory outputs. Although bias does not automatically imply unfairness, it is a result of human behaviour reflected in the data [3]. Fairness generally means not favouring one item/individual/group/group member over another, based on an attribute (-value) [76] which will be further described in Section 2.1.1. Uncovering and identifying bias already is and will be an even more essential step in the data mining and machine learning process since regulatory obligations like the European Union Artificial Intelligence Act (AI-Act) explicitly require investigations for bias which, if not treated properly, is penalized by a severe fine in case of violation.

1.1 Emerging Legal Regulations

In April 2021, the European Commission (EC) published a proposal aiming to regulate placing on the market, putting into service and the use of artificial intelligence systems [37, p.38, Article 1 (a)]. The term "artificial intelligence systems" (further noted as AI systems) covers supervised-, unsupervised-, reinforcement- and deep-learning, logic- and knowledge-based approaches as well as statistical approaches i.e. all systems that interact with humans, provide information or recommendations and may influence a human's environment or decision [37, p.39, Article 3] and Annex I [36, p.1]). The *Annex* is an appendix, which comprises definitions and taxonomics that enable updates in practice posterior. The regulation differentiates between four main types of system risks[2]: *unacceptable-*, *high-*, *limited-* and *no-risk* systems with different obligations. These obligations comprise a variety of required assessments, processes and documents from the early beginning, towards the operation and monitoring of AI systems. Although "potential biases", "biased results", "bias monitoring" and "biased outputs" are literally mentioned in the proposal, it lacks definitions, distinctions and courses of action dealing with bias [37].

1.2 Technical and Development Shortcomings

When performing data mining projects to develop complex AI systems, specific process models like KDD [40], CRISP-DM [27], SEMMA[3] and ASUM-DM[4] provide structure through the whole development process; more general to schedule tasks and specify methods to use [73]. Although there are scientific investigations and comparisons on suitability, applicability, fitness for purpose, strengths and weaknesses according to team-, project-, data- and information management [74], they are missing guidelines for identifying/unveiling/avoiding bias(es). Since interest in bias and fairness in AI systems within the scientific as well as in the non-scientific domain is still unabated, the community is working on libraries and tools to mitigate certain biases but lacks generic and systematic approaches.

[2] https://digital-strategy.ec.europa.eu, accessed 07/04/23
[3] https://tinyurl.com/ms3d4but, accessed 22/09/23
[4] https://public.dhe.ibm.com/software/data/sw-library/services/ASUM.pdf, accessed 22/09/23

1.3 Problem Statement and Motivation

The broad adoption and commercialization of AI technologies are associated with examples of how things can go wrong: COMPAS, as a system for assessing the risk of recidivism, applied in American criminal justice, has shown fatal issues; the false-positive rate was twice as high for black people and the false-negative rate was twice as high for white people. Amazons AI hiring tool favoured men for technical jobs[5]. This led to an increased interest in bias in AI systems and motivated researchers to perform investigations on different types of biases [76, 77, 3], related ethical considerations [83, 82] as well as related to fairness [79, 26, 25]. The ongoing plans and discussions to regulate the domain of AI technologies by the AI-Act [37] (as well as in the U.S.[6] and China[7]) force producers as well as users of AI systems to compliant adoption.

Considering **process models** as a guideline when performing data mining projects, there are some to choose from: KDD (Selection, Pre-Processing, Transformation, Data-Mining, Interpretation/Evaluation) introduced by Fayyad, Piatetsky-Shapiro, and Smyth [40] in 1996 and SEMMA[8] (Sample, Explore, Modify, Model, Assess) introduced by SAS Institute. Both seem to be very similar, are interactive and iterative process models and prepare the base for CRISP-DM (CRoss Industry Standard Process for Data Mining) [27], which is considered as a practical "implementation" of KDD and SEMMA comprising *Business Understanding* in the beginning and *Deployment* in the end (Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment) [8, 27]. Microsoft TDSP[9], based on CRISP-DM, is considered to be a proprietary process model serving Microsoft's services and products. Considering the needs for remotely performed data mining projects, RAMSYS[10], based on CRISP-DM, was introduced to give a clearer guideline according to roles and responsibilities [74]. According to a study from 2014, 43% still refer to CRISP-DM as methodology [74] and although it was developed for data mining projects, it is also successful in the data science domain [43].

Considering **bias detection**, there are different approaches in research as well in the commercial sector for different types of bias detection and mitigation. There are statistical approaches [19, 33, 3, 112, 88] which propose to investigate correlations, distributions etc. but mostly focus on one specific type of bias or issue. There are also tools and libraries released such as Amazon Sage Maker[11], IBM AIF 360[12], Aequitas[13] and Themis-ML[14] to name the most prominent ones. The interest in this field is tremendous and therefore the number of introduced tools and techniques in research and commercial areas is still

[5] https://www.theguardian.com/, accessed 11/04/23
[6] https://www.whitehouse.gov/, accessed 11/04/23
[7] https://digichina.stanford.edu/, accessed 11/04/23
[8] https://tinyurl.com/ms3d4but, accessed 22/09/23
[9] https://learn.microsoft.com, accessed 12/04/23
[10] https://www.researchgate.net/, accessed 12/04/23
[11] https://aws.amazon.com/, accessed 12/04/23
[12] https://github.com/Trusted-AI/AIF360, accessed 12/04/23
[13] https://dssg.github.io/aequitas/, accessed 12/04/23
[14] https://github.com, accessed 12/04/23

growing. Research and commercial interests are converging such as examples of Amazon or IBM showed. We believe that this gap between approaches in academia (statistical studies) and the commercial sector (introduction of tools to detect and mitigate bias that relies on a different "black box") can be closed in the context of this research.

1.4 Research Questions

As bias, which leads to unfairness or discriminatory outputs, becomes more and more an issue in the context of developing and applying AI systems, regulators (on behalf of the EU Commission) were urged to create a legal framework, frame the problem, provide red lines, and challenge AI systems. Actors creating or deploying AI systems need to perform a detailed analysis of their system starting from the purpose of the system, assessing risks, technical documentation, transparency- and human oversight measures through to regular checks during operation. Unfortunately, established and widely used process models lack the capabilities to utilize these requirements in a proper way which forces stakeholders (such as organizations, analysts etc.) to rely on best practice methods in their domain, develop new methods and techniques, hire dedicated experts or consult external experts. Combining these aspects we introduce the overall research question (RQ):

RQ: Which extensions in process models are necessary in order to help analysts uncover bias?

RQ1: Which types of bias are described in the literature, where do they manifest and how (according to the method) can bias be unveiled?

At first, we investigate the definition of bias according to its framed meaning (e.g. mathematical, societal/cultural, ethical etc. [26, p. 18]) as well as different types. Then we will have a look at where bias manifests itself (e.g. the different involved stakeholders in the development process, societal factors, in the data etc.). Once we identified the definition and manner of manifestation of the bias(es), we will gather advisory for unveiling the specific types.

RQ2: At which steps in the process model can bias be identified and which steps are necessary to do so?

Once we define the bias(es), their manifestation and potential ways to deal with it, we are interested in turning this information into an applicable way by a systematic mapping between bias and a corresponding phase and task in a process model. The selected process model for this mapping should therefore be considered as a widely adopted, widely known and easy to adapt for a broad range of developer- and project teams. This would enable analysts from different domains to methodologically search for and discover a broad range of different bias(-es) in the whole development process.

RQ3: To what extent can the additional process steps introduced satisfy and address the requirements set forth in the articles of the EU AI-Act touching on bias-relevant aspects?

As already stated within the problem statement, the EU AI-Act does not mention a specific type of bias that has to be taken aware of, but since a bias has different facets, sources and effects (society, data and learning [54]), it is necessary to assess the regulatory aspects according to bias(es). Therefore, we will investigate the EU AI-Act according to the proposed articles specifically to bias relevant aspects with the course of action and perform a systematic mapping between requirements in the regulation and the corresponding tasks in a process model (according to the requirements stated in RQ2).

1.5 Research Method and Methodological Approach

- RQ1: Tabular/clustered overview of bias types and definitions (according to causative agent, domain or issue), its manifestation (mathematical or domain knowledge) and ways to deal with bias.
- RQ2: Extension of a state-of-the-art process model with additional steps to ensure bias detection (RQ1) as far as possible.
- RQ3: Systematic mapping between identified bias-relevant aspects in the EU AI-Act and the extended process model ensuring bias detection (RQ2).

The combination of the three research questions aims to support analysts to (i) understand the definition and distinction of bias; (ii) search for, uncover and handle bias(es) in a systematic way in the development process of AI systems; (iii) provide guidance on regulatory requirements as well as considerations and implications for the system's development along the EU AI-Act.

RQ1 will be conducted as a *systematic mapping study* (also known as a scoping study)[64] to gather state-of-the-art research and picture the broad variety [84, 64, 6] of definitions and types of biases in a qualitative way. Dedicated clustering properties will be conducted according to the findings from the literature review; we are supposed to consider the source, distinguishing marks, involved actors and methods of discovery. More specifically, we will search for "bias", "AI", "data" and "data mining", focusing on survey papers and their referred sources. The starting point will be research databases (Arxiv, ACM Digital Library and IEEExplore) as well as grey literature (such as respective domain-specific sources and commercially published papers), in case of meaningful, thoughtful and of course reliable contributions. Building upon this review, clustering and mapping based on structural properties will be performed.

As we aim to extend an existing process model, bridging the gap to emerging legal requirements, RQ2 will be conducted as constructive research (design science). The evaluation will be performed in a descriptive way considering informed arguments to demonstrate the artefact's utility on the one hand and enrich them with detailed scenarios according to Hevner et al. [55] for evaluation.

RQ3 will be conducted as a systematic mapping, investigating the fitness for purpose according to legal requirements [84]. Therefore mapping studies are a suitable method in different fields such as Informatics. As part of the method, we perform a mapping of bias-relevant aspects and the fitness for purpose according to the utilization of our extended process model considered in RQ2.

1.6 Reserach Contribution

This interdisciplinary work incorporates typical aspects from the business informatics domain such as process modelling speaking in the context of CRISP-DM with a strong focus on data science methodologies, techniques and tasks which, in our context, are defined to be the aspect of bias discovery and handling in the AI development process. Moreover, this work considers legal obligations and the incorporation of their requirements into process models in a technical context as previously described.

In more detail, the focus is on (i) mapping emerging bias types according to related sections in CRISP-DM, (ii) providing guidance on their emergence and discovery and potential tools or techniques for dealing with them and considering (iii) bias-related aspects of emerging regulatory requirements (AI-Act) in one single process model ensuring compliance with the obligations.

The **resulting artefact**, which is an enriched version of CRISP-DM (RQ2), shall enable analysts and practitioners within the AI development process to understand, interpret, uncover and deal with arising biases (RQ1) as well as ensure compliance to the AI-Act focussing on bias-related aspects (RQ3).

Naming the **delimitations of this work**, the enriched process model does provide a guideline among several bias types, techniques and aspects but it is not a guarantee to discover and handle all biases because of a myriad of relations and potential side effects such as different data types, different modelling techniques or different use cases, domains or contexts. Moreover, due to the technical context of this thesis, we cannot provide a guarantee of full compliance with the requirements stated AI-Act in a legal manner, since the articles stated in the regulation incorporate many dependencies to (product) liability and to several national- and international legislations which would require profound legal expertise.

1.7 Structure of the Thesis

The thesis is structured as follows:

Chapter 2 Related Work establishes a common understanding of bias, assessing different notions and contexts. In the context of bias, protected attributes such as personal identifiers and proxy variables, which are identifiers for protected attributes, are introduced according to the defined understanding of bias. We describe relations and delimitations to fairness resp. discrimination followed by a relation to ethics. We also

provide an introduction to fairness metrics and strategies for mitigating bias. Furthermore, we introduce the European Union Artificial Intelligence Act (AI-Act) and provide its general impact which will be analyzed in Chapter 4. At the end of this chapter, we state process models focussing on CRISP-DM followed by ongoing international standardization in AI process models.

Chapter 3 Assessing Bias in the Literature starts by defining two example use cases which allow a considerate reflection of specific bias types according to human- and non-human related data. Following the mapping study of its particularities and definition of boundaries, we will exhaustively present discovered bias types assessing their definition, explanations, examples as well as emergence within CRISP-DM. At the end of this chapter, an overview of bias types according to the mapped phases in CRISP-DM is presented, which is considered the first core part of this thesis.

Chapter 4 Legal Implications on CRISP-DM describes the AI-Act's Articles 9-15 compulsory for high-risk systems, their mapping to the occurring phases in CRISP-DM, as well as additional articles we believe, are very important in the context of AI development and this thesis. We also highlight the article's updates between the initially proposed version of the European Commission and the amendments made by the European Parliament. The aim of this chapter is to assess the readiness of CRISP-DM according to the AI-Act's requirements which represents the second core part of this thesis.

Chapter 5 Empowering CRISP-DM unites findings from Chapter 3 and Chapter 4 via detailed assessment of the tasks defined in CRISP-DM and the finding gathered in previous chapters. Therefore provided tasks in CRISP-DM are listed according to the defined steps and enriched in the context of bias types in Chapter 3, their methods of discovery and potential strategies for bias handling or mitigation. Additionally, we provide a similar assessment in the context of the AI-Act by proposing additional tasks to incorporate and if necessary align CRISP-DM's outcomes to reach compliance with the requirements stated in the AI-Act.

Chapter 6 Conclusion and Future Work summarizes the main parts of this work, presenting findings and the contribution of the thesis according to the stated research questions. Moreover, the limitations as well as potential future work will be outlined.

CHAPTER 2

Related Work

In this chapter, we will provide a comprehensive overview of related work applied to this thesis. As shown in the introduction, the (research) community has not agreed yet on a common understanding of bias nor fairness or standardized ways to deal with related issues because of different dependent factors such as domain, context, data type and -quality or modelling technique. In case of diverging interpretations or definitions, we argue and define the respective parts in the context of this thesis providing a stringent and clear utilization which is stated accordingly.

2.1 Assessing Bias

As stated in the introduction, "bias" has different definitions, facets and dependencies on the context as well as on the data which requires a separate and dedicated reflection on the term itself. Since bias can lead to unfairness [77], the literature interchangeably uses the terms "bias" and "unfairness" [29] which we agree with Alelyani [3] who claims that algorithms are not intrinsically unfair but rather depend on historical issues in the data. Therefore bias can also be **desired** (intentionally) for correct functionality (e.g. alarm systems – where distinct classes should be sampled higher) or **undesired** (unintentionally) which can negatively affect performance, be problematic and lead to unfairness which is often related to protected attributes [54, 101] (described in Section 2.1) e.g. favouring men over women in men-dominated professional positions [11]. Balayn et al. [11] consider bias to be a "statistical statement on class distributions, and it relies on the human judgment if a given bias is indeed problematic or not; some biases can be non-problematic or irrelevant, while others would require intermediate intervention" while we strongly support this definition, it is not satisfying in practice.

Since bias can emerge during data sampling from the real world, training data, loss function, model architecture and evaluation [70], bias does not automatically imply unfairness. We aim for a rather general but also precise definition of bias which is

introduced by the International Organization for Standardization (ISO) which is defined as:

> *"... a systematic difference in treatment of certain objects, people, or groups in comparison to others where "treatment" is any kind of action, including perception, observation, representation, prediction or decision."* [58]

This definition includes human- as well as non-human-related aspects, it covers potential issues or different treatments incorporated into the data and it does not apriori imply to unfairness. Another important aspect of why we further refer to this accurate definition is to align with ongoing developments in standardizing AI systems by the ISO. Influencing factors such as protected attributes (Section 2.1) and proxy variables (Section 2.1), considerations to fairness and discriminatory effects (Section 2.1.1) as well as ethical aspects (Section 2.1.3) to consider ways of defining a "fair outcome", is considered in the following sections.

Protected Attributes

Implementing fairness includes the consideration of protected (or sensitive) attributes on (un-)privileged groups or individuals whose classification outcome systematically differs on the attribute value, which is mostly based on socio-cultural aspects such as gender, age or race [26]. Protected attributes can causally affect the outcome [72]. To identify protected attributes, the respective legal obligations in the context of anti-discrimination law have to be assessed e.g. the European Union[1] prohibits discrimination based on **sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation**.

Just removing protected attributes does not solve the fairness issue, because correlated features within the data might be utilized for them which are known as "proxy variables" [81].

Proxy Variables

A proxy variable identifies another variable's value indirectly, which could be a single attribute or even a set of attributes i.e. one or more variables that encode the protected attribute with a substantial degree of accuracy. The same issue could arise with proxy groups [26]. This means that removing a sensitive attribute does not solve the problem, it is rather likely to be hidden. As shown in Figure 2.1, demographic information can be a proxy for race since there might be differences in house rents, income etc. This is important especially if the AI system affects certain groups or individuals to avoid just predicting the outcome based on the proxy which can lead to proxy discrimination.

[1] *https://tinyurl.com/4xtbsk4u*, accessed 23/09/23

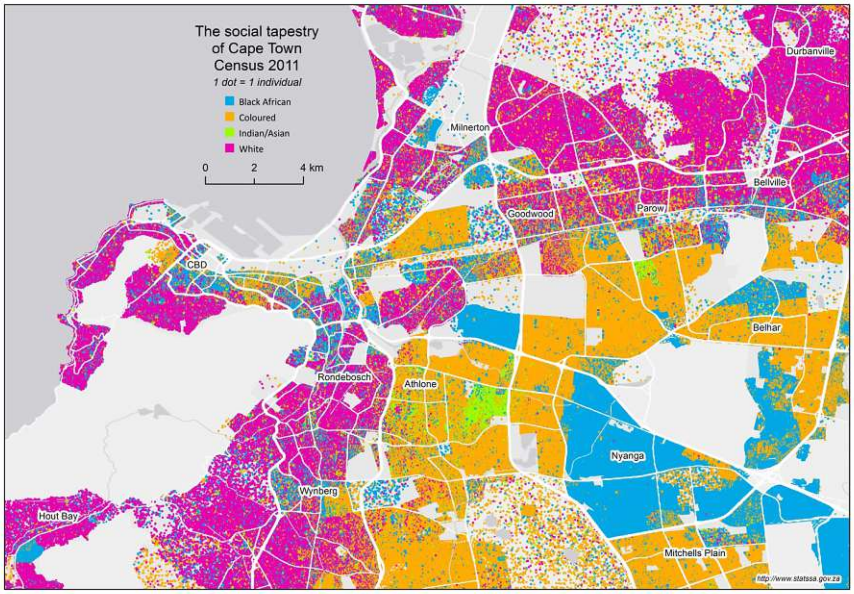

Figure 2.1: Demographic information (e.g. zip-code) can be a proxy for race, wealth etc.
Source: *https* : *//www.statssa.gov.za/?p* = 7678, accessed 25/08/23

2.1.1 Fairness and Discrimination

In the domain of AI the awareness of unconscious bias which can lead to unfair outcomes as well as discrimination, has risen due to the broad adoption of AI systems and the associated prominent examples gone wrong as well. Determining "fair outcomes" requires an understanding of which outcome is considered to be **fair** resp. **unfair**, which further has to be transformed into a measurable representation, by applying an appropriate fairness metric. Although over 20 different definitions of fairness have appeared in the computer science literature [34, 81, 15, 106], there is hardly a clear and common agreement on which definition to apply in each situation [106, 94]. In Section 2.1.1 (Fairness and Discrimination) we give a short overview of the different notions and definitions of fairness. In general, there are different levels of how fairness can be defined when it comes to the subject: (i) individual fairness (same outcome for similar individuals); and (ii) group/subgroup fairness (treating different groups equally) [26]. Depending on the use case one has to decide whether to go for an approach where individuals or members of the same (sub)group should be treated the same. This decision about defining the type of fairness is the first important part and determines the next step of choosing an appropriate metric which is introduced in Section 2.1.1 as well. Yet knowing the different levels of fairness, a common definition of fairness is still missing. As stated in the introduction, Mehrabi et al. interpret fairness as not favouring one item/individual/group/group

member over another based on an attribute (-value). Using the definition of bias stated by ISO [58], we invert the interpretation and understand unfairness in a way where the definition of "treatment" is extended to the above-introduced detail i.e. *unfairness is to favour subjects based on an attribute(-value).* This interpretation aligns with Dwork et al. who state that "any two individuals who are similar with respect to a particular task should be classified similarly" [34].

Discrimination

When speaking about fairness, also unfair cases have to be considered which can lead to discrimination in the worst case. Therefore we refer to the distinction from Mehrabi et al. [76] who differentiate between **direct** and **indirect** discrimination:

- *Direct discrimination* "... happens when protected attributes of individuals explicitly result in non-favourable outcomes toward them" [113].
- *Indirect discrimination* happens if "... individuals appear to be treated based on seemingly neutral and non-protected attributes" [76] e.g. proxy-variables or socio-cultural aspects.

Moreover, the scientific literature distinguishes between (i) *systemic/structural discrimination* which refers to socio-cultural aspects that are incorporated into organizations and therefore incline discrimination against certain (sub-)groups of the population; and (ii) *statistical discrimination* where average group statistics are used to assess an individual of a certain (sub-) group [76, 15].

Besides the literature and developments in computer science providers of AI systems have to deal with legal obligations according to discrimination. According to respective legal frame conditions, affected people do have the right to sue providers of AI systems in case they are convinced to be discriminated against their protected attribute(s), which could lead to tremendous damage to public image or financial loss. Therefore we claim to carefully check the respective (inter-)nationally applicable legal frame conditions such as discriminatory legislations, GDPR as well as the AI-Act which is examined in more detail in Section 2.2.

When it comes to discrimination, one has to consider different aspects aside from sensitive attributes since the boundaries can seem to be very close to business objectives such as explained by Hassani: "if we consider that a customer with a lower revenue is riskier for a bank than a customer with a higher revenue, then correcting the biases by ensuring that social biases are not captured in the data could lead financial institutions to take higher risks" [53].

Fairness Metrics

In Section 2.1.1 (Fairness and Discrimination) we refer to two levels of fairness according to the subject (individual-, group-, and sub-group fairness), which can be utilized in several ways, which is heavily dependent on the use case, the aim and the context of the applied system. However, a recommendation of suitable metrics is out of scope and has to be challenged individually. Caton and Haas [26] provide a comprehensive overview and explanation of different fairness metrics, which can be categorized as follows:

- *Parity-based* (group): considering predicted positive-/negative-rates across groups (if labelled data is available, this can be performed right away) [33, 47, 15].
- *Confusion-based* (group): considering variants of the **predictive positive rate** utilizing different ratios of True-Positive-Rate, True-Negative-Rate, False-Positive-Rate and False-Negative-Rate [33, 47, 52].
- *Calibration-based* (group): equalizing the **predicted probability** (score) among groups [26, 47, 15].
- *Score-based* (group): equalizing the **expected predicted score** for the positive and negative class [26, 47, 15].
- *Individual and Counterfactual fairness* (individual): considering the output for each individual [34].

Regardless of the applied metric, it has to be tested thoroughly to ensure "acceptable" performance among the desired target group, since there is a trade-off between fairness and accuracy [26] especially if there is a high correlation between the class labels and the sensitive attributes, which can lead to unacceptable performance according to the business objectives [112]. Although the combination of multiple protected groups for multiple fairness criteria is shown to be very difficult [79, 26], there are approaches to interpolate between individual- and group fairness metrics [63]. Moreover, it is unlikely that one chosen fairness metric is sufficient across contexts [91]. It is important to understand the notion, boundaries and effects of the chosen metric since we have to keep in mind, that algorithmic systems might make decisions that affect people's lives [76] and therefore acting roles have to take responsibility. Moreover Wachter, Mittelstadt, and Russell [108] claims that statistical methods within technical literature [106] can not conceptualize European non-discrimination law since it does not provide clear defined roles and thresholds, such as required for automated systems.

2.1.2 Bias Mitigation

If a certain bias type is discovered during data analysis, the urge for mitigation (debiasing) arises especially when it comes to fairness criteria along Protected Attributes. Basically, there are three categories of approaches to deal with and mitigate bias:

1. **Pre-processing** techniques transform data upfront before the actual modelling [76, 50, 61] i.e. to train a model on "repaired" data [26] which is a very flexible way according to G. Harris [24]. Different pre-processing approaches are compared in [44] according to performance and fairness in different datasets; which is the most effective stage to address biases [45].

2. **In-processing** techniques focus on modifying the learning algorithms during the training process [76] to find a balance between multiple model objectives [26]; according to G. Harris [45], this is the most efficient stage to handle biases.

3. **Post-processing** techniques aim to assess the model after training without considering any action before which can be assessed by a holdout set which was not involved in the training [76]. Subsequently performing transformations to the model output is seen as the most flexible approach and also suitable for black-box scenarios if the model is not accessible [26]; this may be the ideal point in time to handle biases according to G. Harris [45].

Caton and Haas [26] provide a comprehensive set of research literature resources on different approaches among pre-, in- and post-processing methods in a well-described manner including benefits and drawbacks. While the combination of different fairness metrics seems to be very difficult as described in Section 2.1.1, there are approaches to combine different approaches [71, 45]. In general, which approach to choose depends on the notion of fairness and the context of the application [11, 26].

Speaking about bias mitigation, debiasing "depends, among other factors, on the analyst's background knowledge, quality of reasoning skills and statistical sophistication" [66]. Whether bias should be mitigated depends on different aspects which cannot be generalized. As stated in Section 1.6, bias mitigation per se is not an objective of this work, but we will refer to sources discovered during the mapping study in Chapter 3 within the respective task in CRISP-DM in Chapter 5.

2.1.3 Ethics by Design

Ethics as a philosophic discipline is "the study of what is morally right and wrong or a set of beliefs about what is morally right or wrong" and dealing with questions like "*What is a good action?*", "*What is justice?*" or "*What is the good life?*" [56]. As stated within the definition of ethics, morale is often used in combination, whereas morale is "relating to the standards of good or bad behaviour, fairness, honesty etc. that each person believes in, rather than to laws" which basically focus on "concrete factual patterns of behaviour, the customs, and conventions that can be found in specific cultures, groups, or individuals at a certain time" [56].

Among different (academic) subfields of ethics, we focus on "applied ethics" which considers real-life situations concerning particular and isolated situations challenging what we are obliged, permitted or forbidden to do [56]. Thus "ethics by design" focuses on

the integration of human values into the design and development of technological systems which requires a rich and interdisciplinary set of stakeholders and experts [80]. Therefore theoretical and methodological frameworks such as "value-sensitive-design[2]" or "quality assessment" are supposed to leverage the process of pinning down abstract questions to concrete guidelines. Such guidelines are an indispensable intermediate baseline for further steps in CRISP-DM i.e. from business understanding, development and coding. Therefore we need a comprehensive approach for AI ethics [20].

Ethics by design is not restricted to the development process, it is rather an integral and continuously applied process of evaluation between the defined values, set of rules and norms on the one hand and the way a system operates [80]. Since machine learning and AI in general are based on discovering (statistical) patterns out of large datasets, bias plays an important role in the overall development process. Therefore the most prominent example is uncovering (hidden) biases in different developments.

2.2 The European Union Artificial Intelligence Act

As introduced within the introduction, the European Union laid down a proposal for harmonized rules on AI systems in April 2021. Not least because of fast emerging (generative) AI technologies and their overwhelming performance such as ChatGPT[3] led to raising attention to the powerful potential of (generative) AI systems in broad public. Thus different stakeholders claimed for an even more rigorous legal regulation compared to the proposed regulation published in 2021. Several changes were made to the proposal and submitted to the European Parliament. On June 14th 2023 the European Parliament voted for the adoption of the amendments to the AI-Act with a broad majority[4]. This vote completes the internal processes which enable further negotiation with European Member States[5] about the implementation as well as the development of standards along standardization organizations such as ISO which is planned to be applied in early 2025[6]. We refer to the adapted version of the AI-Act within this work where the amendments are marked in blue.

The regulation focuses on four main types of systems: prohibited systems, high-risk systems, limited-risk systems and minimal (or no-risk) systems:

(i) the prohibition of systems using techniques operating below the level of a person's conscious awareness or that use intentional methods of manipulation or deception with the aim of significantly distorting the behaviour of a person or a group, causing him or her to make a decision that he or she would not otherwise have made, in a manner that causes or is likely to cause substantial harm to that person, another

[2]https://vsdesign.org, accessed 24/08/23
[3]https://chat.openai.com/auth/login, accessed 12/09/23
[4]*https* : *//tinyurl.com/mry*8*v*52*h*, accessed 30/08/2023
[5]*https* : *//tinyurl.com/mreadmwd*, accessed 30/08/2023
[6]https://artificialintelligenceact.eu/standard-setting/, accessed 12/09/23

person or a group; exploiting person's or group's characteristics (personality, social- or economic situation) except for approved therapeutical purposes based on informed consent; biometric categorisation systems according to protected attributes (except informed consent); real-time remote biometric identification systems in publicly accessible spaces (consider exceptions included for law enforcement); **Article 5 par. (1)** [37, 38, p. 43]

(ii) the obligated application of rules and for defined high-risk systems in eight different domains: biometric identification, critical infrastructure, education and vocational training, employment and workers management, access to essential private and public services, law enforcement, migration and administration of justice and democratic processes – **Article 6** [37, p. 45] and Annex III [36, p. 4] with a strong focus is on related risks for health and safety.

