MASTER'S THESIS

A(I) DESIGN STUDY:

A sociotechnical, card-based fairness tool for ADM-development

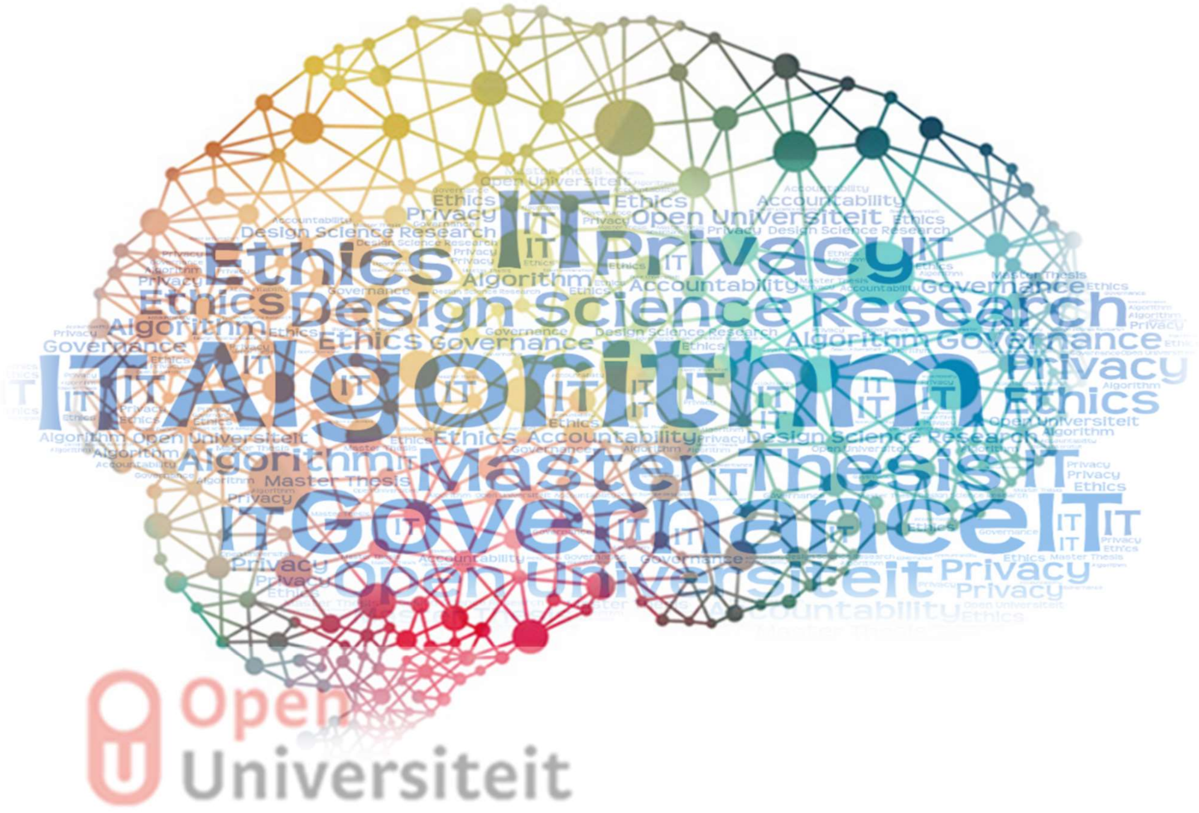
Leunissen, F. L. G.

Award date:
2022

MASTER THESIS| OPEN UNIVERSITEIT

A(I) DESIGN STUDY:

A sociotechnical, card-based fairness tool for ADM-development



Programme:	Open University of the Netherlands, faculty of Science Master of Science Business Process Management & IT (BPMIT)
Course:	IM0602 Voorbereiden Afstuderen BPMIT IM9806 Afstudeeropdracht Business Process Management and IT
Student (ID):	F.L.G. Leunissen
Date:	November 25, 2022
Thesis Advisor:	Dr. L.L.H. Bollen
Co-reader:	Dr. T. Huygh
Version:	Full Version
Status:	**Final Master Thesis / "AF"-report excluding appendices**

Abstract

By using the Design Science Research Methodology a method-agnostic, card-based artefact, i.e. design probe for guiding the embeddedness of fairness in Algorithmic Decision-Making Systems (ADM's) is created on literature groundings, envisioning a sociotechnical approach. The motive is originated in the practitioners' call for guidance, in the shape of light-weight, integrable tools which address (un)fairness in ADM-design. This tool, so called responsible AI:CEID quartet, is demonstrated by using several cards, during a workshop within a single ADM-developers organization, while reflecting upon an artificial Hiring-case founded by Princeton. Elements of Value Sensitive Design, Critical System Heuristics, and the Organizational Justice Theory are incorporated in the card deck to ensure a sociotechnical, and by these means contextual approach in fairness.

Evaluation on perceived usefulness reveals that although the extensive checklist it represents is warmly welcomed, there are several suggested add-ons, and improvement-points for a reiteration in design. Implementing organizational roles, nuancing the fairness-norms used, and including action-points in the form of scenario based hints, are examples of these. A follow-up study could implement these improvements, evolving and maturing the artefact, so the artefact may be of practical value in the future.

Key terms

Algorithmic Decision-making Systems, Fairness-by-Design, AI-Fairness, Machine Learning, Design Science Research, Sociotechnical Approach, Value Sensitive Design, Artefact, Bias Mitigation, CSH.

Summary

Algorithms are increasingly being used in daily decision-making. These so called algorithmic decision-making systems (ADM's) have huge potential, but simultaneously carry notable risks of impactful, ethical harms in them, as algorithms are value-laden. Fairness is one of these ethical values, which are implicitly, or explicitly embedded in these systems. Developers of these ADM