(iii) limited risk systems which are covered by the information provision obligation on interacting with natural persons as well emotion recognition systems and systems that generate or manipulate image, audio or video content – **Article 52** [37, p. 69].

(iv) minimal (or no risk) systems are not affected by the regulation but are suggested to voluntarily submit to defined codes of conduct – **Article 69** [37, p. 80].

These obligations comprise a variety of required assessments, processes and documents from the early beginning towards the operation and monitoring of AI systems. Additionally, the regulation proposes a voluntary submission to gain awareness of the regulation's impact on corporations. As commonly known from GDPR law, the AI-Act also provides penalization provisions ensuring an appropriate application of defined rules. In case of an infringement on prohibited systems, the proposal provides a fee of up to 40 million Euro or 7% of the total worldwide annual turnover of the previous fiscal year (Article 71 par. (3)); non-conformity with Article 10 – Data and Data Governance and Article 13 – Transparency and Provision of Information are penalized with fines up to 20 million Euro or 4% of the annual turnover; supply of incorrect, incomplete or misleading information are penalized with fines up to 5 million Euro or 1% of the annual turnover; [37, p. 82][38]. Misclassifications of prohibited or high-risk systems also lead to penalties according to Article 71 [37][38].

To discuss and answer one of the first questions covered by the regulation: *Is my system affected by the regulation?* – we have to step through the articles and therefore we provide a condensed version of the proposal's text and provide implications on CRISP-DM in more detail in Chapter 4. We especially pay attention to Articles 1-15 of the AI-Act, which give an introduction to the major upcoming topics covered by Title I (general provisions), Title II (prohibited AI practices), Title III (high-risk AI systems) Chapter 1-2 (classification and requirements for high-risk systems). The following excerpt is not exhaustive. It does not substitute the definitions stated in the AI-Act and is used to get a common understanding in order to comprehend the context of this thesis. Articles 1-8, covering the classification of AI systems, are described within the following paragraph. Articles

9-15, covering the requirements on high-risk AI systems, are described in Chapter 4 according to the mapping to the respective phase in CIRSP-DM.

Article 1 and 2 - Subject Matter and Scope: The proposal regulates the (1.a) market placement, service and use of AI systems and shall be applied to providers that place their systems into the market or service in the EU independent from their origin as well as (1.b) deployers of AI systems that have their place of establishment or who are located within the Union and (1.c) providers and deployers in third countries when their output is intended to be used within the EU. The regulation points out certain exceptions like (3) exclusive development or use for military purposes or (4) international cooperation or agreements for law enforcement or judicial cooperation within the EU or with one or more member states and (5a) research, testing and development activities as well as (5e) components that are under free and open-source licence except for high-risk systems or covered by Title II or IV [37, pp. 38–39][38].

Article 3 - Definitions: For clarification purposes, we point out the regulation's most important definitions to understand the context and scope of certain terms:

- **artificial intelligence system** (1) "...means a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions, that influence physical or virtual environments;"[37, p. 39][38].
  - **risk** (1a) "...the combination of the probability of an occurrence of harm and the severity of that harm;" [38]
  - **siginificant risk** (1a) "a risk that is significant as a result of the combination of its severity, intensity, probability of occurrence, and duration of its effects, and its ability to affect an individual, a plurality of persons or to affect a particular group of persons;" [38]
- **provider** (2) "...means a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge;" [37, p. 39].
- **deployer** (4) "...means any natural or legal person, public authority, agency or other body using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity;" [37, p. 40][38]
- **operator** (8) "...the provider, the deployer, the authorised representative, the importer and the distributor;" [37, p. 40][38]
  - **affected person** (8a) "...any natural person or group of persons who are subject to or otherwise affected by an AI system;" [38]

Article 4 - Amendments: operators should do their best according to principles such as 'human agency and oversight', 'technical robustness and safety', 'privacy and data governance', 'transparency', 'diversity, non-discrimination and fairness' and 'social and environmental well-being'. [38].

Article 5 - Prohibited AI Practices: According to the proposal, the prohibited systems are (1a) subliminal systems that purposefully manipulate a person's or a group of persons' behaviour and (potentially) cause that person or a group of persons significant harm; (1b) systems that exploit vulnerable groups or characteristics of persons against age, mental or physical disability and (potentially) take influence on their behaviour cause significant harm to others; (1b/a) biometric categorization systems according to protected or sensitive attributes an the affection of them except given informed consent or therapeutical purposes; (1c) systems (by public authorities) that take advantage of social behaviour or personality characteristics (i.e. social scoring); (1d) systems that use "real-time" remote biometric identification in public spaces; systems that (1d/a) assess risk on criminal reoffenders; (1d/b) create or expand facial recognition databases in relation to facial images from the internet or CCTV; (1d/c) infer emotions in certain domains; [37, pp. 43–45][38].

Article 6 - High-Risk Systems: The classification of high-risk systems is mainly split into two parts: (1) the AI system is part of a safety component of a product or the system is the product itself, covered in Annex II (a comprehensive list with directives respective safety components for e.g. toys or lifts); (2) the AI system is applied in a domain, listed in Annex III which is also comprised in II-C [36, pp. 2–4] if they pose significant risk of harm to health, safety or fundamental rights of natural persons. Both types of AI systems require an overall application of all deliverables requested in the proposal [37, p. 45].

Article 7 – Amendments to Annex III: This article allows the EC to subsequently update and extend the list of high-risk systems (Annex III) according to certain conditions and criteria [36, pp. 2–4]. In more detail, the EC is (according to Article 73 AI-Act [37, pp. 83–84]) enabled to modify Annex I (list of comprised AI techniques and approaches) and Annex III (list of high-risk AI systems) as well as Annex V (declaration of conformity), Annex VI (internal conformity assessment) and Annex VII (conformity assessment of quality assurance and assessment of the technical documentation) whereas V, VI and VII are out of scope since we focus rather on the obligations respect the technical obligations rather than assessing them [37, pp. 45–46][38].

Article 8 - Compliance with Requirements: This article ensures compliance of high-risk systems to Articles 9-15 as well as an explicit note of the importance of the risk-management system and the intended purpose of the high-risk system as well as the adherence of state of the art techniques. [37, p. 46][38].

The accepted amendments on the proposal specify the notion of bias and its emerging within the Recital 44 according to definitions of data quality, requiring training data "...should also have the appropriate statistical properties, including as regards the

persons or groups of persons in relation to whom the high-risk AI system is intended to be used, with specific attention to the mitigation of possible biases in the datasets, that might lead to risks to fundamental rights or discriminatory outcomes for the persons affected by the high-risk AI system. Biases can for example be inherent in **underlying datasets**, especially when historical data is being used, **introduced by the developers** of the algorithms, or **generated when the systems are implemented in real-world settings**. Results provided by AI systems are influenced by such inherent biases that are inclined to gradually increase and thereby perpetuate and amplify existing discrimination, in particular for persons belonging to certain vulnerable or ethnic groups, or racialised communities." [38] which underlines requirements for bias monitoring among the whole AI lifecycle. Since bias does not automatically imply unfairness [3], we appreciate the introduction of "negative bias" within the amendments which is defined as bias that "... that create direct or indirect discriminatory effect against a natural person ... " [38], which underlines the important differentiation of non-harmful biases and bias that lead to unfairness or discrimination [37][38].

2.3 Providing Structure – Process Models

As introduced in Section 1.3, there are several process models available (KDD, SEMMA, CRISP-DM etc.) for projects and there are also comparative studies on the phases itself or different aspects such as project management, data- and information management, risk management etc. [8, 74] such as defined aspects which aspects a "good" process model should incorporate [73].

Although CRISP-DM lacks project management processes, integral processes and organizational processes [73] it is still a good process guide [75] i.e. a defacto standard in data mining (DM) projects. It is vendor- and context-independent which allows the utilization of additional tools and contexts [73]. Therefore CRISP-DM is selected to be the basis process model in the context of this thesis.

2.3.1 Examining CRISP-DM

According to Chapman et al. [27] CRISP-DM can be split and described in four different levels of abstractions according to specificity: phase, generic task, specialized task and process instance. The first two levels are considered to be "stable" which means to cover all process-related aspects according to all possible data mining applications and "complete" according to validity for new modelling techniques; together they are named the "generic" level. The "specialized task" level gives detailed and clear guidance on how certain tasks should be performed and how they may differ in certain situations. The "process instances" and fourth level prescribe the deliverables according to specialized tasks and represent a record of actions and decisions. Levels 3 and 4 (specialized task and process instances) are named "specialized" level [27]. The link between generic- and specialized levels is mapped via the data mining context which consists of (i) the application domain, (ii) the data mining problem type – types of objectives, (iii) the technical aspect – specific

challenges and issues that may arise and (iv) tools and techniques – potential supporting workarounds or domain-specific best practices. This mapping can be further considered as (I) mapping for the present – which utilizes the generic as-is process model for one-time-usage and the (II) mapping for the future – a systematical specialization of the generic level according to a certain context and specific needs for repetitive usage according an organization's dedicated requirements which requires to remove, add, specialize and rename generic contents as needed [27].

The approach and result of this thesis, providing an enriched version of CRISP-DM, follows the "mapping for the future" approach [27, p.8] and modifies the specialized level according to bias as well as fairness issues within a machine learning or an AI project.

As illustrated in Figure 2.2 CRISP-DM is structured in six phases: **(1) Business Understanding** focussing on project objectives and requirements from a business point of view translating identified criteria into a data mining problem definition; **(2) Data Understanding** which comprises data collection, insights into the data, assessing data quality/quantity and forming assumptions according to the business objectives; **(3) Data Preparation** where the base on the raw data, the final dataset for further modelling is prepared such as cleaning, sampling and deriviation of attributes; **(4) Modelling** which comprises various techniques according to the problem as well as calibration to optimal values; **(5) Evaluation** focus on target compliance according to the business objectives; **(6) Deployment** which focus on gaining insights based on the model's outcome as well as the creation of reports or implementing a repeatable data mining process [27].

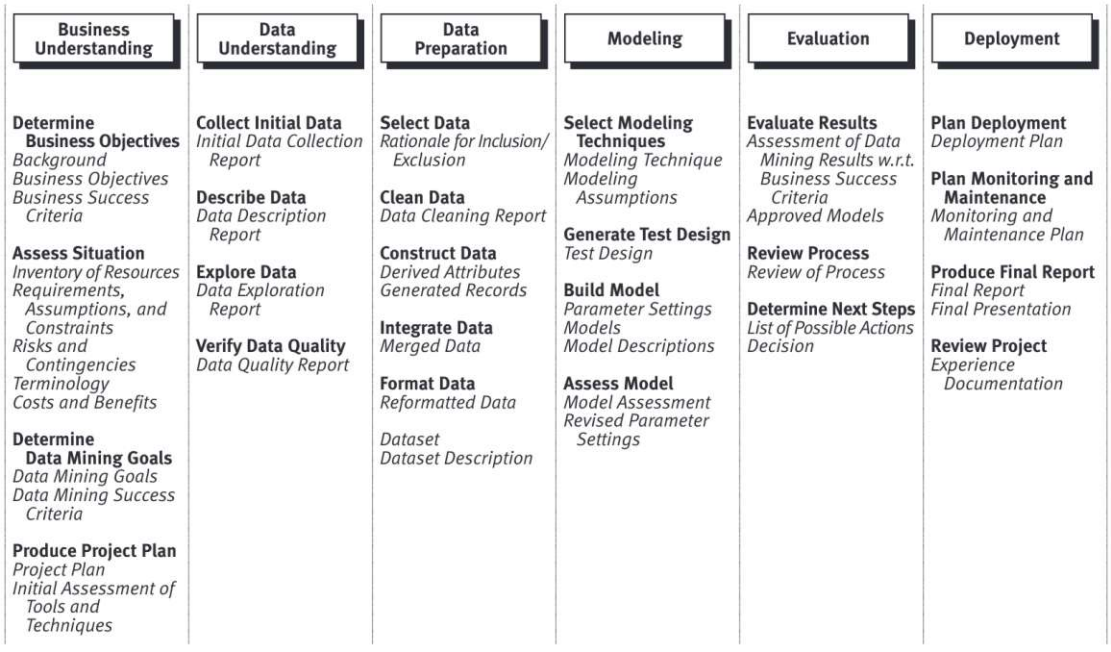

Figure 2.2: Overview of CRISP-DM and its six phases by Chapman et al. [27]

Delimitation 1: In Section 2.4 we will introduce the life cycle model proposed in ISO/IEC 22989 [58] which is different to the process model such as CRISP-DM since process models denote what to do whereas life cycle models denote the order [73] i.e. the stages the product steps through.

Delimitation 2: Nevertheless CRISP-DM was intended for data mining projects and therefore it lacks guidance to machine learning-specific applications [100] focussing on the operation. We will apply the approach of Studer et al. [100] and introduce Monitoring and Maintenance (MM) as Phase 7 which is further described and justified in Chapter 5.

2.4 Emerging ISO Standardization

The International Organization for Standardization (ISO) is an independent and non-governmental international organization subsuming 169 national standardization bodies[7] providing several standards and norms among several domains for its members. ISO recently published a number of documents introducing the standardization of AI systems. In alignment with the AI-Act the published ISO norms provide clearance on definition and distinction in nomenclature as well as guidance in the lifecycle and affecting aspects such as measures in fairness, human oversight, risk management, performance metrics and robustness in AI systems where the three most important ones in the context of this thesis are briefly described:

- *"ISO/IEC TR 24027:2021 – Information technology - Artificial intelligence (AI) - Bias in AI systems and AI aided decision making"*[8] [59] addresses bias in relation to AI systems which assesses sources of unwanted bias in AI systems, bias and fairness in AI systems as well as treatment of unwanted biases throughout the AI system lifecycle. Life Cycle Stage: *60.60 Published* (05/09/23).

- *"ISO/IEC 22989:2022 - Information technology - Artificial intelligence - Artificial intelligence concepts and terminology"*[9] [58] aims to harmonize terms and definition, technologies and techniques, trustworthiness (e.g. robustness and explainability), stakeholders as well as a lifecycle for AI systems. Life Cycle Stage: *60.60 Published* (05/09/23).

- *"ISO/IEC 23894:2023 Information technology - Artificial intelligence - Guidance on risk management"*[10] extends the existing ISO standard of risk management "ISO 31000:2018" to AI systems its scope, risk assessment and -treatment. Life Cycle Stage: *60.60 Published* (05/09/23).

[7] *https://www.iso.org/about-us.html*, accessed 28/08/23
[8] *https://www.iso.org/standard/77607.html*, accessed 28/08/23
[9] *https://www.iso.org/standard/74296.html*, accessed 28/08/23
[10] *https://www.iso.org/standard/77304.html*, accessed 28/08/23

- *"ISO/IEC FDIS 5338 Information technology - Artificial intelligence - AI system life cycle processes"*[11] focus on the standardizing the development process and the lifecycle of AI systems according to ISO/IEC: 22989:2022. Life Cycle Stage: *50.00 Approval* (05/09/23).

2.5 Summary of the Related Work

The different notions of bias depending on domain and context hamper (standardized) methods to unveil bias. Especially when sensitive attributes and their proxies are utilized within an AI system, unfairness, discrimination and ethical aspects are essential to be considered during AI development. The increased and broad utilization of AI systems led European lawmakers to regulatory actions, classifying AI systems such as prohibiting certain systems and calling dedicated requirements for certain systems. The utilization of these requirements in the context of bias requires structured consideration during the development process which cannot be fulfilled by existing process models and it is up to the experience of domain- and technical experts to implement dedicated steps to do so. ISO provided an approach and laid down general naming conventions as well as an AI lifecycle, providing a common base for the overall lifecycle of AI systems but lacking concrete steps.

In Chapter 3 we consider a dedicated assessment of bias and different bias types, provide examples in different contexts and provide a mapping according to respective phases in CRISP-DM. This step is the basis for detailed assessment on task level which is shown in Chapter 5.

[11] *https://www.iso.org/standard/81118.html*, accessed 05/09/23

CHAPTER 3

Assessing Bias in the Literature

3.1 Examples and Use Cases for Examination

A comprehensive understanding of bias types is crucial to take action with appropriate steps. As will be seen in the next section, bias definitions are enriched with examples for a better understanding. Bias per se does not imply unfairness [3] although undesired bias can lead to unfairness and discrimination [11, 76] as introduced in Chapter 2 separately, the identified bias types are further enriched with two defined examples to indicate that being aware of bias affects the performance in general. Although discriminatory effects and legal implications (described in Chapter 4) require regular checks on bias during the whole lifecycle of AI systems. Awareness, identification and mitigation of bias can lead to better-performing AI systems in general, independent from the use case. Therefore the first use case/example deals with non-protected attributes (i) production line data within factories whereas the second example does consider personal and protected attributes as (ii) people's data for a hiring system:

3.1.1 Use Case 1: Production Environment – without protected attributes

In this fictional example, a worldwide operating corporation with different branches in different countries produces packaging material and boxes for different categories of goods. Different base materials (plastics, paper and safe-food synthetics) are used. Each requires special techniques and machinery (injection mold, punching machine). Depending on the bias type, described in Section 3.3, this use case comprises (i) the implementation of a predictive maintenance algorithm to increase productivity and prevent costly outages of production machinery and (ii) an optical image recognition system to detect and reject erroneous products. Potential issues, which will be referred to at the respective bias type in more detail, could be (I) externally: different environmental conditions among the

different countries which affect the production lane; different qualities in the supplied base material, delays in spare part delivery or troubleshooting in case the branch is located in a bad accessible area etc. or (II) internally: arrangement of the production machinery, non-standardized processes among the company or branch, bad management, cultural differences which might affect the amount and output quality etc.

3.1.2 Use Case 2: Hiring Algorithm – including protected attributes

In this example, a company in the digital service sector is acting globally, considering creating an algorithm for hiring new employees to overcome the increasing demand for new employees and therefore to ease the workload and screening process for the HR department. The company has been operating for 30 years and extended its core business several times to fulfil the market's needs. The basis for this algorithm is all the data from previously screened people over the last three decades. Several problems may occur (I) internally: imbalances in genders, inconsistencies in data excerpation and measurement, temporary differences in required skills etc. (II) externally: structural imbalances in applicants, situation on the job market e.g. different balances between demand and supply of skillsets.

3.2 Bias in the Literature – Mapping Study and Analysis

3.2.1 Definition of Keywords and Process

The aim of the mapping study is to identify a broad variety of bias types that have been scientifically investigated, defined and justified. The focus of this work, as well as the strategy of this review, is based on receiving a diverse overview of investigated biases and their context. Therefore, survey papers reflect this variety in a structured manner and provide further literature resources. The considered search terms are "bias", "survey", "bias-types", "data", "data-mining" and "AI". The most important terms are "bias" and "survey" which should be present in the title, combined with the search terms "bias-types", "data", "data-mining" and "AI" which should be in the abstract, resulting in four search terms per research database as described in Section 3.2.2. The chosen research databases and literature archives are as follows:

- **Arxiv**[1] as a community-moderated and open-access database, provides non-peer-reviewed articles across a broad variety of fields such as computer science, mathematics and statistics which justifies the utilization according to the topic of this thesis.
- **ACM Digital Library** [2] is a well-known research database for computer scientists hosting all ACM (Association for Computer Machinery) publications such as journal

[1] https://info.arxiv.org, accessed 15/06/23
[2] https://dl.acm.org, accessed 15/06/23

articles. This library was chosen because of the strong focus on computer science e.g. especially the FAcct Conference (ACM Conference on Fairness, Accountability, and Transparency) where bias in research plays a prominent role.

- **ScienceDirect**[3] as an open-access database offers peer-reviewed research articles among an interdisciplinary set of fields. The most important ones in the context of this work are computer science and mathematics.

3.2.2 Conducting the search and selecting relevant studies

Formalizing the defined keywords to search within the databases results in the following search query which ensures "bias" and "survey" in the title and either "bias types", "data", "data-mining" or "AI" in the abstract:

```
1. [Title: "bias"] AND [Title: "survey"] AND [Abstract: "bias types"]
2. [Title: "bias"] AND [Title: "survey"] AND [Abstract: "data"]
3. [Title: "bias"] AND [Title: "survey"] AND [Abstract: "data-mining"]
4. [Title: "bias"] AND [Title: "survey"] AND [Abstract: "AI"]
```

In order to receive the context and origin of the topic, the search query was split into four single queries. This ensured also a fine granular manipulation of the search query in case the search term has not found any result which is described and documented accordingly.

Search Results: ArXiv

In ArXiv 30 papers were retrieved, although the 3rd query did not gather any results and was adapted to exclude the "survey" attribute and consider the investigation on bias in the context of data mining. The scope was limited to the fields of computer science, mathematics and statistics. The papers' abstracts have been skimmed and assessed, which resulted in 22 relevant papers for the next step in the review. Excluding criteria were defined as (i) no specific type of bias is mentioned or (ii) the context and field of the work are not in the notion of computer science focusing on AI or data mining.

Search Results: ACM Digital Library

In the ACM digital library, 12 papers were retrieved and 8 resources were identified as relevant. The remaining 4 resources were excluded because of criteria (i) and (ii) as described above and (iii) the context of the survey does not consider bias in the context of this work.

[3] https://www.sciencedirect.com, accessed 15/06/23

Search Results: ScienceDirect

The ScienceDirect database differs in search granularity since advanced filtering for specific keyword combinations such as differentiation between title, abstract and content is not possible. Therefore search resulted in 231 resources, containing 3 relevant resources. All other resources were not further considered according to defined exclusion criteria (i), (ii) and (iii).

3.2.3 Analyze Primary Studies

According to defined exclusion criteria in Section 3.2.2, 34 out of 273 retrieved resources were identified as relevant, which were read and assessed to extract and gather various types of biases. Out of these sources, four studies were identified as primary studies, considering different aspects and contexts of bias:

The first identified primary study was conducted by Mehrabi et al. [76, 77] where 23 types of bias were collected. They gathered and described the identified bias types in a universal manner, independent from specific subdomains, technologies and data enriched by demonstrative examples. Furthermore, discriminatory aspects, fairness metrics, methods for fair machine learning and techniques for bias mitigation are described.
Balayn, Lofi, and Houben [11] conducted a survey on bias and unfairness focussing on fairness metrics, fairness identification and mitigation methods in the context of a software engineering approach, including tools and techniques.
Socio-technological causes of bias, how bias is manifested in data and legal aspects of bias and its mitigation are described by Ntoutsi et al. [81] which is essential in terms of General Data Protection Regulation (GDPR) and the European Union Artificial Intelligence Act (AI-Act) set down by the European Parliament.
Kliegr, Bahník, and Fürnkranz [66] consider the socio-technological aspect in terms of human interpretability in the context of cognitive science and how it affects the human understanding of (interpretable) machine learning models which is subsumed as cognitive bias. The twenty different subtypes of cognitive biases, as well as their implications and mitigation strategies, are prescribed by Kliegr, Bahník, and Fürnkranz and all of these specific bias types are purely related to human beings [105, 66].

3.2.4 Ex- and Inclusion Criteria

Utilizing the mapping study, **60 bias types** have been collected. Emphasizing a broad variety of bias types, the collected definitions of bias types range from universal to very applied definitions according to subdomains in AI (Recommender Systems [114, 28], Natural Language Processing [70]), specific types of data (visual datasets [39], survey data [65]) as well as cognitive biases [66]. This broad variety of identified biases, their granularity, technologies and socio-technologic factors can hardly be compared meaningfully upfront which requires further explanation and further refinements in scope. The respective description and justification are covered within the following subsections Section 3.2.4 and 3.2.4.

Cognitive Bias

As stated by Kliegr, Bahník, and Fürnkranz [66] "cognitive bias" was introduced by Amos Tversky and Daniel Kahneman in the 1970s [104] and according to the "Encylopedia of Human Behaviour" [85] defined as "systematic error in judgment and decision-making common to all human beings which can be due to cognitive limitations, motivational factors, and/or adaptations to natural environments". Benson Buster[4] published an article, where more than 200 different subtypes of cognitive bias are listed and shown in Figure 3.1. This underlines the multifaceted influence of human beings working on and with data. Therefore, he distinguishes between four categories: (i) too much information (e.g. confirming one's beliefs), (ii) not enough meaning (e.g. simplifying meanings to grasp the meaning), (iii) the need to act fast (e.g. favouring simple-looking options over ambiguous options) and (iv) important steps to remember (e.g. specifics vs. generalities).

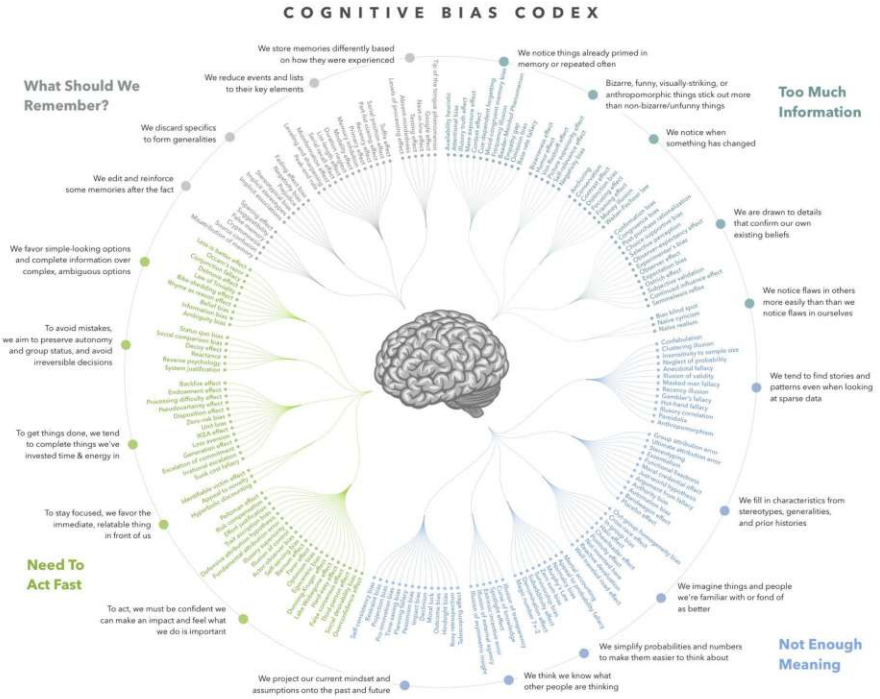

Figure 3.1: Identified cognitive biases from Buster Benson.

For example, **stereotyping** is considered a source of unfairness and can lead to discrimination [76], this bias addresses the socio-technological aspect in a general manner, which is essential, but hard to measure and requires general overthinking of socio-cultural aspects. The fact that there are over 200 different types of cognitive bias, summarized by Buster Benson[4], considering socio-cultural as well as socio-technological aspects, which

[4]https://betterhumans.pub/cognitive-bias-cheat-sheet, accessed 10/08/23

require profound knowledge in social sciences, would go beyond the scope of this work and therefore this bias type including subtypes is defined as out of scope. Although defined as out of scope within this work, cognitive biases are prone to influence stakeholders within the AI development process. Kliegr, Bahník, and Fürnkranz [66] presented a subset of 20 types of cognitive biases including their mitigation strategies.

Specific Application- and Data-based Bias

Defining and detecting bias requires the incorporation of domain-specific knowledge as well as an understanding of which features are considered sensitive/protected [99]. Moreover, it depends on the dataset, the specific model and the application [51] which can be confirmed through the results of the performed mapping study: As mentioned in the introduction, universal- as well as specific bias types were retrieved within the review e.g. in the domain of recommender systems [114], NLP [111] or in specific data such as in visual datasets [39] or survey data [65]. One issue is diverging definitions of identically named bias types was observed e.g. "omitted variable bias" in a universal context by Mehrabi et al. "when one or more important variables are left out of the model" [76] whereas Ferrara [41] interprets it as "when the data collection excludes partly or completely a certain group of people" [39]. Both definitions make sense but strongly depend on the context since the definition of Fabbrizzi et al. would map and subsume this type to "representation bias" defined by Mehrabi et al. ("happens from the way we define and sample from a population") [76]. Another issue is the introduction of additional and very specific bias types which should be considered as one single type of bias: Fabbrizzi et al. [39] defines "chronological bias" for visual datasets as "distortion due to temporal changes in the visual world the data is supposed to represent" which, to our understanding, can be subsumed to "temporal bias" [77, 76] or also relate to time series data.

This shows, that the strong dependency between context, application and data can lead to a diverging and misleading interpretation that requires the exclusion of specific applications as well as bias types defined for specific data in different domains. Furthermore, the aim of this work is to provide a broad overview of bias types in a universal manner and therefore it is hardly feasible to collect all existing bias types among several domains, applications and different types of data which requires relevant and specific expertise in each topic as stated in Section 6.2. Therefore these specific bias types are considered as out of scope within this work.

Summary Ex- and Inclusion Criteria

The introduced bias types and definitions are correlating to the domain (Section 3.2.4) and the stakeholders within the development pipeline (Section 3.2.4). Some bias types tend to be similar e.g. "human evaluation bias" (focus on the human individual during evaluation) is similar to "evaluation bias" (focus on evaluation technique itself). This domain-specific and categorical clustering is valid and is discussed in this section at the dedicated bias type respectively. The importance of a stringent and universal definition

of a specific bias type is indispensable which lays down the basis for metrics and its identification. Since diverging definitions and interpretations were gathered within the mapping study, related literature has been assessed to get a complementary insight ensuring a valid and precise definition of identified bias types.

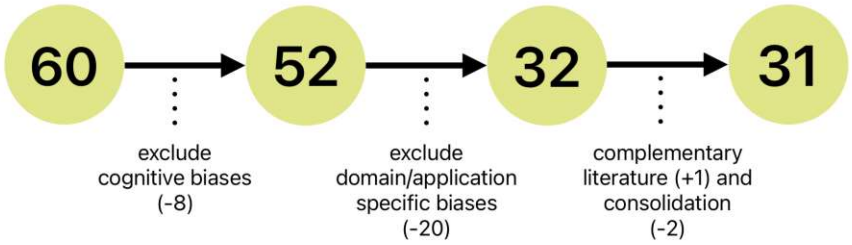

Figure 3.2: High-level overview: the identified number of bias types filtered by further exclusion criteria

Considering the excluded types of bias described before, summarized and sketched in Figure 3.2, **31**[5] **bias types** are remaining which will be presented in the next step.

3.3 Bias in the Literature – Results

To observe structure, we will apply the categorization according to Mehrabi et al. [77] where bias types are grouped into universal emergence and influence as **User to Data**, **Data to Algorithm** and **Algorithm to User**. This categorization shall point out the "Feedback Loop" phenomenon, where biased algorithmic outcomes might impact user experience and further affect data, algorithm and users which can amplify existing sources of bias [77] as shown in Figure 3.3.

[5]Remark: Originally, 32 different types were gathered; Age-, Gender-, and Racial Bias is subsumed in "Attribute Bias" since the emergence does not depend on the semantics but on an attribute (value); "Deployment Bias" was identified in complementary literature, which is essential to consider during the development process. This results in 31 identified bias types covered within this mapping study.

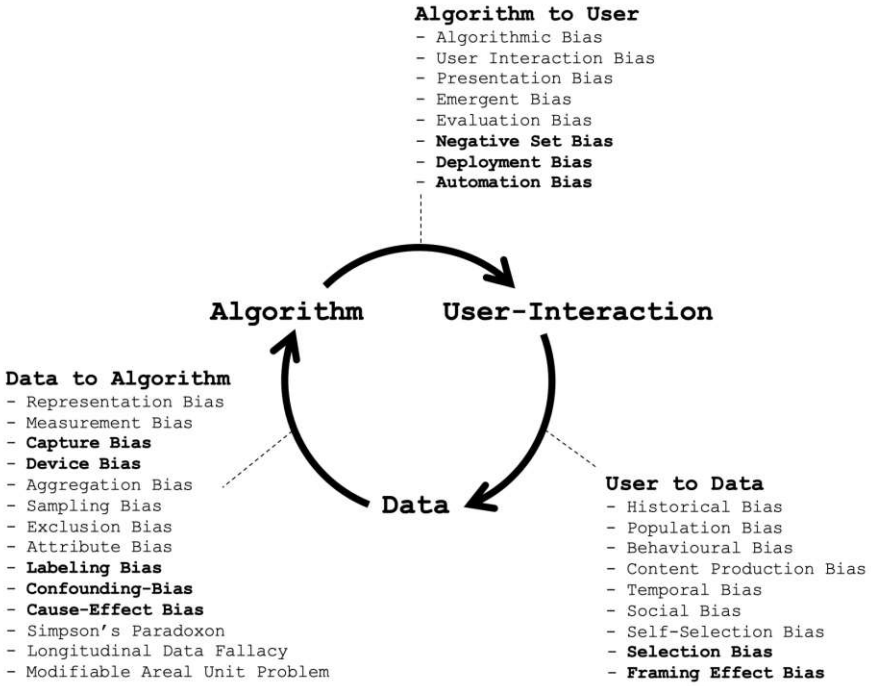

Figure 3.3: Visualized bias categorization according to the "Feedback Loop" defined by Mehrabi et al. [77] bias types marked in bold were based on the same criteria to this schema

The following list shows the key elements and the bias types that are assessed:

- **Bias Type:** Name of the defined bias type
- **Definition:** Definition of identified bias type according to literature. As stated, complementary definitions from related studies were assessed case-by-case to examine the definition's accuracy. In case of an additional definition, a justification of the differences as well as the selected interpretation is marked accordingly.
- **Explanation and Example:** If available within the source, examples are stated and/or enriched by defined examples from Section 3.1. The provision of two different examples with different contexts should give awareness, that bias is not automatically prone to discrimination but can lead to bad outcomes if remaining unhandled.
- **Emergence in CRISP-DM:** According to the stated research questions in Section 1.4 we gathered available information (from mapping study as well as related literature) about how the bias type can be identified according to specific emergence. Therefore a mapping to the respective phase in CRISP-DM is performed,

justified by informed arguments. In Chapter 5 this mapping is refined on the task level of CRISP-DM, providing a more detailed and actionable set of actions for the analyst.

- **Considerations and Notes:** Since much additional information about the different bias types was gathered, additional information such as relations, potential (side-) effects and potential mitigation strategies are denoted here.

Category: User-to-Data | Data-to-Algorithm | Algorithm-to-User

1. **Historical Bias**

   a) **Historical Bias:** *"... arises even if data is perfectly measured and sampled if the world as it is or led to a model that produces harmful outcomes."* [101]

   b) **Explanation and Example:** Utilizing systems, trained on historically biased data can still harm a population, despite the data representing and reflecting the world "as-is" precisely e.g. stereotyping a particular group. Historical bias can be implicitly incorporated which is why stakeholders have to consider whether the data should reflect the world "as-is" or in a "should-be" state [115, 32]. This opens up three approaches: (i) reflecting the world "as-is" and challenging the outcome afterwards (e.g. counterfactual fairness in post-processing) and not manipulating the data upfront which reflects the reality in an unmodified manner ("as is") or (ii) manipulating the data and correcting identified issues (e.g. manipulating label, over- or under-sampling) which transforms the world into "should-be" with unknown effects on performance metrics or (iii) a combined approach. Since historical bias can be present explicitly (e.g. by obviously favoured (protected) attribute values) and/or implicitly (e.g. socio-cultural), the determination of the true baseline (ground truth) is not available or very difficult to determine and more difficult to prove [101]. Biased decisions reflected in the data can also be a reason leading to historical bias [25].

      - **Use Case 1:** Presumed the company is legally obligated to increase the amount of recyclable materials in its products. This obligation could impact the machines, their output and the product's properties (shape or colour). If an algorithm is developed to ensure quality criteria by visual analysis, the incorporation of historical data could impact the performance of such a system e.g. by visual differences when the product is captured such as reflections or colour.
      - **Use Case 2:** Mehrabi et al. [77] claims that socio-technical factors can cause historical bias as well. Considering their example of women being underrepresented in management positions, we extend this example by a legal obligation to hire the same quote of women and men. If an algorithm is trained on historical data where men are highly overrepresented, it

would not be able to incorporate this legal obligation and be likely to suggest men, if no appropriate measure is implemented.

c) **Relevance in CRISP-DM:** Imbalances in the data set due to historical reasons can affect all types of data (nominal, ordinal, discrete and continuous) as well as other types like images, audio, text etc.

- **Phase 1:** Although incorporated in the world, stakeholders have to consider this right away during the definition of objectives and goals on a conceptual level: Assess the use case according to typical stereotypes or known pitfalls e.g. does the business objective consider an (implicit) separation according to a (protected) attribute or (sub-)group? Ask why one decision is favoured over another and ask why past decisions have been made the way it was – challenge them according to ethical guidelines, code of ethics, legal guidelines or codes of conduct as well as socio-cultural aspects e.g. women in technical jobs, gender pay gap, leadership positions or men in maternity leave.
- **Phase 2:** Besides the objectives and goals, the second and more detailed level has to be considered during data exploration, -understanding and -gathering either to prove, reject or adapt assumptions made in Phase 1: Check explicitly on typical stereotypic (protected) attribute values, their labels as well as combinations and check whether there is a significant or systematic difference in labels. Use this basis for further experiments according to protected and unprotected (sub-)groups. Check on deviations and basic statistic differences among (protected) groups and subgroups i.e. is one group or instance more likely to be labelled as positive? (e.g. when catholic people and protestant people have different outcomes? Do (and why do) young male cohorts behave differently than old male cohorts?) Identify potential proxy variable(s) e.g. is one group or instance with a specific attribute value more likely to be labelled as positive? (e.g. when the residence is a proxy for discrimination: (why) are certain districts favoured over others?)
- **Phase 3:** Considering findings from Phase 2 described above, appropriate steps have to be applied. Potential solutions can be over- or undersampling or gathering more data if data should be modified during pre-processing.

d) **Considerations and Notes:**

- Historical bias can lead to Representation bias [94]
- "The problem is that blind application of equalized odds locks in these historical biases without providing a justification for relying on the metric, and thus the historic biases, going forward." [107]

2. **Population Bias**

- **Population Bias:** *"Systematic distortions in demographics or other user characteristics between a population of users represented in a dataset or on a platform and some target population."* [83]
- **Explanation and Example:** Population bias results in non-representational data [77] which leads to bad system performance i.e. when source- and target population is different. In contrast to Aggregation bias, where bias emerges due to bad aggregation choices of (sub) groups, population bias incorporates systematic distortions within the world "as-is" independent from the way the data is selected or sampled.
  - **Use Case 1:** Extending the example described in Historical bias, assuming that the production factories are situated among different climates whereas hot climates with high humidity (tropical climate) require far more maintenance which affects the total output of those factory branches because of shutting down the production lane more often. Applying predictive maintenance algorithms globally without considering this issue, will result in bad performance overall.
  - **Use Case 2:** Branches located in cities with a broad supply of highly educated people e.g. due to a broad range of (specialized) universities, these branches can get more attention because of more supply of potential applicants.
- **Emergence in CRISP-DM:**
  - **Phase 2:** Although systematically incorporated in the data, population bias can be observed during data understanding, when objectives, goals and domain knowledge are required to analyze the data. It is important to consult domain expertise and consider as well as challenge dependencies, (causal) relations and insights.

3. **Behavioural Bias**

- **Definition:** *"... arises from different user behaviour across platforms, contexts, or different datasets"* [76]
- **Explanation and Example:** Since behavioural bias depends on human interactions, it can be subsumed into cognitive biases, although it can affect the data during the data generation process because of concealed user behaviour.
  - **Use Case 1:** The company decides to gather new data to develop a predictive maintenance algorithm for a particular machinery. In case the employee who operates this machinery behaves differently during the data generation process e.g. due to stress and therefore doing special treatments or services to keep the machinery running, this would falsify the data in the end.

  - **Use Case 2:** Within the hiring example, behavioural bias can arise if there are differences in behaviour during the application process compared to the actual performance when they got hired.
- **Emergence in CRISP-DM:**
  - **Phase 2:** Behavioural bias can be observed during data understanding in Phase 2 when objectives, goals and domain knowledge are required to analyze the data. It is important to consult domain expertise and consider as well as challenge dependencies, (causal) relations and insights.

4. **Content Production Bias**

- **Definition:** *"... arises from structural, lexical, semantic, and syntactic differences in the contents generated by users"* [76]
- **Explanation and Example:** User-generated data can be prone to contain different uses of language between people of different populations, ages or genders [76].
  - **Use Case 1:** If workers' data (descriptions or notes made during maintenance) are collected for designing a predictive maintenance algorithm, they can be biased according to different uses of language among the different branches.
  - **Use Case 2:** As stated in the explanation, different usage of language such as within a letter of application can be prone to be biased. Underlying factors can be different cultural habits and also the context of how people apply and write letters for application.
- **Emergence in CRISP-DM:**
  - **Phase 2:** Content production bias should be checked during data understanding in Phase 2 based on statistical measures (such as significant differences between groups and outliers) as well as cross-checking with a domain expert, trying to justify those assumptions.

5. **Temporal Bias**

- **Definition:** *"... arises from differences in populations and behaviours over time"* [76]
- **Explanation and Example:**
  - **Use Case 1 and Use Case 2:** Imagine the COVID-19 crisis, temporal bias would emerge if single branches have to be shut down, governmental restrictions would affect the output (use case 1) or the number of applicants would decrease drastically because of uncertainty and resulting unwillingness to switch- or apply for new jobs (use case 2).

- **Emergence in CRISP-DM:**
  - **Phase 2:** Temporal bias as a data-related bias should be considered during the data understanding task in Phase 2 which requires checking for inconsistencies over time. This requires background checks on external factors accompanied by domain experts to get critical consideration during data gathering.
  - **Operation:** In the case of online learning systems or regular retraining of the algorithm, newly gathered data has to be checked against KPIs from the original dataset to discover inconsistencies or deviations.
- **Considerations and Notes:** Considering the definition, temporal bias should be also seen in relation to Longitunidal Data Fallacy (temporal changes would influence the understanding of groups and their behaviour) and Behavioural bias (according to different characteristics and behaviour among groups or platforms).

6. **Social Bias**

- **Definition:** *"... occur when we unknowingly or deliberately make a judgment about certain individuals, groups, races, opinions, and so on, due to preconceived notions about the group. These can either be positive or negative beliefs and are often instilled in us based on our own culture and environment."* [53]
- **Explanation and Example:** Although social bias can be considered as cognitive bias described in Figure 3.1 its broad definition in context to discriminatory effects in AI yields to include social bias in this analysis. Social biases are naturally and mechanically reflected in the data and incorporated into the algorithm [53]. Potential actions and mitigation has to be considered case by case, but the important part is to discover and unveil social biases. In contrast to Historical bias where data is systematically affected or distorted (by socio-cultural issues) or to Population bias, where the distortions focus on particular groups in a demographic or cultural way, social bias is (in the definition within this work) purely considered to be on the analysts' and stakeholders' side. Therefore social bias can be the trigger for further biases such as Exclusion bias, Selection bias, Aggregation bias, Label bias or Representation bias but purely based on cognitively biased decisions during phase 2, phase 3 and operation. It requires objective and fact-based decisions during data understanding, data preparation and operation such as guidelines and handbooks on how and why decisions within data pre-processing have been made.
  - **Use Case 1:** Within the production lane example, Social bias can arise when favouring particular branches e.g. favouring a particular region by implicitly assuming better productivity among workers.
  - **Use Case 2:** Social bias within the hiring algorithm example can occur when analysts would favour a certain nationality e.g. Chinese people over

European people, just by their own beliefs or consideration that one group would be "better" than another.

- **Emergence in CRISP-DM:**
  - **Phase 2:** Documented favours should be aligned to findings in data understanding and vice versa e.g. potential irregularities should be checked according to business objectives and goals.
  - **Phase 3:** Gained knowledge should also be reflected in data preparation during sampling, deriving attributes and aggregation as well as ensure proper evaluation against harmful (sub-)groups in train-, evaluate- and test-set.
  - **Phase 5:** When evaluating the model, a special focus should be put on previously identified social biases or stereotypes to prevent exposure to harmful (sub-)groups
  - **Operation:** In case of Emergent bias or data shift in online learning systems or retraining, distributions between populations must be checked and aligned with identified or potential issues in previous steps.
- **Considerations and Notes:** Based on the context, there are diverging definitions of social bias such as in web- or platform-generated data [10], which do have their eligibility, but in a different and specialized context.

7. Selection Bias

- **Definition:** *"... is introduced by the selection of individuals, groups, or data for analysis in such a way that the samples are not representative of the population intended to be analyzed."* [99]
- **Explanation and Example:** Selection bias happens due to defining unreflected and uncautious criteria during the data gathering process including neglecting side effects e.g. gathering traffic data only on public holidays will not represent the populations' traffic behaviour.
  Selection bias is interchangeably named and compared as Sampling bias in the literature and therefore different definitions can be found such as Hellström, Dignum, and Bensch [54] equalizes sampling-, selection- and population bias, but we argue for a strict distinction since selection bias arises from the way how to choose from a population, whereas sampling bias defines how the data is split for training purposes, Population bias is systematic distortions in user characteristics and therefore have very distinct causes. Representation bias focus solely on the correct representation of (sub-)groups.
  - **Use Case 1:** Selection bias in the predictive maintenance algorithm example can arise when data is only gathered on Mondays (higher likelihood of errors due to running up machines again) which would result in poor algorithm performance.

  - **Use Case 2:** Within the hiring data example, selection bias can occur when data is gathered just in certain periods of the year or during a period of a hiring stop and therefore not reflect the actual population of applicants.

- **Emergence in CRISP-DM:**
  - **Phase 2:** Selection bias is dependent on data gathering and occurs during data understanding in phase 2. Proper planning of how data is gathered must be aligned with insights into the population (domain experts) as well as business objectives and insights from phase 1.

7.A **Negative Set Bias**

- **Definition:** *"... a consequence of not having enough samples representative of 'the rest of the world'."* [99]
- **Explanation and Example:** This bias can arise due to the lack of positive or negative examples which would result in bad performance. Therefore this has to be considered during data selection within data gathering.
  - **Use Case 1:** Within the production lane example, introducing a visual quality management algorithm, negative set bias can arise due to the lack of examples gone wrong (or good) during the data-gathering process.
  - **Use Case 2:** In contrast, within the hiring example, negative set bias could arise if a few people with very specific skills had not been rejected so far, which would lack data in this case.
- **Emergence in CRISP-DM:** Arises during data collection in phase 2, modelling in phase 4 and operation.

7.B **Self-Selection Bias**

- **Definition:** *"... is a subtype of the selection or sampling bias in which subjects of the research select themselves"* [77]
- **Explanation and Example:** According to different definitions of selection- and sampling bias mentioned in Selection bias, we claim that self-selection bias can affect both of them. Since self-selection bias requires actionable subjects, it requires human interactions which neglects an example for use case 1. Since it is related to the selection process in this example, we subsume it to selection bias.
  - **Use Case 1:** –
  - **Use Case 2:** Self-Selection bias can occur, if applicants, who would suit perfectly for this job, do not apply because they think they will not get the job upfront.

- **Emergence in CRISP-DM:**
  - **Phase 2:** Considered during data gathering, self-selection bias requires domain expert knowledge to investigate dependencies between population and data selection as well as data sampling.

8. **Framing-Effect Bias**

- **Definition:** *"Based on how the problem is formulated and how information is presented, the results obtained can be different and perhaps biased"* [99]
- **Explanation and Example:** A verifiable, achievable and clear problem definition lays down the basis for a successful project which requires clear language and statements. (Cognitively) biased statements and onesided vocabulary are prone to direct the results in this (biased) direction.
  - **Use Case 1:** Framing-Effect bias in production environment example, considering a predictive maintenance algorithm can arise when narrow assumptions among certain countries are made without considering needs and different situations among the rest (assumed that there are no standardized processes established).
  - **Use Case 2:** Within the hiring algorithm example, framing bias can occur, when the focus is purely on hiring full-time equivalents without considering an alternative combination of part-time equivalents or other aspects, such as challenging reasons such as why this is the goal and why this has to be fulfilled by full-time equivalents.
- **Emergence in CRISP-DM:**
  - **Phase 1:** Understanding the problem to its roots is essential to formalize an accurate and precise business objective which is performed in phase 1.
  - **Phase 2:** Precisely defined objectives support proper data understanding in phase 2 to decide which data to gather and how to do so.
  - **Phase 5:** Well-defined objectives and goals are necessary for evaluation to measure the performance that is performed in phase 5.

Category: User-to-Data | Data-to-Algorithm | Algorithm-to-User

9. **Representation Bias**

- **Definition:** *"... occurs when the development sample under-represents some part of the population, and subsequently fails to generalize well for a subset of the use population."* [101, 94]
- **Explanation and Example:** Additionally to the above definition, representation bias can originate "from how (and from where) the data was originally collected or be caused by the biases introduced after collection, either historically, cognitively, or statistically. Representation bias can happen due to

selection bias, i.e. when the sampling method only reaches a portion of the population or the population of interest has changed or is distinct from the population used during model training." [94] which distinct this bias clearly from Selection bias (how and when to gather data) and Sampling bias which focuses on the appropriate sampling when the model is trained. Therefore, representation bias in this context focuses on the training data used in development.

  - **Use Case 1:** When European machinery data is used for training a predictive maintenance algorithm and the system is subsequently applied worldwide.
  - **Use Case 2:** Accordingly, applying European applicants' data for training a hiring algorithm, will not perform well worldwide, since distributions of (protected) groups can change.

- **Emergence in CRISP-DM:**
  - **Phase 3:** "... if it does not reflect the use population" [94] as described in the use cases. "... if contains under-represented groups" [94] e.g. considering an age range within the algorithm, it can perform badly for under-represented groups. "... if the sampling method is limited or uneven" [94] e.g. considering only applicants after the first interview circle (getting a pre-selected sample compared to the target population). Therefore, it has to be precisely aligned with defined business objectives and insights from the data understanding phase.

10. **Measurement Bias**

- **Definition:** *"... happens from the way we choose, utilize and measure a particular feature"* [77] i.e. *"... when choosing and measuring features and labels to use; these are often proxies for the desired quantities. The chosen set of features and labels may leave out important factors or introduce group or input-dependent noise that leads to differential performance"* [102]
- **Explanation and Example:** To prevent measurement bias, specific domain knowledge is required to understand how certain attributes are captured and measured to have meaningful data.
  - **Use Case 1:** Within the predictive maintenance example, measurement bias could arise if wrong granularity, units or scales are used to capture machinery data which can result in bad performance e.g. deciding to measure output time in seconds whereas hundredths would be required to capture the accurate granularity.
  - **Use Case 2:** In case, women's maternity leaves are reflected in the data e.g. observed by less professional experience compared to others, the algorithm might learn this pattern and neglect people within this attribute combination.

- **Emergence in CRISP-DM:**
  - **Phase 2:** Measurement bias can be observed during data understanding in phase 2, especially in data gathering. Therefore domain knowledge and experts have to be consulted to measure the features properly.

10.A **Device Bias**

- **Definition:** Included distortions from the device used to capture the data [99].
  - **Use Case 1:** When different cameras or camera types (resolution, sensors) are used to capture images for developing a visual quality assurance system. Non-visible differences between images captured from those devices can impact performance.
  - **Use Case 2:** When different OCR (optical character recognition) software is used to digitize printed letters of application, OCR system's errors can be incorporated into the data.
- **Emergence in CRISP-DM:** As described within Measurement bias.

10.B **Capture Bias**

- **Definition:** *"... arise from the way a picture or video is captured (e.g., objects always in the centre, exposure, etc.)."* [99]
- **Explanation and Example:** Besides the errors occurring when different devices are used (Device bias), non-standardized methods of measuring the feature can incorporate bias e.g. different lighting, reflections, colours etc.
  - **Use Case 1:** According to the example described in Device bias, capture bias can arise due to different (non-standardized) lightning setups for capturing images.
  - **Use Case 2:** During the digitizing process of printed letters of application, capture bias can arise due to messy scanning e.g. the scanner tends to twist images which can result in systematic distortions.
- **Emergence in CRISP-DM:** As described within Measurement bias.

11. **Aggregation Bias**

- **Definition:** *"... during model construction, when distinct populations are inappropriately combined. In many applications, the population of interest is heterogeneous and a single model is unlikely to suit all subgroups."* [102]
- **Explanation and Example:** Mehrabi et al. [77] summarized the article of [32] very precisely in relation to design choices "such as use of certain optimization functions, regularizations, choices in applying regression models on the data as a whole or considering subgroups, and the general use of statistically biased estimators in algorithms [32], which can all contribute

to biased algorithmic decisions and therefore may bias the outcome of the algorithms" [77]

  - **Use Case 1:** Extending the example described in Historical bias, assuming that the production factories are situated in different climates where hot climate-zones with high humidity (equatorial area) require more maintenance which results in more frequent maintenance and therefore affects the total output of those factory branches. Assuming, factories in the colder north (northern hemisphere) have the same lower output, but based on a different cause (e.g. inefficiency), aggregation just based on their output would lead to wrong assumptions.
  - **Use Case 2:** When applicants of different branches are aggregated based on their group size, although the industries and projects of those branches differ drastically.

- **Emergence in CRISP-DM:**
  - **Phase 3:** According to CRISP-DM this can happen during data aggregation in the data preparation for further modelling. Therefore aggregations have to be challenged according to the goals and objectives as well as insights from data understanding from phase 1 and 2.

11.A **Modifiable Areal Unit Problem**: According to Mehrabi et al. [77] this is a subtype of aggregation bias with a focus on geospatial data "which arises when modelling data at different levels of spatial aggregation. This bias results in different trends learned when data is aggregated at different spatial scales" [77].

11.B **Simpson's Paradoxon** According to Mehrabi et al. [77] Simpson's Paradoxon occurs when "a trend, association, or characteristic observed in underlying subgroups may be quite different from association or characteristic observed when these subgroups are aggregated" [77] which is an effect out of Aggregation bias.

12. **Longitudinal Data Fallacy Bias**

- **Definition:** *"Researchers analyzing temporal data must use longitudinal analysis to track cohorts over time to learn their behaviour. Instead, temporal data is often modelled using cross-sectional analysis, which combines diverse cohorts at a single time point. The heterogeneous cohorts can bias cross-sectional analysis, leading to different conclusions than longitudinal analysis."* [77] "... which may create biases due to Simpson's paradox" [76] from [14]
- **Explanation and Example:** To understand and gather correct insights on the data, an isolated consideration of cohorts and groups has to be performed. Delimitations to Population bias where structural/demographic challenges are reflected in the population, Behavioural bias, where data is affected among

different user behaviour between platforms, longitudinal data fallacy considers changes within a cohort over time in context to others.

  - **Use Case 1:** Applied to the predictive maintenance example, production machinery data from the different production lanes has to be considered, understanding the data in an isolated way over a longer time to avoid misinterpretations e.g. if the machines were adapted or modified which would affect their productivity and therefore falsify conclusions.
  - **Use Case 2:** This bias can arise if the style of writing or amount of information changes in cohorts of younger age groups whereas stays the same among the cohorts of elderly age groups. Combining different cohorts can create a longitudinal data fallacy.

- **Emergence in CRISP-DM:**
  - **Phase 2:** Besides insights from the business understanding phase and its objectives, potential effects in changes of behaviour must be considered during data understanding and gathering which should be challenged by domain experts.

13. **Sampling Bias**

- **Definition:** *"... arises due to non-random sampling of subgroups"* [76]
- **Explanation and Example:** Although it is similar to Representation bias [77] (where the focus is on the presence of representative distributions among the population overall), sampling bias focus on the analyst when data is prepared for training purpose in this definition. Furthermore, a "consequence of sampling bias, the trends estimated for one population may not generalize to data collected from a new population" [77].
  - **Use Case 1 and Use Case 2:** If train-, evaluate- and test set are not randomly sampled among a group e.g. manufacturing year/type within the predictive maintenance (use case 1) or departments and its differences in application type (use case 2). Due to sampling bias, both examples can face a decreased performance whereas use case 2 can lead to discriminatory issues. Consider appropriate sampling techniques according to the type of data e.g. random sampling in time-series data can lead to over-estimation in performance.
- **Emergence in CRISP-DM:**
  - **Phase 3:** As stated, sampling bias should be checked during data preparation in phase 3.
  - **Phase 4:** During modelling sampling bias can occur along the generation of test design.

14. **Exclusion Bias (Omitted Variable Bias)**

- **Definition:** *"... occurs when one or more important variables are left out of the model"* [77]
- **Explanation and Example:** This can happen during data understanding and data preparation if variables are wrongly considered unimportant and therefore skipped for training.
  - **Use Case 1:** In the case of the predictive maintenance example, exclusion bias could arise due to excluding e.g. the temperature within the production hall which could be an indicator for harmful errors during production which would further lead to earlier maintenance intervals in case the temperature exceeds a certain temperature threshold.
  - **Use Case 2:** For the hiring algorithm example, exclusion bias could arise e.g. if the applicant's age is excluded, which can lead to increased weights on proxy variables; exclusion of certain (sub-)groups can further lead to Representation bias.
- **Emergence in CRISP-DM:**
  - **Phase 2, 3, 4:** Arises due to the exclusion of potential valuable variables which can occur upfront during data collection in phase 2, wrong exclusion criteria in phase 3 or even during modelling in phase 4 if the model considers proxy variables for the prediction in place of the excluded attribute.

15. **Attribute Bias (Gender-/Racial-/Age Bias)**

- **Definition:** Within the literature age bias [12], racial bias [3, 95] and gender bias [11] are mentioned, which are biases towards a certain attribute (-value) or group. Therefore we subsume these biases as "Attribute bias" since the effect is indifferent e.g. against a certain age group, a certain gender, religion, language or any other different attribute. Often related to a sensitive attribute (gender, race etc.) as an effect of Historical bias in a socio-cultural context, "Attribute bias" is not limited to protected attributes.
- **Explanation and Example:**
  - **Use Case 1:** In the predictive maintenance example, two types of machines are considered: very new machines with low-quality spare parts that allow for small measurement tolerances, and rather old machines with high-quality spare parts that allow for larger measurement tolerances. With the same output, the algorithm could learn that older machines need to be serviced earlier, even though they would run much longer due to better quality and higher material tolerances. This would have an overall impact on performance.

  - **Use Case 2:** If women have a lower probability of being predicted for a job compared to a man just because of their gender. This can be an effect of Historical bias (e.g. when only men are hired for certain positions which is reflected in the data), Representation bias (e.g. when appropriate training data is available, but women are underrepresented) or Selection bias (e.g. considering only men during data collection).

- **Emergence in CRISP-DM:**
  - **Phase 1:** When business objectives and goals are determined, protected attributes should be identified. Assess known issues on objectives among socio-cultural aspects (e.g. class imbalances or minorities) to identify Historical bias that can lead to Attribute bias.
  - **Phase 2:** During data gathering, appropriate selection of the population should be considered, assumptions from Phase 1 evaluated and proxy variables identified. Groups (and their outcomes according to the defined objectives) have to be opposed and analyzed to understand the reasons why groups with a specific attribute are favoured or disadvantaged. Reflect them according to the data collection process in relation to Selection bias that can lead to Attribute bias.
  - **Phase 3:** In data preparation consider excluding sensitive attributes and define a correlation threshold for proxy variables that may be excluded as well. Exclusion and threshold depend on the use case and have to be defined individually. Consider Representation bias that can lead to Attribute bias.
  - **Phase 4:** During model evaluation Attribute bias can be tested e.g. by counterfactual fairness, assessing the outcome while changing the value of sensitive (or proxy) attributes.
  - **Operation:** Perform regular or automated tests e.g. counterfactual fairness.

16. **Label Bias**

- **Definition:** *"... available labels for particular subpopulations are systematically incorrect"* [31]
- **Explanation and Example:** Label bias can arise due to inconsistent or missing guidelines on how data has to be labelled, due to different cognitive biases of the labellers and due to damaged sensors/equipment/setup when labelling is automated.
  - **Use Case 1:** As mentioned before, an optical quality management system within the production example could be affected by label bias if images are wrongly labelled systematically e.g. due to bad lighting in the production lane. In this case, label bias may arise due to Measurement bias.

  - **Use Case 2:** Within the hiring example, label bias is related to the human resource department, depending on the employee who assesses applicants. Depending on individual (cognitive) biases, this can be incorporated into data and be manifested as label bias which can also come along together with Attribute bias and therefore lead to discriminatory effects.

- **Emergence in CRISP-DM:**
  - **Phase 2:** During data understanding differences in labels have to be assessed carefully together with domain experts to unveil potential distortions in labels.
  - **Phase 3:** In case label bias is present, appropriate steps have to be considered during data preparation to ensure proper training.

17. **Cause-Effect Bias**

- **Definition:** *"... can happen as a result of the fallacy that correlation implies causation"* [76] i.e. "Correlation Fallacy" [98]
  This bias type was not categorized by Mehrabi et al. [77] but since Cause-Effect-bias is strongly dependent on the analyst during data understanding and preparation, we categorize this bias type within the Data-to-Algorithm group.
- **Explanation and Example:**
  - **Use Case 1:** If the output speed from production machinery has to be decreased because of a special and more valuable material-product combination compared to the standard products, an unaware analyst could assume that speed correlates with output, although the material-product combination would not allow a faster production and would further result in bad quality.
  - **Use Case 2:** A typical example within the hiring use case would be the assumption, that more women tend to work part-time because of private caretaking. Therefore one could assume that all women only seeking part-time jobs.
- **Emergence in CRISP-DM:**
  - **Phase 2:** Correlation-causation dependencies have to be assessed during data understanding in context with domain experts as well as in relation to sensitive attributes and proxy variable(s).
  - **Phase 3:** During data preparation an appropriate sampling must be considered for training.
  - **Phase 4:** Avoid misleading interpretation during model assessment.
  - **Phase 5:** Identified dependencies from phase 2 (as well as objectives from phase 1) have to be considered during evaluation to ensure that the model predicts based on eligible attributes.

17.A **Confounding Bias:** *"... occurs when an analyst tries to determine the effect of an exposure on an outcome but unintentionally measures the effect of another factor(s)"* [110]

- **Explanation and Example:** Compared to Cause-Effect bias, where causation is falsely implied by correlation, Confounding bias is the effect of investigating effects between one variable and another but unfortunately measures the effect of (an-)other factor(s). Since not classified by Mehrabi et al. [76], we subsume this bias as a subtype of Cause-Effect bias due to the same approach but a different effect.
  - **Use Case 1:** In case a certain machine produces more faulty output during hot periods, one could assume weaknesses within the machine (e.g. different textures of lubricants), although even the raw material can be affected by the high temperature, utilizing wrong machinery parameter can lead to more faulty products.
  - **Use Case 2:** Investigating the reasons for application for women, could be biased toward confounding bias if the business domain is falsely considered as a driving factor, although the true reason for application can be e.g. the internal kindergarten.
- **Emergence in CRISP-DM:** Due to the relation to Cause-Effect bias we refer to the same phases for this type.

Category: User-to-Data | Data-to-Algorithm | Algorithm-to-User

18. **Algorithmic Bias**

- **Definition:** *"... is when the bias is not present in the input data and is added purely by the algorithm"* [77]
- **Explanation and Example:** Algorithmic bias can arise due to inappropriate modelling techniques and related parameter settings. In the domain of music recommendation Flexer and Schnitzer [42] showed, that different algorithms produce different degrees of hubness which is performed by measuring distances in high dimensional spaces. Depending on the algorithm selection, this can lead to algorithmic bias.
- **Emergence in CRISP-DM:**
  - **Phase 4:** Algorithmic bias arises due to an inappropriate selection of algorithms or due to side effects in optimization e.g. changing weights to improve performance or mitigate different bias types. Integrate audit methods [7] for in-depth model assessment, apply explainability and interpretability methods, consider situation testing (create scenarios and constellations of attributes to determine and define the decision boundaries) and counterfactual testing (explicitly change protected attributes to different values e.g. male to female and observe, whether the output changes).

  - **Operation:** Algorithmic bias can emerge during retraining which requires consideration of steps done in phase 4. Consider logs from CRISP-DM 7.1 to identify edge cases as well as compliance with legal requirements e.g. updates in GDPR or the AI-Act.

19. **User Interaction Bias**

- **Definition:** is emerging from *"the user interface (e.g. via Presentation bias) and the user's own self-selected, biased interaction"* [10].
- **Explanation and Example:** User-Interaction bias can be affected by different bias types such as Presentation bias or Ranking bias (especially arising in the recommender systems domain) [10].
  - **Use Case 1:** User-Interaction bias in the optical image recognition system can occur if the worker has to spend different efforts to confirm positive and negative outcomes e.g. when negative examples require inconvenient or more actions in the user interface (by navigating through different tabs), so the user could tend to mark faulty products as good and vice versa.
  - **Use Case 2:** If the user interface only shows the top five selected applicants while the rest would require scrolling, the HR assessor might tend to focus only on the five presented.
- **Emergence in CRISP-DM:**
  - **Phase 2:** Consider the source of data gathered via user interfaces during data understanding and assess their background (e.g. ground truth if available, domain experts or users).
  - **Operation:** Although sufficient testing (of user interfaces) is required as well, potential changes in forms and further data-gathering strategies have to be assessed regularly.

19.A **Presentation Bias**

- **Definition:** *"... is a result of how information is presented"* [77]
- **Explanation and Example:** Presentation bias can have various facets due to the type of information/data presented (e.g. diagrams, pictograms or images) to the user. It is related to User-Interaction bias.
  - **Use Case 1:** In the case of a visual quality management system, presentation bias can arise if the results presented on the worker's computer do provide ambiguous information about why or where e.g. the particular workpiece is considered as a defect.
  - **Use Case 2:** If the graphical interface of the hiring algorithm highlights the (algorithmic) predicted choice without any explanation or justification, this could lead to presentation bias which can further result in discriminatory effects. Even worse, if the user observes a

pre-ranked list or selection from the algorithm without being communicated or indicated.

- **Emergence in CRISP-DM:**
  - **Phase 1:** Consider how the results will be presented to the end user, determine the need for interaction and assess their particular needs with domain experts and end users. Consider the need for explainability methods as well as usability in case of user-interaction is required.
  - **Phase 2:** Define suitable presentation methods (according to the type e.g. categorical data, time series etc.) along requirements defined in phase 1.
  - **Phase 4:** Assess implementation along defined requirements in Phase 1 and 2 with users and domain experts.
  - **Operation:** In the case of online learning systems, evaluate those techniques and objectives on a regular basis.

19.B **Emergent Bias**

- **Definition:** *"... happens as a result of use and interaction with real users. This bias arises as a result of a change in population, cultural values, or societal knowledge usually sometime after the completion of design"* [76]
- **Explanation and Example:** Although Mehrabi et al. [76, 77] list this as a dedicated bias, we subsume this bias type to User-Interaction bias since this bias arises due to user-interaction and is also mostly observed in user interfaces. Changes in the population are therefore covered within Population bias, as well as cultural values, are reflected in Historical bias. In contrast to user interaction as a superior bias type, emergent bias arises after development during the operation.
- **Emergence in CRISP-DM:**
  - **Operation:** As stated within the definition, this is clearly situated within the operation. Therefore statistical measures have to be assessed on a regular basis to react in case of data drift i.e. different user behaviour.

20. **Evaluation Bias**

- **Definition:** *"... happens during model evaluation ... the use of inappropriate and disproportionate benchmarks for evaluation of applications... "* [76, 77].
- **Explanation and Example:** In this case an example within the use cases would be invalid since this is highly related to the objectives and goals, type of data as well as context. Therefore a set of actions, described below, can help to unveil Evaluation bias.

- **Emergence in CRISP-DM:**
  - **Phase 4:** Evaluation bias occurs at evaluating the model's performance. Ensure the evaluation data matches the target population, perform quantitative comparisons of the models and investigate aggregated groups and measures to avoid hidden subgroup underperformance.

21. **Deployment Bias**

- **Definition:** *"... arises when there is a mismatch between the problem a model is intended to solve and the way in which it is actually used."* [101]
- **Explanation and Example:** Although deployment bias was not gathered during the mapping study, it was collected within assessing further literature. Errornous deployment can clearly affect the algorithm's performance, which is why it is considered in this work and therefore subsumed in the category of Algorithm-to-User.
  - **Use Case 1:** In case a predictive maintenance algorithm is developed based on European machinery data, it might lead to bad performance in case this algorithm is deployed to machines in the southern hemisphere e.g. wrong predictions due to different humidities.
    According to the optical image recognition system example, deployment bias can occur if the system is supposed to give suggestions to the quality assurance worker (who decides on rejecting a product), but then applied without a supervising actor i.e. applying suggestions as decisions.
  - **Use Case 2:** As stated in use case 1, different utilization of the hiring algorithm in different tasks can run into deployment bias e.g. if different countries or areas focus on different industries, they would require different skills.
- **Emergence in CRISP-DM:**
  - **Phase 6:** Deployment bias requires eligibility checks on the system and the context in which the algorithm will be utilized. Provide descriptions and delimitations of the data and algorithm according to the intended target system.

22. **Automation Bias**

- **Definition:** *"... the degree of agreement with erroneous advice provided by an Intelligent Agent ..."* [105]
- **Explanation and Example:** Automation bias is not covered within the groups of Mehrabi et al. [77], but from a technical point of view, it is situated within the Algorithm-to-User group since users (and further the data) can be affected from this bias.
  - **Use Case 1:** Within the visual quality management algorithm marking bad pieces, algorithmic bias could occur if the system is supervised by

a worker who has to confirm the rejection i.e. if the worker trusts the algorithm and confirms also erroneous rejections from the algorithm.

  - **Use Case 2:** Algorithmic bias can also arise within the hiring algorithm, e.g. when recruiters rely on the system and hire the algorithmically suggested candidate without considering equivalent participants (who would also be eligible or even better suited), i.e. when they blindly agree to the algorithmic suggestions.

- **Emergence in CRISP-DM:**
  - **Operation:** Provide regular tests to the utilising users by e.g. showing simulated (false) predictions and examples to assess the user's attention and ability to evaluate the outcome.

A condensed summary of the mapping into the phases of CRISP-DM can be found in Table 3.2, which forms the basis for the enrichment of the individual tasks in CRISP-DM. As shown, most of the identified biases occur during Phase 2 and Phase 3, i.e. during data understanding and data collection as well as data preparation. This again underlines the necessity and importance of a clear understanding of how to identify and deal with bias during development.

No.	Bias Type	P1	P2	P3	P4	P5	P6	P7
1	Historical Bias 1	✓	✓	✓				
2	Population Bias 2		✓					
3	Behavioural Bias 3		✓					
4	Content Production Bias 4		✓					
5	Temporal Bias 5		✓					✓
6	Social Bias 6		✓	✓		✓		✓
7	Selection Bias 7		✓					
7.A	Negative Set Bias 7.A		✓		✓			✓
7.B	Self-Selection Bias 7.B		✓					
8	Framing-Effect Bias 8	✓	✓			✓		
9	Representation Bias 9			✓				
10	Measurement Bias 10		✓					
10.A	Device Bias 10.A		✓					
10.B	Capture Bias 10.B		✓					
11	Aggregation Bias 11			✓				
11.A	Modifiable Areal Unit Probelm Bias 11.A		✓					
11.B	Simpson's Paradoxon 11.B		✓					
12	Longitudinal Data Fallacy 12		✓					
13	Sampling Bias 13			✓	✓			
14	Exclusion Bias 14		✓	✓	✓			
15	Attribute Bias 15	✓	✓	✓	✓			✓
16	Label Bias 16		✓	✓				
17	Cause-Effect Bias 17		✓	✓	✓	✓		
17.A	Confounding Bias 17.A		✓	✓	✓	✓		
18	Algorithmic Bias 18				✓			✓
19	User-Interaction Bias 19		✓					✓
19.A	Presentation Bias 19.A	✓	✓		✓			
19.B	Emergent Bias 19.B							✓
20	Evaluation Bias 20				✓			
21	Deployment Bias 21						✓	
22	Automation Bias 22							✓

Table 3.2: Overview of discovered bias types mapped according to the respective phase in CRISP-DM, including the introduced Monitoring and Maintenance (MM) phase (Phase 7), where they are likely to occur and where they have to be taken into account clustered into categories according to Mehrabi et al. [77]: User-to-Data (1-8), Data-to-Algorithm (9-17.A) and Algorithm-to-User (18-22)

CHAPTER 4

Legal Implications on CRISP-DM

Within this chapter, an assessment of the obligated requirements stated in Article 9-15 [37, pp. 46–52] including the amendments [38] is performed, focusing on "bias" and the identified bias types identified in Chapter 3. Therefore the legal requirements in the AI-Act are described (in the relevant context of this work), followed by relations to CRISP-DM, where a mapping is performed on to the respective phase. Because the AI-Act also covers several technical aspects besides bias, such as human oversight measures, security granting aspects or persistence strategies, we focus on bias and fairness affecting provisions.

4.1 Requirements: Articles 9-15

Since this work incorporates the amendments decided by the European Parliament on June 14th 2023, the respective updates are marked in blue to highlight the differences for further analysis and lookup. Numbers in parentheses refer to the paragraphs within the respective article.

4.1.1 Article 9 – Risk Management System

Legal Claim: A risk management system (RMS) for high-risk systems should (1) be established "throughout the entire lifecycle dots be integrated or even a part of already existing systems" unless it fulfils the requirements of Article 9; (2) consider a "continuous iterative process running throughout the entire lifecycle" requiring regular review and updates to ensure its effectiveness, documentation and significant decisions; (2a) identify, estimate and evaluate known and foreseeable risks that can harm the health or safety of natural persons, fundamental rights, equal access to opportunities or democracy if reasonably foreseeable; (2c) evaluate other significant risks identified in (2a) according to the post-market-monitoring system in Article 61; (3) "mitigate risks effectively while

ensuring appropriate implementation of requirements"; (4) reasonably judge residual risks being acceptable obligating communication to the deployer; (4a) prevent and eliminate identified risks according to technical feasibility involving domain experts; (4b) mitigate and implement control measures addressing significant risks that cannot be excluded; (4c) provide respective information according to Article 13; (7) consider predefined metrics and probabilistic thresholds that align with the intended system's purpose; (8) consider whether the (high-risk) system "is likely to adversely impact vulnerable groups of people or children" [37, pp. 46–48] [38].

To get suitable and sustainable risk management done, we propose to apply a defined standard process to structure these AI-related considerations, ensuring a clear overview of risks and appropriate action items in parallel along the whole lifecycle, if none is already implemented in the organization. Starting with the task "Risks and Contingencies" (CRISP-DM 1.2.3), regular updates and assessments among every single task throughout the Monitoring and Maintenance phase are necessary. Therefore we introduce two options:

An option to achieve a suitable risk management system (RMS) is to utilize frameworks e.g. according to Smart [96] which comprises seven phases:

i Establish the scope of RMS (e.g. according to CRISP-DM: business, data, model, deploy, monitoring)

ii Identify risks among defined topics (e.g. what to do in case of data imbalance and resulting effects?)

iii Quantitative analysis of individual risks (e.g. impact on defined metrics like accuracy? Potential occurrence of bias etc.)

iv Quantitative aggregation of individual risks (e.g. considering internal and external risks according to statistical uncertainty in cost and schedule)

v Risk allocation (e.g. defining responsibilities according to scope and topics)

vi Establish reserve strategy and mitigate risks (e.g. what can be done if risks are occurring?)

vii Monitor and review (e.g. monitoring via previously defined key performance indicators, periodic review and assessment)

Another option belongs to the family of ISO standards: The established ISO 31000 for standardized risk management is complemented in **ISO 23894** "Guidance on Risk Management" described in Section 2.4 and prescribes implications for the development and use of AI systems. Beginning from data collection, processing and operations, human oversight measures, regulatory as well as legal issues towards fairness criteria are considered in this norm. Although a broad range of aspects is covered, it leaves enough space for individual corporate adaptions according to internal processes and tools, which is why we recommend utilizing this norm.
Lee and Singh [68] created a risk identification questionnaire for detecting unintended biases in the machine learning development lifecycle to leverage risk identification for practitioners.

Emergence in CRISP-DM: We argue for **comprehensive risk management over all phases in CRISP-DM including Monitoring and Maintenance**, depending on the specifications and implications of the identified bias type. Examples according to the phases are:

Phase 1: Define actions to be taken when predictions based on business objectives lead to unforeseen or harmful outcomes for individuals or do not meet business success criteria (e.g. when Historical bias reflected in the data lead to discriminatory predictions) and define accountabilities. Gather and challenge known edge cases and prepare an operating plan, defining clear steps for each scenario to prevent ethical dilemmas. Consider legal risks such as non-compliance to GDPR or the AI-Act. Phase 2: Investigate identified risks from phase 1 according to the gathered data and assess their validity. Phase 3: Consider risks according to privacy violations, data security or -leakage in phase 4. Phase 4: Define actions to be taken when selected approach (pre-, in- or post-processing) does not provide the expected results? Phase 5: Identification and documentation of extinguished and residual risks: What to do if they occur? Definition and explanation of residual risk and their impact: Why can a certain risk not be eliminated? Phase 6: Consider risks according to Deployment bias, define limitations in case the system is deployed in a different context or a different use (case). Phase 7: Consider security vulnerabilities and define actions to be taken (e.g. adversarial attacks, data poisoning etc.).

When identifying risks, bias (in a societal and interpersonal context) plays an important role and can have a negative impact on the effectiveness of a risk plan [22]. Within this work, we focus on the bias aspect of risks which should be further incorporated in the chosen RMS.

4.1.2 Article 10 – Data and Data Governance

Legal Claim: (1) Training, test and validation data sets (test and verification for unsupervised- and reinforcement learning) shall (as far as technically feasible) be (2) "subject to suitable data governance" and "appropriate for the context of use as well as the intended purpose" particularly comprising (a) design choices; (a/a) "transparency as regards the original purpose of the data collection"; (b) "data collection processes"; (c) "data preparation and processing (annotation, labelling, cleaning, updating, enrichment and aggregation);" (d) formulation of assumptions according to the data; (e) "assessment of availability, quantity and suitability of data"; (f) "examination of possible biases ... likely affect health and safety, harm fundamental rights or lead to discrimination", especially avoiding 'feedback loops' as well as "appropriate measures to detect, prevent and mitigate possible biases;" (g) "identification of relevant data gaps or shortcomings that prevent compliance" In (3) requires that "training datasets and where they are used, validation and testing datasets, including the labels, shall be relevant, sufficiently representative, appropriately vetted for errors and be as complete as possible in view of the intended purpose" ... "appropriate statistical properties, including, where applicable, as regards the persons" or groups of persons "in relation to whom the high-risk AI system is intended to be used".

According to (5) it is allowed to process special categories of personal data ensuring negative bias monitoring, detection and mitigation for high AI systems (see paragraph 5 a-g); (6a) the deployer is held accountable in case the provider infringes compliance to Article 10 e.g. by missing rights on the data [37, p. 48][38].

Emergence in CRISP-DM: Lacking data- and information management and missing (management) governance are reasons for failures [17, p. 2321] especially dealing with privacy and security [74, p. 4].

Use documentation from the data dictionary, data collection-, description-, exploration- and quality reports from Phase 2 to check compliance on paragraph (2a-g). Additionally consider the implementation of data provenance strategies such as:

- PROV-O[1]
- Datasheets for Datasets [48]
- Dataset Nutrition Label [57]

Use documentation from the data dictionary, dataset description from Phase 3, test design from Phase 4 and from assessment of data mining goals- and success criteria from Phase 5 to check compliance on paragraph (3-6). Define and include legal aspects (e.g. rights on data etc.) that can lead to violating compliance with the AI-Act. Therefore legal aspects on this article should be part of CRISP-DM 1.2.2 Assess Situation to ensure compliance on paragraph (6a). The establishment of a data governance plan is crucial to be compliant with the AI-Act which we recommend performing in CRISP-DM 2.1 Select Data, if not utilized already.

4.1.3 Article 11 – Technical Documentation

Legal Claim: The technical documentation shall be updated regularly and comprise, at minimum, the requirements[2] stated in Annex IV of the AI-Act comprising: (Annex IV/1) a general description including (a) purpose, developer, date, version; (b) interfaces to interacting hard-/software; (c) versions of soft-/firmware(-update); (d) forms in which AI is intended to be placed on the market; (e) description of hardware which runs the AI; (f) marking of internal layout, where the AI system is a component of products; (g) instructions for use and installation instructions. Facilities are granted in the case of SMEs and start-ups, any equivalent documentation meeting the same objectives, subject to approval of the competent national authority.Annex IV/2 require a detailed description of elements and process for its development comprising architecture, design specifications, requirements on data, assessment and testing and validation process as well as a description of human oversight measures (Article 14) considering transparency obligations in (Article 13); Annex IV/3 comprises monitoring and limitatating aspects

[1] https://www.w3.org/TR/prov-o/, accessed 05/09/23

[2] Credit institutions additionally have to align with Article 74, which is not covered within this work.

such as human oversight measures following the RMS as well as operation (Article 14); Annex IV/4 require a detailed description of the RMS: according to (Article 9); Annex IV/5 require a log of development: document every change made in the system. Annex IV/8 require a detailed description of the post-market evaluation system: post-market monitoring system to evaluate the performance to ensure permanent compliance of the system concerning Title III/Chapter 2 which means compliance to "obligations for high-risk systems" (Article 8-15) [37, pp. 74–75].

Emergence in CRISP-DM: According to the requirements listed in paragraph (1) referring to Annex IV [36], required information can be gathered out of the documentation created in CRISP-DM as shown in Table 4.1:

P1	P2	P3	P4	P5	P6	P7
(1) a,b,d,f	-	-	-	-	(1) c,e,g,f	
(2) b,c,e	(2) d	(2) d	(2) a,b,g	(2) e,g		
(3)	-	-	-	-	(3)	(3)
(4)						
-	-	-	-	-	-	(5)
-	-	-	-	-	(6)	-
-	-	-	-	-	(7)	-
(8)	-	-	-	-	(8)	(8)

Table 4.1: Mapping of the requirements listed in Article 13 – Transparency and Provision of Information [Annex IV (1-8) of the AI-Act] according to suitable phases in CRISP-DM

Since several articles extend each other and extend over more than one phase in CRISP-DM (e.g. Article 13 – Transparency and Provision of Information, Article 14 – Human Oversight and Article 9 – Risk Management System) related documentation has to be extracted from the respective article's requirements gathered in the specific task. We annotated that within the respected tasks in CRISP-DM to preserve readability since relevant sources to gather information can be observed in Table 4.1.

4.1.4 Article 12 – Record Keeping

Legal Claim: High-risk systems are obligated to comprise (1) state of the art automatic logging capabilities during operation (2) throughout its lifecycle, including monitoring of operations (Article 29(4)) and post-market monitoring (Article 61) and (2(a)) presenting risks "in the meaning of (Article 65(1))" and (2(b)) "lead to substantial modification of the system". (2a) requires high-risk systems to log energy consumption of energy and environmental impact. Moreover, the logging should contain at minimum: (4a) a "recording of the period of each use (start-, end-, and use date)"; (4b) "reference database against which input data has been checked by the system"; (4c) "input data for which the search has led to a match"; (4d) "identification of natural persons involved" [37, p. 49][38].

Emergence in CRISP-DM: In this context, we refer to GDPR Article 15, which grants users the right of information provision (Data Subject Access Request – DSAR) and enables them to request underlying data/information which led to particular computation/prediction. Logging is aimed at storing the state of the system and significant occurrences in a system at a certain point in time to monitor the system and ease anomaly detection and defect management [21]. We propose to start focusing on logging in the CRISP-DM 4.1 Select Modelling Technique until the operation, covered via the CRISP-DM 7.1 Monitor to align design decisions along potential logging capabilities.

4.1.5 Article 13 – Transparency and Provision of Information

Legal Claim: Ensure that (1) the system's operation is "sufficiently transparent to enable users to interpret the system's output, system's functioning" (how it works, its processes and what data it processes) according to the intended purpose and use it appropriately . . . "allowing the user to explain the decisions taken by the AI system to the affected person"; (2) ". . . include concise,correct, clear and to the extent possiblecomplete information thathelps to operate and maintaining the AI system as well as supporting informed decision-making by users and is reasonably relevant";

(3b) "characteristics, capabilities and limitations of performance including" (3b/i) intended purpose; (3b/ii) level of accuracy, robustness and cybersecurity (Article 15 – Accuracy, Robustness and Cybersecurity) and "any circumstance that may have an effect on them"; (3b/iii) "effects and foreseeable risks on "its intended purpose under conditions of reasonably foreseeable misuse which lead to risks to the health and safety or fundamental rights or environment;" (3b/iiia) "the degree to which the AI system can provide an explanation for decisions it takes;" (3b/iv) performance on intended persons or groups; (3b/v) specifications of used input data which affects the intended purpose of AI system and "user actions that may influence system performance" (type or quality of data); (3d) human oversight measures (Article 14 – Human Oversight); (3e) expected lifetime, necessary maintenance, care measures to ensure proper functioning, software updates; (3ea) "a description of the mechanisms included within the AI system that allows users to properly collect, store and interpret the logs in accordance with Article 12 – Record Keeping(1)" [37][38].

Emergence in CRISP-DM: According to the requirements, appropriate methods of interpretability and explainability should be concerned in CRISP-DM 1.3.1 to meet paragraph (1) and (2). Incorporating documentation about the model and its performance created in CRISP-DM 4.3 Build Model, CRISP-DM 4.4 Assess Model, CRISP-DM 5.1 Evaluate Results, data dictionary to meet (3).

Delimitations: Some remaining points require additional documentation from related articles such as documentation of implemented security and robustness measures required (3b/ii) in Article 15 – Accuracy, Robustness and Cybersecurity, documentation of implemented human oversights measures (3d) required in Article 14 – Human Oversight,

the risk management system (3b/iii) from Article 9 – Risk Management System and the logging capabilities (3ea) Article 12 – Record Keeping.

4.1.6 Article 14 – Human Oversight

Legal Claim: Human oversight shall (1) include human-interface tools to enable oversight by natural persons "as proportionate to the risk associated with those systems"; (2) "prevent or minimizing the risk to health, safety or fundamental rights or environment " and "where decisions based solely on automated processing by AI systems produce legal or otherwise significant effects on the persons or groups of persons on which the system is to be used." (3) be ensured either (3a) by the provider; or (3b) by the user (4a) "fullycunderstand the capacities and limitations of the high-risk AI system . . . monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible;" (4b) be aware of Automation bias to "provide information or recommendations for decisions to be taken by natural persons"; (4c) "correctly interpret the system's output"; (4d) "be able to decide, in any particular situation, not to use the high-risk AI system . . . override or reverse the output"; (4e) "interrupting the system through a *"stop-button"* or similar technique" ensuring a safe shutdown; "except if the human interference increases the risks or would negatively impact the performance in consideration of generally acknowledged state-of-the-art" (5) apply the "four-eye principle" in point 1(a) Annex III ("systems . . . used for the real-time and post remote biometric identification of natural persons")[37][38].

Emergence in CRISP-DM: Requirements are **not considered by default** whithin the CRISP-DM process.

Delimitations: With respect to (1), (2), (3), (4) and (5) this is within the responsibility of the provider and the user, depending on the context and domain of the system, why we suggest to formulate and include these requirements within CRISP-DM 1.1.2 Determine Business Objectives to consider further steps along the CRISP-DM process. In contrast to human oversight, where a human has the privilege to interrupt the system, validate the system's output before the decision becomes effective or reject outcomes [67], related to CRISP-DM 4.1 Select Modelling Technique and CRISP-DM 7.1 Monitor, human in the loop is used during development e.g. when not much data is available, costs for errors are too high or the required data is rare or not available [109].

4.1.7 Article 15 – Accuracy, Robustness and Cybersecurity

Legal Claim: High risks systems shall (1) implement the security by design principle by default and "archive an appropriate level of accuracy, robustness, safety and cybersecurity to perform consistently" through its lifecycle (according to state-of-the-art measures . . .); (2) levels of accuracy and metrics should be defined in *instructions of use* in a clear language; (3) be "resilient as possible regarding errors, faults or inconsistencies" within the system and its environment, but especially with natural persons and other

systems; ... be robust "through technical redundancy solutions", backups and fail-safe plans; ... address mitigation measures against feedback loops and malicious manipulation of inputs in learning during operation for online-learning systems; (4) be "resilient as regards attempts by unauthorised third parties altering use, behaviour, outputs or performance;" ... ensure cybersecurity "appropriate to the risk and circumstances" ... address vulnerabilities, measures to "prevent, detect, respond to, resolve and control for attacks" to training dataset "('data poisoning'), or pre-trained components used in training ('model poisoning')", or causing the model to fail "('adversarial examples' or 'model evasion'), confidentially attacks or model flaws, which lead to harmful decision-making[37][38].

Emergence in CRISP-DM: Required documentation (2) and (3) can be utilized from CRISP-DM 4.3 Model Descriptions, from CRISP-DM 6.2 Plan Monitoring and Maintenance as well as from CRISP-DM 7.1 Monitor concerning the operation. Requirements regarding security stated in (1) and (4) should be addressed in CRISP-DM 1.2.2 Requirements, Assumptions and Constraints regarding requirements to "security" as well as in CRISP-DM 4.1 Select Modeling Technique considering e.g. membership inference attacks, model/attribute inversion attacks, adversarial attack or model stealing.

4.2 Additional Articles

In this section, we refer to relevant aspects with respect to the obligations for high-risk systems stated in Section 4.1 which grant exceptions or required additional information about how a certain article has to be implemented.

4.2.1 Article 52: Transparency obligations for certain AI systems

Refers to a specific group of systems which are not classified as high-risk but obligated to transparency obligations. Providers must guarantee that AI systems meant to engage with individuals are created in a manner ensuring that the person interacting with the AI system is promptly, clearly and understandably informed that they are indeed interacting with an AI system unless this fact is evident from the circumstances and the context of the interaction (e.g. which functions are AI-enabled, human oversight, responsibility within the decision-making process, legal situation or **allow natural persons to object against the application ... including their right to seek an explanation**) [37][38].

Special obligations for **(2) emotion recognition systems** or **(3) generative AI systems (text, audio and video)** are stated in this article.

4.2.2 Article 53: AI regulatory sandboxes

As measures in support of innovation member states shall "establish a controlled environment to test innovative technologies for a limited time on the basis of a testing plan agreed with the competent authorities" [37].

4.2.3 Article 61: Post-market monitoring system

Providers shall (1) establish and document a post-market monitoring system in a manner that is proportionate to the nature of the artificial intelligence technologies and the risks of the high-risk AI system; (2) actively and systematically collect, document and analyse relevant data provided by deployers or collected through other sources on the performance of high-risk AI systems throughout their lifetime, and allow the provider to evaluate the continuous compliance of AI systems with the requirements in Title III, Chapter 2 AI-Act and "include an analysis of the interaction with other AI systems environment, including other devices and software taking into account the rules applicable from areas such as data protection, intellectual property rights and competition law"; (3) be part of the technical documentation Article 10 [37][38].

4.2.4 Article 69: Codes of Conduct

Fostering the voluntary application of obligations for high-risk systems shall strengthen the trustworthiness of AI systems according to the defined aspects in paragraph (2) lit. a-g [38].

4.3 Concluding Remarks on Legal Provisions

Table 4.3 shows a mapping between the AI-Act's requirements and the corresponding phases in CRISP-DM. To ensure compliance with the requirements, regular checks have to be performed e.g. risk assessment (Article 9, which calls for the need for a "monitoring and maintenance" phase, introducing a continuous process ensuring compliance with the requirements during operation which will be introduced in Chapter 5. Despite this mapping and guidance, we suggest consulting additional legal expertise to check on compliance, especially in edge cases [37].

Article	P1	P2	P3	P4	P5	P6	P7
Article 9 – Risk Management System	✓	✓	✓	✓	✓	✓	✓
Article 10 – Data and Data Governance		✓	✓	✓	✓		✓
Article 11 – Technical Documentation	✓	✓	✓	✓	✓	✓	✓
Article 12 – Record Keeping				✓	✓	✓	✓
Article 13 – Transparency and Provision of Information	✓			✓	✓		
Article 14 – Human Oversight	✓			✓			✓
Article 15 – Accuracy, Robustness and Cybersecurity	✓			✓	✓	✓	✓

Table 4.3: Overview of mapped articles form the AI-Act [37, 38] according to the affected phase in CRISP-DM including the introduced Monitoring and Maintenance (MM) (Phase 7)

CHAPTER 5

Empowering CRISP-DM

After ascertaining the risk classification and obligations for high-risk systems, we provide guidance on when to take which steps to detect bias(es), potential mitigation strategies as well as guidance on considering bias-relevant aspects in high-risk systems in relation to the AI-Act mapped to CRISP-DM.

5.1 Extending CRISP-DM: Monitoring and Maintenance

As shown in Chapter 3, some bias types are likely to just emerge during operation (Automation bias) or are likely to persist throughout the development process (Emergent bias) which can affect the performance during operation. The AI-Act considerately addressed this issue and requires a post-market monitoring system for high-risk systems according to Article 61 during operation [37, 38].

This led to considerate steps in standardization (Section 2.4), which put a strong focus on operation and monitoring, continuous validation, re-evaluation and retirement [58]. This prompted us to introduce an additional phase to CRISP-DM to meet emerging requirements and standards. Therefore we introduced the already mentioned "monitoring and maintenance" phase introduced by Studer et al. [100] as Phase 7 and enriched it according to the context of the research questions to leverage proper monitoring and observation during operation.

5.2 Incorporating Findings into CRISP-DM

In this section, we extend the existing phases and respective tasks in CRISP-DM whereas black text is the original CRISP-DM specification [27] with green and blue text denoting extension to address bias issues, and compliance with the AI-Act, respectively which looks as follows:

phase.section.task Title of the task: Deliverable output

Original CRISP-DM specification [27].
Adoptions accoring to findings regarding bias from Chapter 3 including informed arguments according to [55].
Adoptions according to findings regarding the AI-Act focussing on bias relevant aspects from Chapter 4 including informed arguments according to [55].

Justification and explanations to the additional steps according to the mapping within the respective position in CRISP-DM following the evaluation within the design science approach based on informed arguments from Hevner et al. [55]. Footnotes within the CRISP-DM task relate to the correlating tasks presented.

5.2.1 Phase 1 – Business Understanding

Phase 1 – Business Understanding: Determine Business Objectives

1.1.1 Determine business objectives: Background

- Organization
  - Develop organizational charts identifying divisions, departments, and project groups.
  - The chart should also identify managers' names and responsibilities
  - Identify key persons in the business and their roles
  - Identify an internal sponsor (financial sponsor and primary user/domain expert)
  - Indicate if there is a steering committee and list members
  - Identify the business units that are affected by the data mining project (e.g., Marketing, Sales, Finance or legal department)
  - Identify ethical advisors and ideally form an ethical committee e.g. an Artificial Intelligence - Risk Board (AI-RB) including ethical-, legal-, business-, domain- and technical experts as well as decision-makers[1]
- Problem area
  - Identify the problem area (e.g., marketing, customer care, business development, etc.)
  - Describe the problem in general terms
  - Check the current status of the project (e.g., Check if it is already clear within the business unit that a data mining project is to be performed, or whether data mining needs to be promoted as a key technology in the business)
  - Clarify prerequisites of the project (e.g., What is the motivation of the project? Does the business already use data mining?)
  - If necessary, prepare presentations and present data mining to the business
  - Identify target groups for the project result (e.g., Are we expected to deliver a report for top management or an operational system to be used by naive end users?)
  - Identify the users' needs and expectations
  - Identify known issues regarding bias, fairness or discrimination internally (within the organization) and externally (known within the domain) within the defined problem area[2]
  - Identify legal frame conditions (e.g. known internal or external issues, regulations or dependencies) known within the problem area, identify legal obligations such as GDPR or AI-Act providing indication whether and how the domain is affected[3]

- Current Solution
  - Describe any solution currently used to address the problem
  - Describe the advantages and disadvantages of the current solution and the level to which it is accepted by the users
  - Describe the current solution's properties and attributes highlighting specifically ethical issues e.g. are people involved/affected in the problem/current solution or use case? Are there disadvantaged groups or minorities identified? Are there irregularities according to fairness? Are there any potential issues or known hazards according to discrimination whether direct or indirect? Are there any known edge cases to avoid in the future?[2]
  - Describe the current solution's properties and attributes according to legal issues or conflicts (GDPR, discrimination act or lawsuits) in relation to identified aspects in the "problem area" above[3] and document

[1] According to Blackman [20], organizations seriously dealing with ethical risks are required to have an ethics committee or an Artificial Intelligence - Risk Board (AI-RB) which should have the responsibility to recommend against developing the solution, confirm that the solution does not pose ethical risks and propose feature changes. We argue for including stakeholders from different domains to ensure a broad assessment of the developed project, which is done during the very early stage of a project such as in CRISP-DM 1.1.1.
[2] Background assessment is important to uncover Historical bias understanding the domain, the context and problems (in the current situation as well as the domain). Results have to be documented thoroughly and be the basis for a data dictionary.
[3] In alignment with the argumentation in [2], identified legal issues (e.g. in alignment with legal experts) have to be documented according to the utilized RMS in Section 4.1.1 to track and trace risks over the lifecycle (Article 9).

1.1.2 Determine business objectives: Business objectives

- Informally describe the problem to be solved
- Specify all business questions as precisely as possible
- Specify any other business requirements (e.g., the business does not want to lose any customers)
- Specify expected benefits in business terms
- Identify whether the specified requirements above (problem area and current) align with legal compliance such as GDPR, discrimination act, AI-Act) and redefine otherwise[1].
  - Define requirements according to human oversight measures required in Article 14 – Human Oversight
- If individuals are involved, define protected attribute(s) and identify protected class(es)[2]
- Identify whether/why and how the specified requirements above (problem, ...) involve unbalanced-, minority-, or protected groups as well as social stereotypes (races, gender ...), protected attribute(s) prone to be a basis for discrimination, potential biases. Do the requirements consider different treatment of entities which might be ethically questionable or not fair?[3]

- Consider potential alternatives or redefinitions and annotate/extend business requirements with use cases or user stories (best/average/worst case scenarios), red lines (e.g. what must not happen) – define the desired outcomes considering business interests and fair outcome and describe the differences according to the objectives[4]
- Identify the risk classification according to the proposed AI-Act [37] identifying the feasibility (prohibited-, high- and low-risk systems) and obligated steps (according to classification); if classified as high-risk, then Articles 9-15 are obligated[5]
- Beware of setting unattainable goals - make them as realistic as possible.

[1] Extend and update assessment as stated in CRISP-DM 1.1.1 [2].
[2] When business objectives are specified, the domain and subjects are defined which allows a determination of whether individuals or groups of individuals are affected.
[3,4] Refinement and extension of considerations made in CRISP-DM 1.1.1 according to [2] check whether the objectives identified might give rise to ethical issues (disadvantaged groups, ...) and legal aspects [3]. The introduction of use cases along with potential outcomes can ease the identification of risks, potential issues, refinement on individuals/groups and consideration of bias- and fairness aspects.
[5] The most reasonable point for detecting the regulation's applicability is after assessing the business objectives in CRISP-DM 1.1.2 since the problem, objectives and target group are defined.

Am I affected by the regulation? – A structured and timely detection of the obligated scope during the development process is crucial to save resources, minimise costs and set up a project plan. Precisely defined business objectives, concrete definitions should provide suitable information on whether the system is in scope or prohibited and whether it is a high-risk system, required to apply transparency obligations (Article 52) or even considered as a low- or no-risk system. Consider our assessment guideline in Figure 5.1 for structured decision-making and classification:

1. **Creation of facts and objectives:** CRISP-DM 1.1.1-2 as well as the required information for technical documentation of Annex IV (1.) [36, pp. 6–7], especially lit. a, b, d, f.
2. **Scope-Check:** Check whether the desired system is in the scope of Article 2 (e.g. an AI system that improves the productivity of a production system – would be in the scope of the regulation); if this is not the case, exceptions of Article 2 have to be checked.
3. **Prohibition-Check:** Check if the system is prohibited according to Article 5. If it is prohibited per definition, it is forbidden to be put on the European market. Therefore one should reconsider the system according to defined business goals.
4. **High-Risk Check:** Check whether the system is classified as high-risk according to Article 6 [37, p. 45] and Annex III [36, p. 4]; if the system is considered as high-risk, Articles 8-15 have to be applied – illustrated by the red box in Figure 5.1.
5. **Transparency-Check:** If the system is neither prohibited nor classified as high-risk, the environment of the AI system has to be assessed whether an interaction with natural persons, emotion recognition or manipulation of images, audio or video is planned as stated in Article 52 [37, p. 69] as well as systems obligated to transparency provisions – illustrated by the orange box in Figure 5.1.

6. **Codes-of-Conduct-Check:** In case the planned system is in the scope of the regulation according to Article 2 as well as Annex I [36], these can voluntarily submit to the same rules as for high-risk systems as a code-of-conduct recommendation. According to the proposed regulation, this may have further, wide-ranging reasons like environmental sustainability, accessibility to stakeholders, building trust etc. which could be utilized internally (operative or strategic) or externally (competitive) – illustrated by the green box in Figure 5.1.

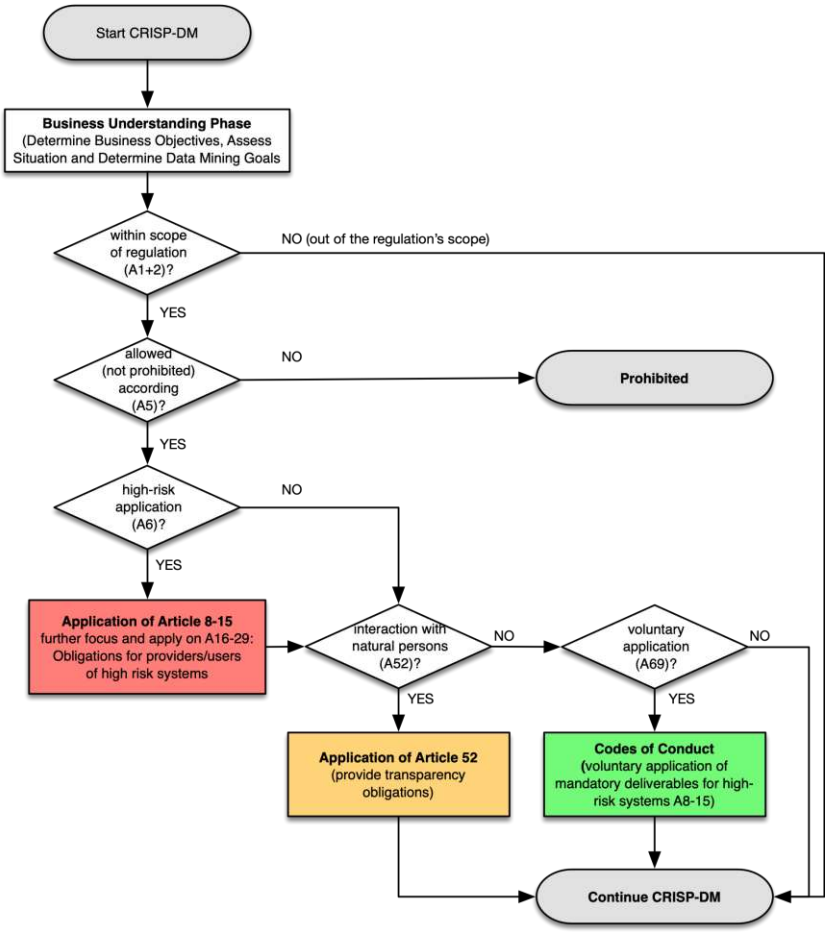

Figure 5.1: Scope- and risk classification according to the AI-Act in Phase 1 – Business Understanding

1.1.3 Determine business objectives: Business success criteria

- Specify business success criteria (e.g., Improve response rate in a mailing campaign by 10 percent and sign-up rate by 20 percent)
- Redefine and extend defined use cases and user stories in order to challenge the objectives with respect to operation in order to prevent Framing-Effect bias[1] and challenge underlying information (numbers, charts etc.) to challenge Presentation bias[2]
- Explicitly discuss the implementation of fairness (e.g. does and how would the organization benefit from a fair algorithm?)[3]
- identify who assesses the success criteria

[1] Framing-Effect bias can be prevented by having a clearly defined problem statement and objectives e.g. challenged by the AI-RB.
[2] Presentation bias can be prevented by challenging the presented information from different aspects and challenging the way how and why the information is presented in the way it is, preventing Framing-Effect bias stated in [1].
[3] Based on CRISP-DM 1.1.2 [2] discussion about potential biases leading to unfairness and discrimination should be started together with the AI-RB.

Phase 1 – Business Understanding: Assess Situation

"This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors that should be considered in determining the data analysis goal and project plan." [27]

1.2.1 Assess Situation: Inventory of resources

- Hardware resources
  - Identify the base hardware
  - Establish the availability of the base hardware for the data mining project
  - Check if the hardware maintenance schedule conflicts with the availability of the hardware for the data mining project
  - Identify the hardware available for the data mining tool to be used (if the tool is known at this stage)
- Sources of data and knowledge
  - Identify data sources
  - Identify type of data sources (online sources, experts, written documentation, etc.)
  - Identify knowledge sources
  - Identify type of knowledge sources (online sources, experts, written documentation such as a data dictionary or data governance documentation etc.)
  - Check available tools and techniques
  - Describe the relevant background knowledge (informally or formally)

- Personnel sources
  - Identify project sponsor (if different from internal sponsor as in Section 1.1.1)
  - Identify system administrator, database administrator, and technical support staff for further questions
  - Identify market analysts, data mining experts, and statisticians, and check their availability
  - Check availability of domain experts for later phases

The CRISP-DM basis provides sufficient guidance such as documents on the data or the identification of domain experts which is crucial for further bias assessment and does not require further adaptions.

1.2.2 Assess Situation: Requirements; assumptions; and constraints

- Requirements
  - Specify target group profile
  - Capture all requirements on scheduling
  - Capture requirements on comprehensibility, accuracy, deploy ability, maintainability, and repeatability of the data mining project and the resulting model(s)
  - Capture requirements on security, legal restrictions, privacy, reporting, and project schedule; consider security aspects as required for Article 15 – Accuracy, Robustness and Cybersecurity (par. 1, 4); consider legal aspects on data and its usage for Article 10 – Data and Data Governance [1]
- Assumptions
  - Clarify all assumptions (including implicit ones) and make them explicit (e.g., to address the business question, a minimum number of customers with age above 50 is necessary)
  - List assumptions on data quality (e.g., accuracy, availability)
  - List assumptions on external factors (e.g., economic issues, competitive products, technical advances)
  - Clarify assumptions that lead to any of the estimates (e.g., the price of a specific tool is assumed to be lower than $1,000)
  - List all assumptions regarding whether it is necessary to understand and describe or explain the model (e.g., how should the model and results be presented to senior management/sponsor)
- Constraints
  - Check general constraints (e.g., legal issues, budget, timescales, and resources)
  - Check access rights to data sources (e.g., access restrictions, password required)
  - Check technical accessibility of data (operating systems, data management system, file or database format)
  - Check whether relevant knowledge is accessible
  - Check budget constraints (fixed costs, implementation costs, etc.)

[1] CRISP-DM considers security aspects per default, why we recommend stating security-related aspects here.

1.2.3 Assess Situation: Risk and Contingencies

- Identify risks
  - Identify business risks (e.g., competitor comes up with better results first)
  - Identify organizational risks (e.g., department requesting project doesn't have funding for the project)
  - Identify legal risks (e.g. wrong classification of the system, not fulfilling an obligation, being not compliant, facing a lawsuit from a regulatory instance or private instance); if classified as high-risk: obligation to Article 9 – Risk Management System; identify risks potentially occurring during operation preventing reasonably foreseeable misuse (Article 9) AI-Act[1]
  - Identify financial risks (e.g., further funding depends on initial data mining results) and estimate costs for violating legal obligations, erroneous/false predictions according to biased or unethical outcomes (e.g. customer complaints, potential relating lawsuits or regulatory penalties)[2].
  - Identify technical risks
  - Identify risks that depend on data and data sources (e.g., poor quality and coverage)
  - Identify risks according to bias (and unfairness) along business objectives (e.g. the effects of an unfair or discriminatory outcome on different stakeholders, business departments, customers and subjects) along the project (Phase 1-6) as well as operation (Phase 7)[3]
  - Establish a risk management/governance system (e.g. ISO 31000, ISO 23894 or any other risk management system (RMS) or framework applied in the organization) and re-evaluate the risks after each CRISP-DM phase as well in production and if required, establishing compliance to Article 9 [4]
- Developing contingency plans
  - Determine conditions under which each risk may occur
  - Develop contingency plans

[1] Depending on the risk classification done in CRISP-DM 1.1.2, respective legal risks have to be assessed according to Article 9.
[2] In alignment with [1], resulting financial penalties violating the regulation as well as correlating laws (GDPR or non-discrimination laws) have to be considered.
[3] Working with data based on individuals may lead to Attribute bias resulting in risks affecting them during operation which should be further along the lifecycle.
[4] As stated in Article 9 requires the implementation of an appropriate risk management system (RMS) tracing the risks among the whole lifecycle.

Risk management considering bias and fairness: Since bias is an issue that can negatively impact the effectiveness of a risk plan [22] one has to consider and think through the consequences of not being prepared for worst-case scenarios like human tragedies, financial shortages or even the organizations' assets [20] which per se requires holistic consideration of the overall lifecycle closely attached to each phase of CRISP-DM. Considering a RMS from a (software) project management view, several frameworks exist as described in Section 4.1.1.

1.2.4 Assess Situation: Terminology

- Check prior availability of glossaries; otherwise begin to draft glossaries and incorporate prior findings such as bias(es), protected attributes and protected classes as well as initial considered notions of fairness and use cases from CRISP-DM 1.1.1-2[1]
- Talk to domain experts to understand their terminology
- Become familiar with the business terminology

[1] In order to gain a common understanding of terminology a well-hosted glossary is crucially required for Article 10 – Data and Data Governance, Article 11 – Technical Documentation and Article 13 – Transparency and Provision of Information.

1.2.5 Assess Situation: Costs and Benefits

- Estimate costs for data collection
- Estimate costs of developing and implementing a solution
- Identify benefits (e.g., improved customer satisfaction, ROI, and increase in revenue)
- Estimate operating costs

CRISP-DM documentation is sufficient.

Phase 1 – Business Understanding: Determine Data Mining Goals

1.3.1 Determine Data Mining Goals: Data mining goals

- Translate the business questions to data mining goals (e.g., a marketing campaign requires segmentation of customers in order to decide whom to approach in this campaign; the level/size of the segments should be specified).
- Specify data mining problem type (e.g., classification, description, prediction, and clustering). For more details about data mining problem types
- Specify criteria for fairness respective business objectives in alignment with the AI-RB[1]:
  - Specify the use case(s) according to minority groups (small number of data instances), advantaged or disadvantaged groups and outcome (according to favoured group(s))
  - Specify ethical and fairness criteria according to objectives, goals and user stories and outcomes for best-/average- and worst cases: What should the system predict including the boundaries (values, classes etc.)?
  - Check whether the task or use case is permissible or an ethically sensitive task
  - Define boundaries and dedicated non-goals e.g. where (geographically, users or providers) is the system prohibited (according to other attribute values or -ranges and socio-cultural differences)?
  - Specify criteria for explainability and interpretability (listed in CRISP-DM 4.3)
- Specify legal obligations and requirements[2]:

  - Recheck the data mining goals according to risk classification and its obligations according to Article 5 and Article 6 assessed in CRISP-DM 1.1.2
  - Consider transparency and provision of information required in Article 13 – Transparency and Provision of Information par. (1, 2)

[1] Although the selection of fairness metrics also depends on the data (Phase 2) and the model (Phase 4), (Phase 1) defines the problem area and objectives which is sufficient to provide a common (verbal) understanding, how fairness should be implemented. In CRISP-DM 1.3.2 appropriate success criteria have to be described and in Section 5.2.2 and Section 5.2.4 further steps are determined. Clearly define success metrics including AI-RB according to the defined problem in CRISP-DM 1.1.3.
[2] In order to avoid financial sanctions under Article 71 [37], early clarification of obligations has to be determined which is why we suggest to check again the data mining goals along the AI-Act's scope.

1.3.2 Determine Data Mining Goals: Data Mining Success Criteria

- Specify criteria for model assessment (e.g., model accuracy, performance and complexity).
- Define benchmarks for evaluation criteria as well as for fairness criteria (high-level)[1]
- Specify criteria that address subjective assessment criteria (e.g., model explainability and data and marketing insight provided by the model);
- Specify criteria for fairness respective its model assessment in alignment with the AI-RB[1]:
  - Define clear and measurable success criteria preventing Framing-Effect bias; assess, identify and document socio-technical issues to uncover Historical bias as well as toward potential Attribute bias upfront which should be challenged in CRISP-DM 2.1
  - Check whether success criteria (accuracy, performance, benchmarks etc.) are permissible from an ethical point of view (e.g. comparison of individuals based on a certain attribute)
- Specify legal obligations and requirements[2]:
  - Specify success criteria for ensuring the possibility for safe intervention: human oversight measures (Which measures are required? When and how does a human observe the process and how does a human interact/intervene in the process?) required in Article 14 – Human Oversight

[1] As stated within the business objectives in CRISP-DM 1.3.1, data mining success criteria have to be considered in this step.
[2] In order to avoid financial sanctions under Article 71 [37], appropriate required intervention techniques according to (Articles 14) have to be considered within the success criteria.

Phase 1 – Business Understanding: Produce Project Plan

1.4.1 Produce Project Plan: Project Plan

- Define the initial process plan and discuss the feasibility with all involved personnel
- Combine all identified goals and selected techniques in a coherent procedure that solves the business questions and meets the business success criteria
- Schedule a dedicated task of assessing implemented fairness metrics ensuring appropriate performance along the identified individual-, group- or subgroup metrics.[1]
- Plan time and resources for appropriate documentation and legal assessment of the AI system by regulators[2]
- Setup technical documentation Article 11 – Technical Documentation
- Estimate the effort and resources needed to achieve and deploy the solution. (It is useful to consider other people's experience when estimating timescales for data mining projects. For example, it is often postulated that 50-70 percent of the time and effort in a data mining project is used in the Data Preparation Phase and 20-30 percent in the Data Understanding Phase, while only 10-20 percent is spent in each of the Modeling, Evaluation, and Business Understanding Phases and 5-10 percent in the Deployment Phase.)
- Identify critical steps
- Mark decision points
- Mark review points
- Identify major iterations

[1] A thorough testing of fairness measures is necessary to ensure non-discrimination, gathering corresponding benchmarks and information for documentation, as well as being compliant with the regulation's requirements stated in [2].
[2] Article 9 – Risk Management System require to provide appropriate documentation on the implemented system and the utilized data. Therefore information gathered in [1] has to be implemented in Article 10 – Data and Data Governance, Article 11 – Technical Documentation and Article 12 – Record Keeping, Article 13 – Transparency and Provision of Information which should be done in a dedicated task.

1.4.2 Produce Project Plan: Initial assessment of tools and techniques

- Create a list of selection criteria for tools and techniques (or use an existing one if available)
- Choose potential tools and techniques
- Evaluate the appropriateness of techniques
- Review and prioritize applicable techniques according to the evaluation of alternative solutions
- Consider security aspects as required for Article 15 – Accuracy, Robustness and Cybersecurity

5.2.2 Phase 2 – Data Understanding

2.1 Collect initial data: Initial data collection report

- Data requirements planning
  - Plan which information is needed (e.g., only for given attributes, or specific additional information)
  - Check if all the information needed (solve the data mining goals) is actually available
  - Create a data dictionary including meta data (-information) and context from a domain expert, incorporating findings from Phase 1 (protected attributes, considered notions of fairness, potential discrimination etc.) and use this as the basis for considering ethical- or fairness challenges needed for data gathering[1]
  - Apply a data governance system and strategy to ensure compliance with Article 10 – Data and Data Governance
- Selection criteria
  - Specify selection criteria (e.g., Which attributes are necessary for the specified data mining goals? Which attributes have been identified as being irrelevant? How many attributes can we handle with the chosen techniques?) and check if the permission for using the information is given (informed consent on GDPR) and whether needed information is allowed to be computed[2]
  - Select tables/files of interest
  - Select data within a table/file
  - Think about how long a history one should use (e.g., even if 18 months of data are available, only 12 months may be needed for the exercise)
  - Consider arising biases:
    - Historical bias: If indicated in business objectives/goals or existing solutions consider gathering more data among underrepresented groups or individuals with a particular attribute [3]
    - Population bias: Consider if the population to be sampled from differs in characteristics from others and challenge if this could lead to an issue; document this in the data dictionary[4]
    - Behavioural bias: Consider differences in user behaviour: on the web across platforms; different data sources; different user characteristics and user environment during data gathering. Challenge potential causal relationships that can lead to Population bias: as well as background information for aggregation that can lead to Aggregation bias. Document observed differences within the data dictionary[5]

* Content production bias: In relation to Behavioural bias ensure uniform naming conventions, clear statements, guidelines and definitions ensuring common understanding. Denote observed differences in the data dictionary to prevent Aggregation bias in Phase 3[6]
* Temporal bias: Identify external factors that influence data generation and document potential side effects on the data within the data dictionary[7]
* Social bias: Define clear data-gathering guidelines that meet the business objectives, formulate them as inclusive as possible and address potential prejudices against protected attributes or other known aspects against minorities. This should be considered in a broad way including the AI-RB ensure different contexts, social- and cultural aspects[8]
* Selection bias: Define the data (-subjects) according to the business objectives in alignment with the data collection and sampling process' properties (e.g. time(-span), date, duration, place, season, point of extraction/measurement) to collect data across all defined target group ensuring inclusion of defined (protected) groups. Be as precise as possible but as inclusive as necessary to gather a correct and representative data sample of the population. Ensure objective data collection to prevent Self-Selection bias [9]
* Framing-Effect bias: Based on poorly defined problems and business objectives in Phase 1, Framing-Effect bias can progress during data collection which requires a thorough consideration of the business objectives and the required data to challenge which data(-attribute) is well suited to solve the problem[10]
* Measurement bias: Use standardized units and preserve high fidelity (e.g. raw data without loss), prevent derivations or aggregations of attributes to be measured, utilize calibrated or standardized (measurement-)equipment to prevent Device bias, define clear guidance about how a particular feature or situation has to be captured to prevent Capture bias and provide documentation on how features are measured in a detailed way
* User interaction bias: In case of online gathered data the corresponding form has to be set in accordance with the AI-RB to avoid subjective data or biased data according to the structure and context

[1] Creation of a data dictionary is crucial to have all relevant (meta) information of the data in a dedicated place, understanding relations, context and background information which is further required to be used for Article 10 – Data and Data Governance, Article 11 – Technical Documentation, Article 12 – Record Keeping and Article 13 – Transparency and Provision of Information. Consider human-related data (employees in different environments, demographics, countries etc.) and non-human data (different categories of certain subjects e.g. spare parts of machines) which also are prone to bias. Identify and document, protected attributes, their values and ranges. This can be done within a living document/wiki along customary ways, already done within the organization. Identify causal relations and potential correlations (expectations and assumptions which should be checked in CRISP-DM 2.1.2).
[2] Consider legal permissions to gather data as well as the necessity to process certain personal attributes/data according to the desired purpose (personal data must be anonymized/pseudonymized which can lead to problems such as verifiability of successful anonymization or decrease of system performance).
[3] Refers to the pre-processing mitigation strategy from [115] getting from

the world "as-is" to "should-be" by systematically oversampling of data; if a "groud-truth" is available, comparisons should be checked.
[4] Carefully think about the target population, the sampled population and define differences according to meet business objectives [83].
[5, 6] Most likely to be observed during data gathering and -generation in CRISP-DM 2.2.
[7] External factors might influence the data generation process over time which affects the data's representativeness during this period. In case data is already available, this has to be checked in CRISP-DM 2.2.
[8] Clear and well-defined guidelines during data collection are essential to have inclusive data which requires holistic consideration of the target group (including protected attributes, -groups and -individuals). This is necessary for further steps in CRISP-DM 2.2 and Phase 3, assessment of bias and fairness as well as obligations according to Article 13 – Transparency and Provision of Information.
[9] We localize Selection bias within the data collection process. It is crucial to thoroughly consider, define and document this process to prevent downstream Representation bias in CRISP-DM 2.2. Avoid restricting data analysis to truncated portions of dataset which can lead to unwanted selection bias [99].
[10] Emerging in CRISP-DM 1.1.3 this bias would persist without thoughtful challenging defined problems, objectives and required data.

Consider aspects on data quality (fitness for purpose) and quantity (number of instances and distribution of labels across protected groups) as well as missing values and their impact e.g. which level of missing data is acceptable? Do missing values have an impact on fairness metrics?

Based on critical/risky decisions depending on protected groups, edge cases, ethical- and fairness guidelines, define quantities and qualities e.g. identify and define certain ratios among attributes i.e. data should reflect the environment the system is deployed in. Identify and define certain attributes/values for fairness metrics i.e. which attribute (combinations) do we consider in challenging what is fair or not?

Further options for creating a data dictionary:

- **"Datasheets for Datasets"**[48]: documenting composition, collection process etc. ensuring better communication between dataset creators and user providing transparency and ensuring accountability.
- **"Dataset Nutrition Label"**[57]: standardize data analysis and harmonizing insights, facts and properties of a dataset such as strengths and weaknesses on data/model/API and potential biases,
- **"Semantic data dictionary"**[86]: provides a machine-readable variant for

exploration and further processing as well as enabling standardization and harmonization across diverse datasets.

2.2 Describe Data: Data Description Report

- Volumetric analysis of data
  - Identify data and method of capture
  - Access data sources
  - Use statistical analyses if appropriate
  - Report tables and their relations
  - Check data volume, number of multiples, complexity
  - Note if the data contains free text entries
  - Assess group- and subgroup sizes ensuring appropriate representation of the population to be represented to prevent Representation bias (correct properties and distributions of the target population) and Attribute bias (having enough samples containing a certain attribute value to have appropriate balance in the data)[1,2]
  - Check the necessity for synthetic data[2]
- Attribute types and values
  - Check accessibility and availability of attributes
  - Check attribute types (numeric, symbolic, taxonomy, etc.)
  - Identify protected attributes, minority-, advantaged- and disadvantaged (sub-)groups according to findings from Phase 1 and CRISP-DM 2.1 and document it within the data dictionary; highlight especially human-related data which is subject in CRISP-DM 2.3[3]
  - Check attribute value ranges
  - Analyze attribute correlations; check for correlations with protected attributes; check distributions and correlations within and between (minority) groups and subgroups, check for causality and proxy variables unveiling protected attributes or affiliation to (dis-)advantaged groups (e.g. isolated consideration of (sub-)groups applying correlation matrices and formulating assumptions which should be checked in Phase 4)[4]
    - Cause-Effect bias: challenge and carefully differ between causal relations and correlations and document findings within the data dictionary
  - Understand the meaning of each attribute and attribute value in business terms; consult domain experts for context- and domain knowledge, challenge identified correlations and assumed causal relations and document findings in data dictionary
  - For each attribute, compute basic statistics (e.g., compute distribution, average, max, min, standard deviation, variance, mode, skewness, etc.)
  - Analyze basic statistics and relate the results to their meaning in business terms
  - Decide if the attribute is relevant for the specific data mining goal; prevent Exclusion bias by not excluding attributes in a very early stage; highlight protected attributes that are prohibited to use for computation or can lead to unfair bias to prevent Attribute bias[5]

[1] In case a certain (sub-)group or (protected) attribute value is systematically undersampled or missing, establishing appropriate balance in data is proposed by [23]. Therefore it depends on the use case and whether to ensure a cor-

rect representation of the population or to oversample a certain (sub-)group which refers to a pre-processing mitigation approach from [115] to transform the world's properties from an "as-is" state to "should-be" referring to [2] which should be done in CRISP-DM 3.3.
[2] Synthetic data allows to generation of artificial data according to predefined properties or to replicate a given dataset as well as anonymize data according to GDPR or even leveraging companies' barriers to release their data for externals. In case of unbalanced or a lack of data, this technique enables analysts to generate it artificially. Potential tools would be Data Synthesizes[a], Synthetic Data Vault[b], Synthpop[c] or Gretel Synthetics[d].
[3] When attributes are assessed, previously defined assumptions and findings should be finally incorporated into the data dictionary.
[4] Identification of dependent features associated with the target is crucial [99] as well as proxy attribute detection, is required to identify dependent attributes which exemplarily is prescribed in [26] since features that are associated with both input and output can lead to biased estimates [99]. Utilising *FairTest*[e] can provide insights into correlations and the effects on the label.
[5] If not legally prohibited, exclusion of variables should be done reasonably depending on the value added for computation.

[a] *https://github.com/DataResponsibly/DataSynthesizer*, accessed 31/08/23
[b] *https://sdv.dev*, accessed 31/08/23
[c] *https://github.com/hazy/synthpop*, accessed 31/08/23
[d] *https://synthetics.docs.gretel.ai/en/stable*, accessed 31/08/23
[e] https://github.com/columbia/fairtest, accessed 12/09/23

2.3 Explore Data: Data Exploration Report

- Data exploration
  - Analyze properties of interesting attributes in detail (e.g., basic statistics, interesting sub-populations)
  - Identify characteristics of sub-populations considering specifically minority groups and define desired-/undesired outcomes across groups or among individuals and derive relevant attributes considered to be included within a fairness metric (individual or (sub-)group)
  - Extend and challenge defined use cases (objectives) according to best-, average- and worst case CRISP-DM 1.1.2
  - Consider arising biases[1]:
    - Modifiable Areal Unit Problem: If data is already aggregated, related attributes and -values of aggregated groups have to be assessed and challenged to be reasonable as well as ensure appropriate preservation of granularity
    - Simpson's Paradoxon: In relation to identified correlations in CRISP-DM 2.2 differing user characteristics in sub-groups have to be assessed as well as the effect of other (protected) attributes which needs to be challenged within the AI-RB to unveil context and identify potential vulnerabilities of (sub-)groups

    * Longitudinal Data Fallacy: Avoiding misleading interpretations based on aggregated views on data (cross-sectional vs. longitudinal) background assessment has to be performed: if timestamps are available, data (users) should be assessed to understand the context (and change) of cohorts over time whereas snapshots may provide limited insights in the data (users) [14]
    * Attribute Bias: Define decision boundaries and red lines (examples) for what is considered as (un-)fair and which attribute value combinations can cause an unfair outcome based on objectives defined in CRISP-DM 1.3.1. The defined examples shall be considered during the model assessment to avoid such explicit outcomes
    * Label bias: Check systematical distortions between subgroups and attribute-label combinations and compare/challenge the statistics e.g. do the differences in labels depend on a sensitive attribute? Is there another correlating or corresponding attribute combination that could explain the distortion in labels? (distinguish between potential discriminatory effects and business objectives based on economic reasons e.g. granting someone a loan should depend on the salary but not on the gender – which is related to Attribute Bias but reflected in the labels) [60][2]
    * Confounding bias: Analysts have to beware of subjective assumptions and to provide, replicable analysis and reasonable sensemaking about their conclusions according to data semantics which should be documented carefully and in the best case peer-reviewed

- Form suppositions for future analysis
  - Consider and evaluate information and findings in the data description report
  - Form a hypothesis and identify actions
  - Transform the hypothesis into a data mining goal, if possible
  - Clarify data mining goals or make them more precise. A "blind" search is not necessarily useless, but a more directed search toward business objectives is preferable.
  - Perform basic analysis to verify the hypothesis

[1] According to findings in Section 2.1, bias types and their emergence, we argue for investigation during "Data Exploration" in Phase 2.
[2] Jiang and Nachum [60] provides an approach to uncover and correct label bias by re-weighting without changing the labels.

2.4 Verify Data Quality: Data Quality Report

- Review keys, attributes
  - Check coverage (e.g., whether all possible values are represented) according to considerations made in CRISP-DM 2.2
  - Check keys
  - Verify that the meanings of attributes and contained values fit together
  - Identify missing attributes and blank fields
  - Establish the meaning of missing data
  - Check for attributes with different values that have similar meanings (e.g., low fat, diet) and consult domain experts if important semantic differences are among identified attribute values
  - Check spelling and format of values (e.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter)

- Check for deviations, and decide whether a deviation is "noise" or may indicate an interesting phenomenon
- Check for plausibility of values, (e.g., all fields having the same or nearly the same values)
- Evaluate identified risks from Phase 1 and update them according to gained knowledge from the data exploration and integrate them within the introduced RMS
- Check data quality criteria in AI-Act according to Article 10 – Data and Data Governance

5.2.3 Phase 3 – Data Preparation

3.1 Select data: Rationale for inclusion/exclusion

- Collect appropriate additional data (from different sources—in-house as well as externally)
- Perform significance and correlation tests to decide if fields should be included
- Reconsider Data Selection Criteria (see CRISP-DM 2.1) in light of experiences of data quality and data exploration (i.e., may wish include/exclude other sets of data)
- Reconsider Data Selection Criteria (see CRISP-DM 2.1) in light of experience of modelling (i.e., model assessment may show that other datasets are needed)
- Select different data subsets (e.g., different attributes, only data which meet certain conditions)
- Consider the use of sampling techniques (e.g., A quick solution may involve splitting test and training datasets or reducing the size of the test dataset if the tool cannot handle the full dataset. It may also be useful to have weighted samples to give different importance to different attributes or different values of the same attribute.)
- Document the rationale for inclusion/exclusion but be aware of Exclusion bias, consider utilization of multiple datasets whether potentially excluded attributes contribute positively to the results; assess the impact of attributes in evaluation Phase 4
- Check available techniques for sampling data and prevent Sampling bias by random sampling of (sub-)groups and consider further steps in CRISP-DM 4.2
- Good Idea: Based on Data Selection Criteria, decide if one or more attributes are more important than others and weight the attributes accordingly. Decide, based on the context (i.e., application, tool, etc.), how to handle the weighting.

3.2 Clean data: Data cleaning report

- Reconsider how to deal with any observed type of noise
- Correct, remove, or ignore noise such as Label bias which can be an issue in imbalanced datasets [23] and caused by e.g. Temporal bias or Behavioural bias [90]
- Decide how to deal with special values and their meaning. The area of special values can give rise to many strange results and should be carefully examined. Examples of special values could arise through taking the results of a survey where some questions were not asked or not answered. This might result in a value of 99 for unknown data. For example, 99 for marital status or political affiliation. Special values could also arise when data is truncated—e.g., 00 for 100-year-old people or all cars with 100,000 km on the odometer.

- Reconsider Data Selection Criteria (see CRISP-DM 2.1) in light of experiences of data cleaning (i.e., you may wish to include/exclude other sets of data).
- Good Idea: Remember that some fields may be irrelevant to the data mining goals and, therefore, noise in those fields has no significance. However, if noise is ignored for these reasons, it should be fully documented as the circumstances may change later.

3.3 Construct data: Derived attributes | Generated records

- Implement fairness [1]
  - Translate gathered fairness aspects from Phase 1 and Phase 2 into technical resp. mathematical metrics based on the objectives and defined scenarios
  - Decide which metric (Section 2.1.1) to use in accordance with the potential modelling technique defined in (CRISP-DM 4.1), document, why this metric is used and define thresholds and success criteria for successful utilization of the metric. State this also in the data dictionary
  - Based on the selected fairness approach (pre-, in- or post-processing), apply pre-processing approaches during this step (in-processing approaches in modelling (CRISP-DM 4.3) and post-processing approaches in model assessment (CRISP-DM 4.4))
- Construct data
  - Check available construction mechanisms with the list of tools suggested for the project
  - Decide whether it is best to perform the construction inside the tool or outside (i.e., which is more efficient, exact and repeatable)
  - Reconsider Data Selection Criteria (see CRISP-DM 2.1) in light of experiences of data construction (i.e., you may wish to include/exclude other sets of data)
  - Decide the necessity to balance the data according to considerations made in CRISP-DM 2.2 [1] (consider the modelling techniques to be used as well as utilized fairness metrics in CRISP-DM 4.1) preventing Representation bias i.e. if the data does not represent the population, balancing the data can be a solution; in case in-processing techniques are applied (weighting underrepresented instances) balancing may be not applied in this step[2]
- Derived Attributes
  - Decide if any attribute should be normalized (e.g., when using a clustering algorithm with age and income, in certain currencies, the income will dominate)
  - Consider adding new information on the relevant importance of attributes by adding new attributes (for example, attribute weights, weighted normalization)
  - How can missing attributes be constructed or imputed? [Decide type of construction (e.g., aggregate, average, induction).]
  - Add new attributes to the accessed data
  - Good Idea: Before adding Derived Attributes, try to determine if and how they ease the model process or facilitate the modeling algorithm. Perhaps "income per person" is a better/easier attribute to use than "income per household." Do not derive attributes simply to reduce the number of input attributes.
  - Good Idea: Another type of derived attribute is the single-attribute transformation, usually performed to fit the needs of the modeling tools.
  - Challenge derived attributes' contribution to increasing the level of information to prevent Cause-Effect bias e.g. by introducing an explicit proxy

- Single-attribute transformations
  - Specify necessary transformation steps in terms of available transformation facilities (for example, change a binning of a numeric attribute)
  - Perform transformation steps
  - Good Idea: Transformations may be necessary to change ranges to symbolic fields (e.g., ages to age ranges) or symbolic fields ("definitely yes," "yes," "don't know," "no") to numeric values. Modeling tools or algorithms often require them.
- Generated Records
  - Check for available techniques if needed (e.g., mechanisms to construct prototypes for each segment of segmented data).
- Dealing with potential biases:
  - Historical Bias: If indicated in business objectives/goals, in existing solutions or in CRISP-DM 2.1, randomization techniques can be applied to avoid historically biased data shown in [89]
  - Social bias: Apply objective and reasonable decision-making during the data preparation process and document every decision and transformation of the data to keep the process transparent and traceable for further adaptions or refinements

---

[1] The determination and selection of an appropriate fairness metric should be accomplished within the AI-RB to discuss all potential influencing factors according to [20] including responsible data analysts providing insights along the created data dictionary. According to the high number of fairness metrics, not every metric does the intended job for a given purpose, is compatible with another one, is applied at the same stage and has different trade-offs when it comes to the overall performance of the system [20, 62].

[2] Imbalances in classes can negatively impact the performance in classification tasks, especially if minority classes do not represent certain (minority) groups:
Example 1: When detecting brain tumours, we do not want to reflect the ratio of positive/negative examples in reality, preventing the algorithm to learn this ratio. Additionally, this is a very crucial domain which requires very high accuracy and lowest FNR.
Example 2: When predicting customer behaviour, where the data population does not reflect the real population (Representation Bias) further steps are required: investigating the ground truth, distributions and ratios; over- or undersampling; a collection of additional data; creation of synthetic data [23]. What to choose depends on the objective, the level of imbalance and the model where Brownlee [23] provided answers such as the utilization of particular performance metrics, resampling methods, the utilization of Microsoft Azure AI SMOTE[a] (Synthetic Minority Over-sampling TEchnique), different utilization of algorithms or penalized models.

[a] *https://azure.microsoft.com/de-de/solutions/ai*, accessed 31/08/23

3.4 Integrate Data: Merged data

- Check if integration facilities are able to integrate the input sources as required
- Integrate sources and store results
- Prevent Aggregation bias by combining only comparable (sub-)groups
- Reconsider Data Selection Criteria (see CRISP-DM 2.1) in light of experiences of data integration (i.e., you may wish to include/exclude other sets of data)

3.5 Format Data: Reformatted data

- Rearranging attributes
- Reordering records
- Reformatted within-value
  - These are purely syntactic changes made to satisfy the requirements of the specific modelling tool
  - Reconsider Data Selection Criteria (see CRISP-DM 2.1) in light of experiences of data cleaning (i.e., you may wish to include/exclude other sets of data)

CRISP-DM documentation is sufficient.

5.2.4 Phase 4 – Modelling

4.1 Select modeling technique: Modeling technique | Modeling assumptions

- Select modeling technique
  - Consider if the technique is compatible with the chosen fairness metric and the overall fairness goal defined in CRISP-DM 3.3[1]
  - Consider obligations to the AI-Act:
    - Consider security aspects as required for Article 15 – Accuracy, Robustness and Cybersecurity par. (1, 4)
    - Consider human oversight aspects as required for Article 14 – Human Oversight and assess the compatibility with the selected modelling technique
    - Consider potential logging-techniques required in Article 12 – Record Keeping and the compatibility with the selected modelling technique[2]
    - Consider transparency obligations according to Article 13 – Transparency and Provision of Information methods of explainability and interpretability along considerations made in CRISP-DM 1.3.2 and align to suitable modelling techniques with respect to defined steps in Article 12, 13 and 15
  - Apply in-processing fairness approaches
- Modeling assumptions
  - Define any built-in assumptions made by the technique about the data (e.g., quality, format, distribution)
  - Compare these assumptions with those in the Data Description Report

- Make sure that these assumptions hold and go back to the Data Preparation Phase, if necessary

[1] As described in Section 2.1.2 the utilization of fairness metrics depend on the modelling techniques which should be aligned right at the beginning of Phase 4. Mehrabi et al. [76, pp. 13–15] provided a list of literature resources, dealing with bias according to different domains of AI algorithms.
[2] Therefore we provide additional questions: Does the logging meet privacy obligations? Does the logging have notable risks or side effects? Is the log entry meaningful? Which logging technique is suitable? Does the model/modelling technique have an impact on logging capabilities? Are there specific circumstances e.g. encryption and how can logging be achieved then? How can it be traced accordingly?

4.2 Generate test design: Test design

- Check existing test designs for each data mining goal separately
- Decide on necessary steps (number of iterations, number of folds, etc.)
  - In case of imbalanced data, consider an appropriate split technique (e.g. stratified k-fold cross-validation) [23] and the impact on the performance metric (e.g. accuracy can become unreliable if the class distribution is fierce [23]) to prevent Sampling bias
  - Assess the number of positive- and negative samples among protected attributes or (sub-)groups to ensure appropriate balance between labelled outcomes to prevent Negative Set bias
- Prepare data required for test

4.3 Build model: Parameter settings | Models | Model description

- Parameter settings
  - Set initial parameters[1]
  - Document reasons for choosing those values
- Models
  - Run the selected technique on the input dataset to produce the model
  - Post-process data mining results (e.g., edit rules, display trees)
- Model description
  - Describe any characteristics of the current model that may be useful for the future
  - Record parameter settings used to produce the model
  - Give a detailed description of the model and any special features
  - For rule-based models, list the rules produced, plus any assessment of per-rule or overall model accuracy and coverage as required for Article 15 – Accuracy, Robustness and Cybersecurity par. (2, 3)

- For opaque models, list any technical information about the model (such as neural network topology) and any behavioural descriptions produced by the modeling process (such as accuracy or sensitivity)
- Describe the model's behaviour and interpretation; Consider to implement explainability methods (SHAP, Lime etc.)[2] and provide insights on interpretability with dedicated plots (PDP-Plot, ICE-Plot, ALE-Plots)[3] on the impact of certain attributes to prevent Attribute Bias[3]; as required for information provision in Article 13 – Transparency and Provision of Information par. (4)
- State conclusions regarding patterns in the data (if any); sometimes the model reveals important facts about the data without a separate assessment process (e.g., that the output or conclusion is duplicated in one of the inputs)

---

[1] Streamline the desired experiments utilizing *Weights and Biases*[a] toolkit.
[2] Understanding the model's output respective to the output's composition can be a crucial part when it comes to bias, fairness and resulting discrimination since we claim traceability. Explainability can be one step to assess the model's output according to causal relations, reasonability and finally ensuring creditability i.e. why the prediction was computed as is. This may help to prevent the model from considering protected attributes and unveil proxy variables as well. We distinguish between local- and global explainability methods [78]:
SHAP[b] *(SHapley Additive exPlanations)* as a game theory based explainability method for any machine learning model can be used to describe a model's output locally (impact of variables on a decision: waterfall and force graph) and globally (describes the general impact (ratio) of considered variables). Further options are TreeSHAP (for tree-based ML models) and DeepShap (a combination of SHAP and DeepLift for deep learning models);
Lime[c] *(Local Interpretable Model-Agnostic Explanations)* generates explanations for the different machine learning model's behaviour. Lime supports explanations for text and table-based data or images.

[3] Interpretability methods leverage the understanding of machine learning models unveiling how a prediction was computed: This can be realized by **PDP-Plots** (showing the effect one or two features have on the predicted outcome), **ICE-Plot** (PDP-Plot but for individual data points), **ALE-Plot** (how features influence the prediction on average) [78]

[4] A classifier's output decision should be the same across sensitive characteristics, given what the correct decision should be and therefore a "classifier, $f\Theta(x)$, is biased if its decision changes after being exposed to additional sensitive feature inputs. In other words, a classifier is fair with respect to a set of latent features(z), if: $f\Theta(x) = f\Theta(x, z)$." [4]

[a] https://wandb.ai/site, accessed 12/09/23
[b] https://shap.readthedocs.io/en/latest/generated/shap.Explainer.html, accessed 05/09/23
[c] https://github.com/marcotcr/lime, accessed 05/09/23

4.4 Assess model: Model assessment | Revised parameter settings

- Model assessment
  - Prevent Evaluation bias: Ensure benchmark data represents the use population; instead of using aggregated measures only, evaluate subgroup performance and compare potential trade-offs between performance metrics; apply multiple metrics and confidence intervals or utilize targeted data augmentation [101]
  - Prevent Presentation bias: Ensure data type and chosen visualization is compatible to avoid wrong conclusions based on skewed or unsuited type of visualization
  - Check Algorithmic Bias: Which is purely added by the algorithm but often interchangeably used with attribute bias in the literature such as negative impact towards a certain attribute [1]
  - Evaluate results with respect to evaluation criteria
  - Test result according to a test strategy (e.g.: Train and Test, Cross-validation, bootstrapping, etc.)
  - Compare evaluation results and interpretation
  - Create ranking of results with respect to success and evaluation criteria
  - Select best models
  - Apply post-processing fairness approaches [2]
  - Evaluating fairness criteria according to success criteria defined in CRISP-DM 3.3 along global (data mining and business) success criteria in Phase 1 and Phase 2. Assess performance among defined (sub-)groups, among defined (protected) attributes according to global model performance. Apply intersectional evaluation (performence resp. subgroup intersection) or disaggregated evaluation (performance resp. different subgroups) [4]
  - Interpret results in business terms (as far as possible at this stage)
  - Get comments on models by domain or data experts to prevent Confounding bias and Cause-Effect bias
  - Check plausibility of model
  - Check effect on data mining goal
  - Check model against given knowledge base to see if the discovered information is novel and useful
  - Check reliability of result
  - Analyze potential for deployment of each result
  - If there is a verbal description of the generated model (e.g., via rules), assess the rules: Are they logical, are they feasible, are there too many or too few, do they offend common sense?
  - Assess results; provide information to meet Article 13 – Transparency and Provision of Information par. (3)
  - Get insights into why a certain modeling technique and certain parameter settings lead to good/bad results
  - Good Idea: "Lift Tables" and "Gain Tables" can be constructed to determine how well the model is predicting.
- Revised parameter settings
  - Adjust parameters to produce better models.

[1] According to Danks and London [32], algorithmic bias strongly depends on several frame conditions which results in a lack of guidance to mitigate algorithmic bias. We propose to test on Attribute bias e.g. contractual fairness approach.
[2] Once the model is created, post-processing metrics can be applied to the model to assess bias and fairness.

5.2.5 Phase 5 – Evaluation

5.1 Evaluate results: Assessment of data mining results with respect to business success criteria | Approved models

- Assessment of data mining results with respect to business success criteria
  - Understand the data mining results
  - Interpret the results in terms of the application
  - Check effect on for data mining goal
  - Check the data mining result against the given knowledge base to see if the discovered information is novel and useful
  - Evaluate and assess results with respect to business success criteria (i.e., has the project achieved the original Business Objectives); Consider a neutral and objective assessment of the results and validate (sub-)groups as well as individuals equally to avoid Social Bias and Framing-Effect bias. Consider causal and correlating aspects to prevent Cause-Effect bias and challenge results according to the initial questions for reasonable sensemaking to prevent Confounding bias
  - Compare evaluation results and interpretation
  - Rank results with respect to business success criteria
  - Check effect of result on initial application goal
  - Determine if there are new business objectives to be addressed later in the project, or in new projects
  - State recommendations for future data mining projects
  - Gather information about transparency and provision of information required in Article 13 – Transparency and Provision of Information (par. 3)
- Approved models
  - After accessing models with respect to business success criteria, select and approve the generated models that meet the selected criteria.

5.2 Review process: Review of process

- Review of process
  - Provide an overview of the data mining process used
  - Analyze the data mining process. For each stage of the process ask:
  - Was it necessary?
  - Was it executed optimally?
  - In what ways could it be improved?

- Identify failures
- Identify misleading steps
- Identify possible alternative actions and/or unexpected paths in the process
- Review data mining results with respect to business success criteria
- Check on compliance with defined fairness criteria according to applied metrics and thresholds defined in business objectives
- Identify residual risks and document them in the RMS ensuring compliance to Article 9 – Risk Management System

5.3 Determine next steps: List of possible actions | Decision

- List of possible actions
  - Analyze the potential for deployment of each result
  - Estimate potential for improvement of current process
  - Check remaining resources to determine if they allow additional process iterations (or whether additional resources can be made available)
  - Recommend alternative continuations
  - Refine process plan
- Decision
  - Rank the possible actions
  - Select one of the possible actions
  - Document reasons for the choice also in the context of bias and fairness metrics[1]

[1] Ensure knowledge gain out of utilized fairness metrics and prepare documentation for monitoring and updates in Phase 7 – Monitoring and Maintenance.

5.2.6 Phase 6 – Deployment

6.1 Plan deployment: Deployment plan

- Deployment plan
  - Summarize deployable results
  - Develop and evaluate alternative plans for deployment
  - Decide for each distinct knowledge or information result
  - Determine how knowledge or information will be propagated to users
  - Decide how the use of the result will be monitored and its benefits measured (where applicable); according to considerations made in Phase 7
  - Decide for each deployable model or software result
  - Establish how the model or software result will be deployed within the organization's systems; ensure (as far as possible) the system (model) deployed is utilized according to its purpose along defined business objectives to avoid Deployment bias

- Determine how its use will be monitored and its benefits measured (where applicable) and how the system can be interrupted (in case of harmful outcomes) by human instances to fulfil obligations in Article 14 – Human Oversight
- Identify possible problems during deployment (pitfalls to be avoided)

6.2 Plan monitoring and maintenance: Monitoring and maintenance plan

- Monitoring and maintenance plan
  - Check for dynamic aspects (i.e., what things could change in the environment?)
  - Decide how accuracy will be monitored
  - Determine when the data mining result or model should not be used any more. Identify criteria (validity, threshold of accuracy, new data, change in the application domain, etc.), and what should happen if the model or result could no longer be used. (update model, set up new data mining project, etc.) as required in Article 15 – Accuracy, Robustness and Cybersecurity par. (2, 3)
  - Will the business objectives of the use of the model change over time? Fully document the initial problem the model was attempting to solve.
  - Develop monitoring and maintenance plan.

6.3 Produce final report: Final report | Final presentation

- Final report
  - Identify what reports are needed (slide presentation, management summary, detailed findings, explanation of models, etc.)
  - Analyze how well initial data mining goals have been met
  - Identify target groups for report
  - Outline structure and contents of report(s)
  - Select findings to be included in the reports
  - Write a report specifically including the process of considering fairness [1]
  - provide "instructions of use" according to Article 15 – Accuracy, Robustness and Cybersecurity par. (2)
- Final presentation
  - Decide on target group for the final presentation and determine if they will already have received the final report
  - Select which items from the final report should be included in final presentation

[1] Include all relevant information according to made decisions, constraints/conditions and further justify why they have been made; why certain fairness metrics were chosen (e.g. advantages and disadvantages from other metrics, importance of error-types); which edge cases are crucial; which trade-offs have been decided and which arguments were discussed; which departments/single-point-

of-contacts have to be considered when adaptions are made; why and which thresholds are considered as sufficient; which use cases are forbidden and why?

6.4 Review project: Experience documentation

- Experience documentation
  - Interview all significant people involved in the project and ask them about their experience during the project
  - If end users in the business work with the data mining result(s), interview them: Are they satisfied? What could have been done better? Do they need additional support?
  - Summarize feedback and write the experience documentation
  - Analyze the process (things that worked well, mistakes made, lessons learned, etc.)
  - Document the specific data mining process (How can the results and the experience of applying the model be fed back into the process?)
  - Internalization of outcomes or learned lessons with respect to bias and fairness. Which ethical/fairness/data insights are crucial to incorporate outside the project? [1]
  - Generalize from the details to make the experience useful for future projects

[1] Which insights/definitions/outcomes can be used as a guideline for strategic/-operative business development as well as further projects or IT department (data management etc.)? Which adaptions according to the process model should be incorporated for future projects?

5.2.7 Phase 7 – Monitoring and Maintenance

The illustration of the introduced Monitoring and Maintenance (MM) phase is highlighted in red since tasks, output and activities are not part of the original CRISP-DM by Chapman et al. [27]. Therefore green text refers to considerations according to bias relevant aspects whereas blue textmarks requirements along the AI-Act.

7.1 Monitor: Metrics; Risks and Compliance Check

- Continuous input assessment[1]
  - Monitor **(operational) input data** and compare statistics (distributions, correlations, labels etc.) from training data; consider assumptions and actions considered in CRISP-DM 6.2 [100]
  - Define boundaries and thresholds that are considered leading to erroneous output and trigger appropriate (update-)notifications if they are exceeded (automatically or manually) [100]; consider reasons (why) and effects (what) leading to data-drift to identify temporary Emergent bias; in case online data is gathered, the interface could lead to bias such as User-Interaction bias; in case humans are interacting with the system and add or manipulate data manually which is incorporated in the data collection (during operation) potential emerging Social bias
  - Evaluate related risks and estimate costs of updates along the defined boundaries

  - Utilize libraries or tools to leverage (automated) assessment e.g. automated data validation tools e.g. *Deequ*[92] to analyse and compare datasets or *Microsoft Azure ML DataDriftDetector Class*

- Continuous model, bias, fairness and output assessment
  - Monitor the model's performance metrics (precision, recall, specificity/sensitivity, F-Score, (balanced) accuracy etc.) from training [26, 9]; consider assumptions and actions in CRISP-DM 6.2
  - Define boundaries and thresholds of performance metrics (precision, recall, specificity/sensitivity, F-Score, (balanced) accuracy etc.) that are considered to be unacceptable (e.g. defined success criteria in CRISP-DM 5.1) leading to updates and trigger appropriate notifications if they are exceeded (automatically or manually) and document them according to Article 12 – Record Keeping respective Article 61: Post-market monitoring system
  - Monitor fairness metrics' performance along defined boundaries and thresholds (CRISP-DM 1.3.2, CRISP-DM 2.3 and CRISP-DM 4.4) e.g. implement (automated) assessment of fairness performance (e.g. counterfactual fairness
  - Monitor output

- Continuous risk assessment according to Article 9 – Risk Management System
  - Update validity of risks e.g. model- and data drift/shift, data quality, performance- and fairness metrics
  - Update validity of legal risks e.g. updated obligations or requirements, ongoing lawsuits or issues within the domain/community
  - Update risks of organizational risks e.g. change of personnel, departments, leadership, organizational goals – strategic or operative

- Continuous compliance assessment
  - Check requirements along Article 10 – Data and Data Governance
  - Check requirements along Article 11 – Technical Documentation in case of updates
  - Check requirements along Article 12 – Record Keeping
  - Check requirements along Article 13 – Transparency and Provision of Information
  - Check requirements along Article 14 – Human Oversight
  - Check requirements along Article 15 – Accuracy, Robustness and Cybersecurity(2)(3)

[1] As suggested by Studer et al. [100] input has to be monitored according to meet statistics discovered in training triggering a notification if those change. This process can be automated by applying dedicated libraries [92].

7.2 Maintenance: Model and Documentation

- Consider hard- and software updates on the productive systems, different versions of utilized tools and libraries or other internal and external side-effects that influence the performance
- Update and re-train
  - Update and collect new data (consider labelling process) [100]; consider appropriate amount of positive- and negative labeled data to prevent Negative Set bias
  - Re-train the model (according to changed data distribution) by fine-tuning the existing model to the new data [100]; be aware of Algorithmic bias leading to potential Attribute bias
  - Re-evaluate the re-trained model and ensure a better performance [100]
  - Consider positive- or negative feedback-loops e.g. when computations are left out the model and therefore not part of the dataset used for re-training [100] to prevent Automation bias
- Deployment
  - Consider a fallback scenario minimizing the risk of erroneous models [100]
  - Consider automation such as continuous training and -deployment, including automated validation ensuring quality gates [16]
- Compliance and Documentation
  - Update and document findings in the data dictionary

5.3 Bias- and Fairness Toolkits

In Table 5.2 we introduce tools and libraries dealing with bias. Developed by commercial organizations as well as the research community, they aim to leverage to uncover bias in the data and the model by providing several pre-, -in and post-processing techniques, applicable to a broad range of use cases. Nevertheless, the utilization of the tools is no guarantee for bias-free data or algorithms but shall ease the way of unveiling and assessing potential biases.

Tool Name [*rel.Lit.*]	**Description and Focus**
AIF360[1] [1, 18]	Comprehensive open-source toolkit including metrics to test for biases, explain those metrics and mitigate bias in data and models utilizing pre-, in- and post-processing techniques [26].
Aequitas[4]	Open source bias audit toolkit incorporating parity- and confusion-based metrics assessing (sub-)groups of a protected attribute. It leverages the identification of suitable metrics for a particular situation [26, 77, 18].
Amazon SageMaker Clarify [51]	Detecting bias in data (group balance and label distribution) and models utilizing explainability and reports as well as pre- and post-processing techniques[2].
Fairlearn[3]	Open source Python framework producing fair model focussing on group fairness by focussing on parity-based measures. Fairlearna[3] contains post-processing (transformation) and reduction (re-weighting) algorithms and comprises a dashboard and a set of algorithms in binary classification and regression [26].
FairPut[13]	Provides an approach to implement fair outputs at the individual and group level and provides pre- and post-processing methods. the aim is to enhance interpretability, robustness and fairness while ensuring a reasonable level of accuracy [97].
FairTest[5]	Python toolkit identifying unwarranted associations between an algorithm's outcomes and (sensitive) user attributes [26, 103, 87, 44].
FairnessMeasures[6]	Quantifies definitions of discrimination in classification by testing the model on different data Caton and Haas, Friedler et al.
auditAI[7]	Open Source Python tool for bias testing according to comparisons to other known populations i.e. testing with respect to a "ground-truth".
ML-fairness-gym[8]	Based on Open AI's "gym" framework, fairness-gym[9] simulates the long-term outcomes of fairness measures in a model.
REPAIR[12][69]	Mititages Representation bias by reweighting the underlying dataset and penalizing the classifier.
themis[11] [5, 13]	Open source bias toolbox for measuring potential discrimination focussing on causality in discriminatory behaviour and group-based fairness within binary classification. The approach is based on pre- and post-processing techniques [46, 5, 87, 18, 44].
What-If Tool[10]	Provides a dashboard for analyzing black-box classification and regression models by assessing the model's behaviour and output via various inputs. It includes interactive features (editing data points, searching for nearest counterfactuals) and techniques for assessing fairness and performance among different subgroups. Data types are suited to categorical and image classification applied as post-processing techniques.

Table 5.2: Identified tools to detect and mitigate bias or/and ensure fairness

[1] https://github.com/Trusted-AI/AIF360, accessed 05/09/23
[2] https://aws.amazon.com/de/sagemaker/clarify, accessed 12/09/23
[3] www.fairlearn.org, accessed 12/09/23 [3a] https://tinyurl.com/8jcx93ev, accessed 12/09/23
[4] https://dssg.github.io/aequitas/, accessed 12/09/23
[5] https://github.com/columbia/fairtest, accessed 12/09/23
[6] https://fairnessmeasures.github.io, accessed 12/09/23
[7] https://github.com/pymetrics/audit-ai, accessed 12/09/23
[8] https://github.com/google/ml-fairness-gym, accessed 12/09/23
[9] https://tinyurl.com/49jubps5, accessed 12/09/23
[10] https://pair-code.github.io/what-if-tool, accessed 19/09/23
[11] https://github.com/LASER-UMASS/Themis, accessed 19/09/23
[12] https://github.com/JerryYLi/Dataset-REPAIR, accessed 19/09/23
[13] https://github.com/firmai/ml-fairness-framework, accessed 19/09/23

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion

In this section, we summarize our findings and conclusions related to the research questions stated in Section 1.4 followed by identified gaps leading to future work in Section 6.2.

> *RQ1: Which types of bias are described in the literature, where do they manifest and how (according to the method) can bias be unveiled?*

As shown in Chapter 3 we identified **60 different bias types** within the mapping study performed in Section 2.1. Considering further investigations along the context of this thesis including cognitive biases, more than 260 different bias types were gathered. The types refer to subjects such as participating actors and stakeholders which can be biased in a cognitive manner and objects such as different types of data, algorithms, domains, contexts and use cases. Therefore, they manifest in (i) the real world per se such as Historical bias reflecting cultural socio-economic and socio-technical peculiarities (ii) the involved actors processing the data (iii) the data, domain and the way it is collected, manipulated, interpreted and processed (iv) the algorithm and the way of recognizing patterns and calculating predictions and (v) the interaction during operation and how predictions are utilized and incorporated.

We noticed, that different (sub)domains utilize derived definitions of similar bias types according to their specific requirements in their domain and therefore establish fine granular definitions of additional bias types according to their field, context or data. Based on identified key papers, multiple definitions of bias types with the same name are defined differently not because of a diverging domain, but the different focus and viewpoint of the studies e.g. Selection Bias and Sampling Bias. This mismatch between

the definitions hampers a clear and streamlined method of bias detection and potential mitigation strategies.

The methods to unveil bias comprise (i) human capabilities such as sensemaking by identifying causal relationships and socio-economic resp. socio-technical reasoning (ii) mathematical and statistical reasoning by investigating correlations and distributions and (iii) technical capabilities by implementing explainability, and interpretability as well as profound knowledge in peculiarities of modelling techniques and how to adapt them according to the required use.

> *RQ2: At which steps in the process model can bias be identified and which steps are necessary to do so?*

As mentioned in RQ1, the emergence of bias stems from the reflection of bias within the society incorporated in the development process, affecting throughout the use and application of the system. Therefore bias can be identified at every single step within CRISP-DM throughout to the operation. Since different bias types have different causes, they differ in methods and tools to identify, deal with and mitigate which is shown in Chapter 5. As mentioned in RQ1 the necessary steps to identify bias comprise human-related aspects, mathematical and statistical capabilities as well as technical finesse. Moreover, the research community and big (commercial) tech players created tools, libraries and platforms to assess data according to bias considering different approaches.

> *RQ3: To what extent can the additional process steps introduced satisfy and address the requirements set forth in the articles of the EU AI-Act touching on bias-relevant aspects?*

By incorporating findings from Chapter 4 in Chapter 5 we showed, that the basic CRISP-DM per default covers a broad aspect of the AI-Act, but on a general level: CRISP-DM do consider assessing legal dependencies as a task, but specific guidelines and action items are missing. This is reasonable with respect to the intended aim of data mining more than twenty years ago without considering emerging legal requirements such as GDPR. This argument is valid for technical developments in the field of AI and its vulnerability in terms of security and robustness, dealing with attacks such as data-poisoning or model-stealing which became important due to the broad utilization of AI systems in the economy. Missing aspects are proper data governance strategies, logging capabilities to assess provided decisions and human oversight measures all required within the AI-Act. Therefore we proposed mapped steps in CRISP-DM to consider appropriate steps to implement these missing aspects. CRISP-DM as a data mining process does hardly consider state-of-the-art AI techniques that incorporate integrative pipelines such as automated continuous development and integration which led us to introduce the Monitoring and Maintenance (MM) phase as "Phase 7" attached to CRISP-DM to ensure proper operation and compliance to the AI-Act's obligations on monitoring the system

during operation. In Chapter 7 we provide a final overview of the detailed mapping between identified bias types, the tasks in CRISP-DM and available tools and literature available in the context of this thesis.

Concluding Remarks In Chapter 3 we argued that bias[1] does not necessarily lead to unfairness or discrimination (Section 2.1.1) by assessing a specific bias type (Section 3.3) according to exemplary use cases (Section 3.1) and whether the result is discriminatory or not, depends on the attribute's semantics. This aligns with findings by Alelyani who claims that bias does not imply unfairness because it depends on the underlying data (characteristics) affected by humans, which is reflected by the model that is used to learn patterns and relations from the data since data is biased by nature due to cognitive bias of human brains [3].

Which level of bias is acceptable? As described in Chapter 5 we argue for comprehensive testing on different types of bias to uncover and understand potential issues in the data as well as the model's outputs to increase overall performance and prevent harmful effects. Therefore the goal of total fairness and unbiasedness "may be an impossibly high computational and statistical hurdle for any algorithm" [30] especially "if we want to mimic human intelligence, we need to accept bias" [3]. This underlines findings from Wachter, Mittelstadt, and Russell who claims that no system can be perfectly calibrated and future-proof since AI systems using real-world data incorporate a small amount of bias [108]. Therefore we argue, that removing bias completely is beyond reach, but to prevent worst-case scenarios, a considerate contemplation of use cases, domains, contexts and risks along with profound bias testing along our process can help to avoid worst-case scenarios and increase performance. Moreover, a considerate definition of the problem has to be stated, assessing the effect and outcome of defined business objectives weighting the balance between business objectives and fairness [99].

6.2 Future Work

Including domain-specific requirements: In-depth analysis and mapping of different domain-specific needs such as identification of bias types in recommender systems or natural language processing, as different data-specific needs such as image processing or survey data would complement this work. A starting point could be the work from Chen et al. [28], who describes a broad overview of bias specific to recommender systems including mitigation strategies and related literature. Therefore one big picture could be to create "layers" (e.g. the "recommender systems layer" as described previously) that focus on the described aspects to create a modular version of CRISP-DM, that can be enriched by adding specific guidelines according to the organization's needs and focus.

[1] In the notion defined in Section 2.1 as a *"systematic difference of treatment . . . "* (objects, people or groups) *"... in comparison to others where treatment is any kind of action"* (perception, observation, representation, prediction or decision) [58].

Unveil potential gaps in regulatory obligations: The U.S. as well as China proposed some guidelines and regulations of AI in their countries e.g. China unveiled regulation on recommendation algorithms in 2021[2] and the U.S. created an *AI Bill of Rights*[3]. Therefore a comparison of international regulatory requirements in the AI domain could obtain information about similarities or potential pitfalls during utilization, development and production. Besides the different cultural values among these players, the identified differences could further be mapped to CRISP-DM which would leverage the compliant development for global acting organizations.

Emphasis on dedicated AI-Act Articles: Since the focus of this work is limited to bias-related aspects, we consider a gap between the dedicated articles and specific guidelines to ensure compliance beside bias-related aspects e.g. we would appreciate for providing dedicated guidelines, solutions and tools for record-keeping and logging (Section 4.1.4) or human oversight measures (Section 4.1.6).

Taxonomy and clustering of bias types: As shown in Section 2.1, specific bias definitions depend on the context and domain in which they are considered, which leads to the inheritance of sub-bias-types justified just by semantic difference. We suggest investigating similarities in techniques uncovering bias and whether a dedicated technique is able to unveil more than one specific bias type at once e.g. by applying dedicated statistical methods. This would allow analysts and practitioners to detect different types of bias by a few dedicated tests.

Bridging the gap between CRISP-DM and ISO: As mentioned in Section 2.4, the ISO aims to standardize terms and processes among the AI lifecycle. Therefore an assessment of emerging standardization and its effect on established development processes could unveil the fitness of an organization's utilized processes and process models for ISO standardization.

Effectiveness of mitigation strategies according to specific bias types: Yet, some bias types are defined and described without providing answers or actions to unveil or mitigate those. A mathematical investigation of dedicated bias types could be performed to deliver metrics and techniques to do so.

Mapping between bias types and stakeholders: As shown, different bias types are dependent on different factors such as data bias (e.g. reflected in the real world), cognitive bias (e.g. different treatment according to human beliefs) or bias related to the analyst (e.g. Sampling bias). An investigation resp. mapping of identified bias types according to stakeholders within the development process could be established to increase the awareness among factors of influence in the context of bias and increase focus on bias mitigation by design.

[2] https://tinyurl.com/33ur6ur3, accessed 11/09/23
[3] https://www.whitehouse.gov/ostp/ai-bill-of-rights, accessed 11/09/23

CHAPTER 7

Appendix

CRISP-DM Phase	**Phase 1**				**Phase 2**				**Phase 3**					**Phase 4**				**Phase 5**			**Phase 6**				**Ph. 7**		
Task / Bias Type	Determine Business Objectives	Assess Situation	Determine Data Mining Goals	Produce Project Plan	Collect Intital Data	Describe Data	Explore Data	Verify Data Qualitay	Select Data	Clean Data	Construct Data	Integrate Data	Format Data	Select Modeling Techniques	Generate Test Design	Build Model	Assess Model	Evaluate Results	Review Process	Determine Next Steps	Plan Deployment	Plan Monitoring and Maintenance	Produce Final Report	Review Project	Monitor	Maintenance	Tools and Techniques
Historical Bias 1	✓		✓		✓						✓																
Population Bias 2					✓																						
Behavioural Bias 3					✓					✓																	
Content Production Bias 4					✓																						
Temporal Bias 5					✓					✓																	
Social Bias 6					✓						✓							✓							✓		
Selection Bias 7					✓																						
Negative Set Bias 7.A															✓											✓	
Framing-Effect Bias 8	✓		✓		✓													✓									
Representation Bias 9											✓																REPAIR
Measurement Bias 10					✓																						
Device Bias 10.A																											
Capture Bias 10.B																											
Aggregation Bias 11												✓															
Modifiable Areal Unit Probelm Bias 11.A							✓																				
Simpson's Paradoxon 11.B							✓																				
Longitudinal Data Fallacy 12							✓																				
Sampling Bias 13									✓						✓												
Exclusion Bias 14						✓			✓																		
Attribute Bias 15			✓			✓	✓									✓										✓	see Section 5.3
Label Bias 16							✓			✓																	[60]
Cause-Effect Bias 17						✓					✓						✓	✓									
Confounding Bias 17.A							✓										✓	✓									
Algorithmic Bias 18																	✓									✓	AIF360 [93, 4]
User-Interaction Bias 19					✓																				✓		
Presentation Bias 19.A	✓					✓											✓										
Emergent Bias 19.B																									✓		
Evaluation Bias 20																	✓										
Deployment Bias 21																					✓						
Automation Bias 22																										✓	

Table 7.1: Detailed overview of mapped bias types from Chapter 3 to CRISP-DM [27].

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence. xiii, 1, 7, 24

AI-Act European Union Artificial Intelligence Act. xi, xiii, 2, 5, 7, 12, 21, 26, 53, 57, 63–65, 90, 105

AI-RB Artificial Intelligence - Risk Board. 64, 65, 68, 71, 72, 75, 78, 82

API Application Programming Interface. 76

CRISP-DM Cross Industry Standard Process Model for Data Mining. xi, xiii, 7, 14, 19

DM data mining. 19

EU European Union. xiii, 5, 17

GDPR General Data Protection Regulation. 12, 16, 26, 58, 64, 65, 70, 74, 78, 96

ISO International Organization for Standardization. 10, 12, 15, 21, 54, 98

MM Monitoring and Maintenance. 21, 51, 54, 55, 61, 90, 96, 105

RMS Risk Management System. 54, 55, 57, 65, 70, 80, 88

Glossary

data dictionary is a "collection of names, definitions, and attributes about data elements that are being used or captured in a database, information system, or part of a research project. It describes the meanings and purposes of data elements within the context of a project and provides guidance on interpretation, accepted meanings and representation. A data dictionary also provides metadata about data elements. The metadata included in a data dictionary can assist in defining the scope and characteristics of data elements, as well the rules for their usage and application" [1]. 56, 58, 65, 68, 74, 75, 77, 82, 92

explainability concentrates on providing justifications for individual predictions made by a model i.e. why does the model compute this output for the particular input. 46, 48, 58, 71, 83, 85, 96

human in the loop (HITL) refers to a mode of human-computer interaction where humans are actively and directly involved in a computational process or decision-making loop alongside an AI system. In HITL systems, the human and AI collaborate in real-time, with the human providing input, guidance, validation, or oversight as part of the overall workflow. 59

human oversight is the practice of having humans actively monitor and supervise the actions and decisions made by artificial intelligence systems. It is an important element in the development and deployment of AI technologies, especially in situations where AI systems can have significant consequences, ethical implications, or safety concerns. 59

interpretability concerns the overall behaviour within the model's decision-making process to make it transparent and understandable i.e. how does the model work in general. 46, 58, 71, 83, 85, 96

protected attribute refers to characteristics or features of individuals that are considered sensitive or potentially discriminatory. These attributes are typically protected by laws and regulations to prevent unfair or biased treatment as described in Section 2.1. 65

[1] https://library.ucmerced.edu/data-dictionaries, accessed 27/09/23

protected class is a group of individuals who share a common characteristic that is legally protected from discrimination by laws, regulations, and ethical guidelines and usually related to protected attribute(s). In the context of AI systems, especially marginalized groups should be considered. 65

proxy variable a variable that is used as a substitute or stand-in for a different variable of interest. Proxy variables are often used when the original variable might be difficult to measure directly or might not be available in the dataset, but the proxy variable is often correlated with or indicative of the original variable and therefore has the capability to identify the original variable as described in Section 2.1. 10, 32, 45

risk management system (RMS) is a "structured, organized, and documented system for dealing with risks" [35]. 53, 54, 70

significant risk a "risk that is significant as a result of the combination of its severity, intensity, probability of occurrence, and duration of its effects, and its the ability to affect an individual, a plurality of persons or to affect a particular group of persons;" [38]. 53, 54

Bibliography

[1] Tor H Aasheim, Knut T Hufthammer, Sølve Ånneland, Håvard Brynjulfsen, and Marija Slavkovik. „Bias mitigation with AIF360: A comparative study“. In: *Proceedings of the NIKT-2020*. Norwegian conference for ICT-research and education. University of South-Eastern Norway: Bibsys Open Journal Systems, Nov. 2020. URL: https://hdl.handle.net/11250/2764230.

[2] Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V. Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M. Pohl. „Representation Learning with Statistical Independence to Mitigate Bias“. In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. 2021 IEEE WACV. Waikoloa, HI, USA: IEEE, Jan. 2021, pp. 2512–2522. ISBN: 978-1-66540-477-8. DOI: 10.1109/WACV48630.2021.00256.

[3] Salem Alelyani. „Detection and Evaluation of Machine Learning Bias“. In: *Applied Sciences* 11(14).6271 (July 7, 2021). ISSN: 2076-3417. DOI: 10.3390/app11146271.

[4] Alexander Amini, Ava P. Soleimany, Wilko Schwarting, Sangeeta N. Bhatia, and Daniela Rus. „Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure“. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19. Honolulu HI USA: ACM, Jan. 27, 2019, pp. 289–295. ISBN: 978-1-4503-6324-2. DOI: 10.1145/3306618.3314243.

[5] Rico Angell, Brittany Johnson, Yuriy Brun, and Alexandra Meliou. „Themis: automatically testing software for discrimination“. In: *Proceedings of the 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ESEC/FSE '18. Lake Buena Vista FL USA: ACM, Oct. 26, 2018, pp. 871–875. ISBN: 978-1-4503-5573-5. DOI: 10.1145/3236024.3264590.

[6] R. Armstrong, B. J. Hall, J. Doyle, and E. Waters. „'Scoping the scope' of a cochrane review“. In: *Journal of Public Health* 33.1 (Mar. 1, 2011), pp. 147–150. ISSN: 1741-3842, 1741-3850. DOI: 10.1093/pubmed/fdr015. URL: https://academic.oup.com/jpubhealth/article-lookup/doi/10.1093/pubmed/fdr015 (visited on 07/25/2023).

[7] Banu Aysolmaz, Nancy Dau, and Deniz Iren. „Preventing Algorithmic Bias in the Development of Algorithmic Decision-Making Systems: A Delphi Study“. In: *Proceedings of the 53rd Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences. Vol. 53. Jan. 7, 2020, pp. 5268–5276. DOI: 10.24251/HICSS.2020.648.

[8] Ana Azevedo and Manuel Filipe Santos. „KDD, SEMMA and CRISP-DM: A parallel Overview“. In: *Proceedings of the IADIS European Conference on Data Mining* (Jan. 2008), pp. 182–185. ISSN: 978-972-8924-63-8. URL: http://dblp.uni-trier.de/db/conf/iadis/dm2008.html#AzevedoS08.

[9] Marley Bacelar. *Monitoring bias and fairness in machine learning models: A review*. May 6, 2021. DOI: 10.14293/S2199-1006.1.SOR-.PP59WRH.v1. (Visited on 08/24/2021).

[10] Ricardo Baeza-Yates. „Bias on the web“. In: *Communications of the ACM* 61.6 (May 23, 2018), pp. 54–61. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3209581. URL: https://dl.acm.org/doi/10.1145/3209581 (visited on 07/27/2023).

[11] Agathe Balayn, Christoph Lofi, and Geert-Jan Houben. „Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems“. In: *The VLDB Journal* (May 5, 2021). ISSN: 1066-8888, 0949-877X. DOI: 10.1007/s00778-021-00671-8.

[12] Agathe Balayn, Jie Yang, Zoltan Szlavik, and Alessandro Bozzon. „Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature“. In: *ACM Transactions on Social Computing* 4.3 (Sept. 30, 2021), pp. 1–56. DOI: 10.1145/3479158.

[13] Niels Bantilan. *Themis-ml: A Fairness-aware Machine Learning Interface for End-to-end Discrimination Discovery and Mitigation*. Oct. 18, 2017. URL: http://arxiv.org/abs/1710.06921 (visited on 07/26/2023).

[14] Samuel Barbosa, Dan Cosley, Amit Sharma, and Roberto M. Cesar-Jr. „Averaging Gone Wrong: Using Time-Aware Analyses to Better Understand Behavior“. In: *Proceedings of the 25th International Conference on World Wide Web*. Apr. 11, 2016, pp. 829–841. DOI: 10.1145/2872427.2883083.

[15] Solon Barocas, Moritz Hardt, and Arvind Narayanan. „Fairness in Machine Learning“. In: *www.fairmlbook.org* (2019), p. 181. URL: http://www.fairmlbook.org.

[16] Denis Baylor et al. „TFX: A TensorFlow-Based Production-Scale Machine Learning Platform“. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17. Halifax NS Canada: ACM, Aug. 13, 2017, pp. 1387–1395. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098021. (Visited on 09/05/2023).

[17] David K. Becker. „Predicting outcomes for big data projects: Big Data Project Dynamics (BDPD): Research in progress“. In: *Proceedings of the IEEE International Conference on Big Data (Big Data)*. Boston, MA: IEEE, Dec. 2017, pp. 2320–2330. ISBN: 978-1-5386-2715-0. DOI: 10.1109/BigData.2017.8258186. (Visited on 09/29/2021).

[18] Rachel K. E. Bellamy et al. „AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias“. In: *arXiv:1810.01943 [cs]* (Oct. 3, 2018). URL: http://arxiv.org/abs/1810.01943 (visited on 10/06/2021).

[19] Philippe Besse, Eustasio del Barrio, Paula Gordaliza, Jean-Michel Loubes, and Laurent Risser. „A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set“. In: *The American Statistician* 76(2) (Apr. 6, 2020), pp. 1–25. DOI: 10.1080/00031305.2021.1952897.

[20] Reid Blackman. *Ethical Machines - Your Concise Guide to Totally Unbiased, Transparent and Respectful AI*. Harvard Business Review Press, 2022. ISBN: 978-1-64782-281-1.

[21] Nathan Bosch and Jan Bosch. „Software Logs for Machine Learning in a DevOps Environment“. In: *Proceedings of the 46th Euromicro Conference on Software Engineering and Advanced Applications*. SEAA. Portoroz, Slovenia: IEEE, Jan. 24, 2020, pp. 29–33. URL: http://arxiv.org/abs/2001.10794 (visited on 08/02/2021).

[22] Patrick Bradley. „Risk management standards and the active management of malicious intent in artificial superintelligence“. In: *AI & SOCIETY* 35.2 (June 2020), pp. 319–328. ISSN: 0951-5666, 1435-5655. DOI: 10.1007/s00146-019-00890-2.

[23] Jason Brownlee. *Imbalanced Classification with Python*. Jason Brownlee, 2020.

[24] Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. „Optimized Pre-Processing for Discrimination Prevention“. In: *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS '17. Dec. 2017, pp. 3995–4004. URL: https://dl.acm.org/doi/10.5555/3294996.3295155.

[25] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, and Daniele Regoli. *The zoo of Fairness metrics in Machine Learning*. June 11, 2021. URL: http://arxiv.org/abs/2106.00467 (visited on 08/26/2021).

[26] Simon Caton and Christian Haas. „Fairness in Machine Learning: A Survey“. In: *ACM Computing Surveys* (Aug. 2023). DOI: https://doi.org/10.1145/3616865.

[27] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. „Step-by-step data mining guide“. In: *SPSS Inc* (2000), p. 76.

[28] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. „Bias and Debias in Recommender System: A Survey and Future Directions“. In: *ACM Transactions on Information Systems* 1.3 (Dec. 29, 2021), pp. 1–39. DOI: 10.1145/3564284.

[29] Sam Corbett-Davies and Sharad Goel. *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*. Aug. 14, 2018. URL: http://arxiv.org/abs/1808.00023 (visited on 07/15/2021).

[30] Bo Cowgill and Catherine Tucker. „Algorithmic Bias: A Counterfactual Perspective“. In: *NSF Trustworthy Algorithms* (Dec. 2017), p. 3.

[31] Jessica Dai and Sarah M Brown. „Label Bias, Label Shift: Fair Machine Learning with Unreliable Labels“. In: *34th Conference on Neural Information Processing Systems*. Workshop on Consequential Decision Making in Dynamic Environments (2020), p. 8.

[32] David Danks and Alex John London. „Algorithmic Bias in Autonomous Systems“. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. Melbourne, Australia, Aug. 2017, pp. 4691–4697. ISBN: 978-0-9992411-0-3. DOI: 10.24963/ijcai.2017/654.

[33] Vanesa Arjonilla Díez, Maider Fernández Egido, Juan Jesús García Sánchez, José Carlos Baquero Triguero, Pablo González Fuente, Alexander Benítez Buenache, and Antón Makarov Samusev. „Analysis and mitigation of bias in Machine Learning“. In: *Master Thesis (Mathematical Engineering) – Universidad Complutense Madrid* (2019), p. 21.

[34] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. „Fairness Through Awareness“. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. Jan. 2012, pp. 214–226. DOI: 10.1145/2090236.2090255.

[35] Peter J. Edwards, Paulo Vaz Serra, and Michael Edwards. *Managing project risks*. Hoboken, NJ: Wiley-Blackwell, 2020. 431 pp. ISBN: 978-1-119-48976-4.

[36] European Commission, Directorate-General for Communications Networks, Content and Technology. *ANNEXES to the Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative Acts*. COM(2021) 206 final. Apr. 21, 2021. URL: https://ec.europa.eu/transparency/documents-register/detail?ref=COM(2021)206&lang=en.

[37] European Commission, Directorate-General for Communications Networks, Content and Technology. *PROPOSAL for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative Acts*. COM(2021) 206 final. Apr. 21, 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206.

[38] European Parliament. *Artificial Intelligence Act – Texts Adopted*. June 14, 2023.

[39] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. „A Survey on Bias in Visual Datasets“. In: *Journal on Computer Vision and Image Understanding* 223.103552 (June 23, 2022). DOI: 10.1016/j.cviu.2022.103552.

[40] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. „From Data Mining to Knowledge Discovery in Databases“. In: *Communications of the ACM* 39.11 (Nov. 1, 1996), pp. 27–34. DOI: 10.1145/240455.240464.

[41] Emilio Ferrara. *airness and Bias in Artificial Intelligence: A brief Survey of Sources, Impacts and Mitigation Strategies*. 2023. DOI: 10.48550.

[42] Arthur Flexer and Dominik Schnitzer. „A MIREX META-ANALYSIS OF HUBNESS IN AUDIO MUSIC SIMILARITY“. In: *Proceedings of the 13th International Society for Mu- sic Information Retrieval Conference*. ISMIR 2012. Porto, Portugal, 2012, pp. 175–180. ISBN: 978-972-752-144-9.

[43] Bill Franks. *97 Things About Ethics Everyone in Data Science Should Know*. OCLC: 1225958431. 2020. ISBN: 978-1-4920-7266-9.

[44] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. „A comparative study of fairness-enhancing interventions in machine learning“. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta GA USA: ACM, Feb. 13, 2018, pp. 329–338. ISBN: 978-1-4503-6125-5. DOI: 10.1145/3287560.3287589.

[45] Christopher G. Harris. „Mitigating Cognitive Biases in Machine Learning Algorithms for Decision Making“. In: *Companion Proceedings of the Web Conference 2020*. WWW '20. Taipei Taiwan: ACM, Apr. 20, 2020, pp. 775–781. ISBN: 978-1-4503-7024-0. DOI: 10.1145/3366424.3383562.

[46] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. „Fairness testing: testing software for discrimination“. In: *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. ESEC/FSE'17. Paderborn Germany: ACM, Aug. 21, 2017, pp. 498–510. DOI: 10.1145/3106237.3106277.

[47] Pratyush Garg, John Villasenor, and Virginia Foggo. „Fairness Metrics: A Comparative Analysis“. In: *Proceedings of the 2020 IEEE International Conference on Big Data (Big Data)*. Atlanta, GA, USA: IEEE, Dec. 10, 2020, pp. 3662–3666. ISBN: 978-1-72816-251-5. DOI: 10.1109/BigData50022.2020.9378025.

[48] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. „Datasheets for Datasets“. In: *Communications of the ACM* 64.12 (2021), pp. 86–92. DOI: 10.1145/3458723.

[49] Diana F. Gordon and Marie Desjardins. „Evaluation and selection of biases in machine learning“. In: *Machine Learning* 20.1 (1995), pp. 5–22. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/BF00993472.

[50] Sara Hajian, Francesco Bonchi, and Carlos Castillo. „Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining“. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM SIGKDD. San Francisco California USA, Aug. 13, 2016, pp. 2125–2126. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2945386.

[51] Michaela Hardt et al. „Amazon SageMaker Clarify: Machine Learning Bias Detection and Explainability in the Cloud“. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. ACM SIGKDD. Aug. 14, 2021, pp. 2974–2983. DOI: 10.1145/3447548.3467177.

[52] Moritz Hardt, Eric Price, and Nathan Srebro. „Equality of Opportunity in Supervised Learning“. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Oct. 7, 2016, pp. 3323–3331. DOI: 10.48550/arXiv.1610.02413. (Visited on 08/12/2021).

[53] Bertrand K. Hassani. „Societal bias reinforcement through machine learning: a credit scoring perspective“. In: *AI and Ethics* 1 (Dec. 18, 2020), pp. 239–247. ISSN: 2730-5953, 2730-5961. DOI: 10.1007/s43681-020-00026-z. URL: 239%E2%80%93247%20(2021) (visited on 07/16/2021).

[54] Thomas Hellström, Virginia Dignum, and Suna Bensch. „Bias in Machine Learning – What is it Good for?“ In: *CEUR Workshop Proceedings*. Vol. 2659. Sept. 20, 2020, pp. 3–10. URL: http://arxiv.org/abs/2004.00686 (visited on 07/15/2021).

[55] Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. „Design Science in Information Systems Research“. In: *MIS Quarterly* Vol. 28 No. 1 (Mar. 2004), pp. 75–105. URL: https://www.researchgate.net/profile/Alan-Hevner/publication/201168946_Design_Science_in_Information_Systems_Research/links/5405d4670cf23d9765a75fc2/Design-Science-in-Information-Systems-Research.pdf (visited on 04/11/2023).

[56] HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE. *Ethics Guidelines for Trustworthy AI*. Apr. 8, 2019.

[57] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. „The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards“. In: *Workshop on Dataset Curation and Security*. NeurIPS 2020. May 2018. DOI: 10.5040/9781509932771. URL: http://securedata.lol/camera_ready/26.pdf (visited on 06/01/2021).

[58] ISO/IEC 2022. *ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*. July 2022.

[59] ISO/IEC 2022. *ISO/IEC TR 24027:2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making*. Nov. 2021.

[60] Heinrich Jiang and Ofir Nachum. „Identifying and Correcting Label Bias in Machine Learning“. In: *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics* 108 (2020), p. 10.

[61] Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. „Quantifying explainable discrimination and removing illegal discrimination in automated decision making“. In: *Knowledge and Information Systems* 35.3 (June 2013), pp. 613–644. ISSN: 0219-1377, 0219-3116. DOI: 10.1007/s10115-012-0584-8.

[62] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. „An Empirical Study of Rich Subgroup Fairness for Machine Learning“. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Jan. 2019, pp. 100–109. DOI: 10.1145/3287560.3287592.

[63] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. „Preventing Fairness Gerrymandering:Auditing and Learning for Subgroup Fairness“. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. 2018, pp. 2564–2572. URL: https://proceedings.mlr.press/v80/kearns18a.html (visited on 09/26/2023).

[64] Barbara Kitchenham and Stuart M. Charters. *Guidelines for performing Systematic Literature Reviews in Software Engineering*. EBSE Technical Report Version 2.3. Jan. 2007. URL: https://www.researchgate.net/publication/302924724 (visited on 09/29/2023).

[65] René F. Kizilcec. „Reducing non-response bias with survey reweighting: applications for online learning researchers“. In: *Proceedings of the first ACM Conference on Learning @ scale conference*. L@S 2014. Atlanta Georgia USA: ACM, Mar. 4, 2014, pp. 143–144. ISBN: 978-1-4503-2669-8. DOI: 10.1145/2556325.2567850.

[66] Tomáš Kliegr, Štěpán Bahník, and Johannes Fürnkranz. „A review of possible effects of cognitive biases on interpretation of rule-based machine learning models“. In: *Artificial Intelligence* 295 (June 2021), p. 103458. ISSN: 00043702. DOI: 10.1016/j.artint.2021.103458. URL: https://linkinghub.elsevier.com/retrieve/pii/S0004370221000096 (visited on 06/12/2023).

[67] Riikka Koulu. „Proceduralizing control and discretion: Human oversight in artificial intelligence policy“. In: *Maastricht Journal of European and Comparative Law* 27.6 (Dec. 2020), pp. 720–735. ISSN: 1023-263X, 2399-5548. DOI: 10.1177/1023263X20978649.

[68] Michelle Seng Ah Lee and Jantinder Singh. „Risk Identification Questionnaire for Detecting Unintended Bias in the Machine Learning Development Lifecycle“. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '21. July 2021, pp. 704–714. DOI: 10.1145/3461702.3462572.

[69] Yi Li and Nuno Vasconcelos. „REPAIR: Removing Representation Bias by Dataset Resampling“. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 IEEE/CVF. Long Beach, CA, USA: IEEE, June 2019, pp. 9564–9573. ISBN: 978-1-72813-293-8. DOI: 10.1109/CVPR.2019.00980.

[70] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. „Quantifying and alleviating political bias in language models“. In: *The journal of Artificial Intelligence* 304.103654 (Mar. 2022). ISSN: 0004-3702. DOI: 10.1016/j.artint.2021.103654.

[71] Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. „Bias Mitigation Post-processing for Individual and Group Fairness“. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ICASSP 2019. Brighton, United Kingdom: IEEE, May 2019, pp. 2847–2851. DOI: 10.1109/ICASSP.2019.8682620.

[72] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. „Fairness Through Causal Awareness: Learning Latent-Variable Models for Biased Data“. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Dec. 2, 2018, pp. 349–358. DOI: 10.1145/3287560.3287564.

[73] Oscar Marbán, Javier Segovia, Ernestina Menasalvas, and Covadonga Fernández-Baizán. „Toward data mining engineering: A software engineering approach“. In: *Journal of Information Systems* 34.1 (Mar. 2009), pp. 87–107. ISSN: 03064379. DOI: 10.1016/j.is.2008.04.003.

[74] Iñigo Martinez, Elisabeth Viles, and Igor G. Olaizola. „Data Science Methodologies: Current Challenges and Future Approaches“. In: *Journal on Big Data Research* 24.100183 (May 2021). ISSN: 22145796. DOI: 10.1016/j.bdr.2020.100183.

[75] Fernando Martínez-Plumed, Lidia Contreras-Ochando, Cèsar Ferri, Peter Flach, José Hernández-Orallo, Meelis Kull, Nicolas Lachiche, and María José Ramírez-Quintana. „CASP-DM: Context Aware Standard Process for Data Mining“. In: (Sept. 19, 2017). URL: https://arxiv.org/abs/1709.09003 (visited on 09/26/2023).

[76] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. *A Survey on Bias and Fairness in Machine Learning*. Sept. 17, 2019. URL: https://arxiv.org/abs/1908.09635v2 (visited on 05/31/2021).

[77] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. „A Survey on Bias and Fairness in Machine Learning“. In: *ACM Computing Surveys* 54.6 (July 13, 2021), pp. 1–35. DOI: 10.1145/3457607.

[78] Christoph Molnar. „Interpretable Machine Learning“. In: (Feb. 21, 2019), p. 251.

[79] Arvind Narayanan. *Tutorial: 21 fairness definitions and their politics*. YouTube. 55:20 minuts. Mar. 1, 2018. URL: https://www.youtube.com/watch?v=jIXIuYdnyyk (visited on 02/22/2022).

[80] National Pilot Comitttee for Digital Ethics (CNPEN). *Opinion N°3 – ethical issues of conversational agents*. page 19-29. Sept. 15, 2021. URL: https://www.ccne-ethique.fr/sites/default/files/2022-05/CNPEN%233-ethical_issues_of_conversational_agents.pdf (visited on 01/03/2023).

[81] Eirini Ntoutsi et al. „Bias in Data-driven AI Systems – An Introductory Survey“. In: *WIREs Data Mining and Knowledge Discovery Journal* 10.3 (May 2020). ISSN: 1942-4787, 1942-4795. DOI: 10.1002/widm.1356.

[82] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. „Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries“. In: *Journal on Frontiers in Big Data Sec. Data Mining and Management* 2.13 (July 11, 2019). DOI: 10.3389/fdata.2019.00013.

[83] Alexandra Olteanu, Emre Kıcıman, and Carlos Castillo. „A Critical Review of Online Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries“. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM 2018. Marina Del Rey CA USA: ACM, Feb. 2, 2018, pp. 785–786. DOI: 10.1145/3159652.3162004.

[84] Carol L. Perryman. „Mapping studies“. In: *Journal of the Medical Library Association* 104 (Jan. 2016), pp. 79–82. ISSN: 1536-5050, 1558-9439. DOI: 10.3163/1536-5050.104.1.014.

[85] Vilayanur S. Ramachandran, ed. *Encyclopedia of human behavior*. 2. ed. London: Academic Press, 2012. ISBN: 978-0-12-375000-6.

[86] Sabbir M. Rashid, James P. McCusker, Paulo Pinheiro, Marcello P. Bax, Henrique O. Santos, Jeanette A. Stingone, Amar K. Das, and Deborah L. McGuinness. „The Semantic Data Dictionary – An Approach for Describing and Annotating Data“. In: *Data Intelligence* 2.4 (Oct. 2020), pp. 443–486. ISSN: 2641-435X. DOI: 10.1162/dint_a_00058.

[87] Andreas Rauber, Maximilian Staats, and Cornelia Michlits. *Report on Fair, Open and Robust Algorithms - Bias in Data Analytics Processes*. Aug. 2020.

[88] Aida Sharif Rohani and Ricardo Baeza-Yates. *Measuring Bias - Bias in small datasets*. 2022.

[89] Drew Roselli, Jeanna Matthews, and Nisha Talagala. „Managing Bias in AI“. In: *Companion Proceedings of The 2019 World Wide Web Conference*. WWW '19. San Francisco USA: ACM, May 13, 2019, pp. 539–544. ISBN: 978-1-4503-6675-5. DOI: 10.1145/3308560.3317590.

[90] Denis Rothman. *Hands-On explainable AI (XAI) with Python: interpret, visualize, explain, and integrate reliable AI for fair, secure, and trustworthy AI apps*. Birmingham Mumbai: Packt Publishing, 2020. 428 pp. ISBN: 978-1-80020-813-1.

[91] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. „How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations“. In: *Journal on Artificial Intelligence* 283.103238 (June 2020). ISSN: 00043702. DOI: 10.1016/j.artint.2020.103238.

[92] Sebastian Schelter, Felix Biessmann, Dustin Lange, Tammo Rukat, Phillipp Schmidt, Stephan Seufert, Pierre Brunelle, and Andrey Taptunov. „Unit Testing Data with Deequ“. In: *Proceedings of the 2019 International Conference on Management of Data*. SIGMOD/PODS '19. Amsterdam Netherlands: ACM, June 25, 2019, pp. 1993–1996. ISBN: 978-1-4503-5643-5. DOI: 10.1145/3299869.3320210.

[93] Rahul Sethi, Vedang Ratan Vatsa, and Parth Chhaparwal. „Identification and Mitigation of Algorithmic Bias through Policy Instruments“. In: *International Journal of Advanced Research* 8.7 (July 31, 2020), pp. 1515–1522. ISSN: 23205407. DOI: 10.21474/IJAR01/11418.

[94] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. „Representation Bias in Data: A Survey on Identification and Resolution Techniques“. In: *ACM Computing Surveys* (Mar. 17, 2023), p. 3588433. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3588433.

[95] Moninder Singh and Karthikeyan Natesan Ramamurthy. *Understanding racial bias in health using the Medical Expenditure Panel Survey data*. Nov. 4, 2019. URL: http://arxiv.org/abs/1911.01509 (visited on 06/12/2023).

[96] Christian B. Smart. *Solving for Project Risk Management*. McGraw-Hill, 2021. ISBN: 978-1-260-47383-4.

[97] Derek Snow. „FairPut: A Light Framework for Machine Learning Fairness with LightGBM“. In: *SSRN Electronic Journal* (2020). ISSN: 1556-5068. DOI: 10.2139/ssrn.3619715.

[98] Eva Soleimany. „MIT 6.S191: AI Bias and Fairness“. Jan. 2021. URL: https://www.youtube.com/watch?v=wmyVODy_WD8&list=PLtBw6njQRU-rwp5__7C0oIVt26ZgjG9NI&index=8 (visited on 08/04/2021).

[99] Ramya Srinivasan and Ajay Chander. „Biases in AI systems“. In: *Communications of the ACM* 64.8 (Aug. 2021), pp. 44–49. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3464903.

[100] Stefan Studer, Thanh Binh Bui, Christian Drescher, Alexander Hanuschkin, Ludwig Winkler, Steven Peters, and Klaus-Robert Mueller. „Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology“. In: *Machine Learning and Knowledge Extraction* 3.2 (Feb. 24, 2021), pp. 392–413. ISSN: 2504-4990. DOI: 10.3390/make3020020.

[101] Harini Suresh and John V. Guttag. „A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle“. In: *Equity and Access in Algorithms, Mechanisms, and Optimization* 17 (June 15, 2021), pp. 1–9. ISSN: 978-1-4503-8553-4. DOI: 10.1145/3465416.3483305.

[102] Harini Suresh and John V. Guttag. „A Framework for Understanding Unintended Consequences of Machine Learning“. In: *EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization* 17 (Feb. 17, 2020), pp. 1–9. URL: http://arxiv.org/abs/1901.10002 (visited on 06/01/2021).

[103] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. „FairTest: Discovering Unwarranted Associations in Data-Driven Applications“. In: *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. 2017 IEEE European Symposium on Security and Privacy (EuroS&P). Paris: IEEE, Apr. 2017, pp. 401–416. ISBN: 978-1-5090-5762-7. DOI: 10.1109/EuroSP.2017.29.

[104] Amos Tversky and Daniel Kahneman. „Judgment under Uncertainty: Heuristics and Biases“. In: 185 (1974).

[105] Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller, and Liz Sonenberg. „The Effects of Explanations on Automation Bias“. In: *Artificial Intelligence* 322 (Sept. 2023). ISSN: 00043702. DOI: 10.1016/j.artint.2023.103952.

[106] Sahil Verma and Julia Rubin. „Fairness definitions explained“. In: *Proceedings of the International Workshop on Software Fairness*. ICSE '18: 40th International Conference on Software Engineering. Gothenburg Sweden: ACM, May 29, 2018, pp. 1–7. ISBN: 978-1-4503-5746-3. DOI: 10.1145/3194770.3194776.

[107] Sandra Wachter, Brent Mittelstadt, and Chris Russell. „Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non- Discrimination Law“. In: *West Virginia Law Review* 123.3 (Jan. 15, 2021). DOI: 10.2139/ssrn.3792772.

[108] Sandra Wachter, Brent Mittelstadt, and Chris Russell. „Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI“. In: *Journal on Computer Law & Security Review* 41.105567 (July 2021), pp. 1–31. ISSN: 02673649. DOI: 10.1016/j.clsr.2021.105567.

[109] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. „A Survey of Human-in-the-loop for Machine Learning“. In: *Journal on Future Generation Computer Systems* 135 (Oct. 2022), pp. 364–381. DOI: https://doi.org/10.1016/j.future.2022.05.014.

[110] Brit Youngmann, Michael Cafarella, Yuval Moskovitch, and Babak Salimi. „NEXUS: On Explaining Confounding Bias“. In: *Companion of the 2023 International Conference on Management of Data*. SIGMOD '23. Seattle, WA, USA, June 18, 2023, pp. 171–174. ISBN: 978-1-4503-9507-6. DOI: 10.1145/3555041.3589728. arXiv: 2210.02943[cs].

[111] Desen Yuan. *Language bias in Visual Question Answering: A Survey and Taxonomy*. Nov. 16, 2021. URL: http://arxiv.org/abs/2111.08531 (visited on 06/12/2023).

[112] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. „Fairness Constraints: Mechanisms for Fair Classification“. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. Vol. 54. Proceedings of Machine Learning Research. ISSN: 2640-3498. PMLR, Apr. 10, 2017, pp. 962–970. URL:

https://proceedings.mlr.press/v54/zafar17a.html (visited on 03/21/2022).

[113] Lu Zhang, Yongkai Wu, and Xintao Wu. „Achieving Non-Discrimination in Data Release“. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17: The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax NS Canada: ACM, Aug. 13, 2017, pp. 1335–1344. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098167.

[114] Yaochen Zhu, Jing Ma, and Jundong Li. *Causal Inference in Recommender Systems: A Survey of Strategies for Bias Mitigation, Explanation, and Generalization*. Jan. 2, 2023. URL: http://arxiv.org/abs/2301.00910 (visited on 06/12/2023).

[115] James Zou and Londa Schiebinger. „Design AI so that it's fair“. In: *Journal Nature* 559 (July 19, 2018), pp. 324–226. ISSN: 1476-4687. URL: https://www.nature.com/articles/d41586-018-05707-8.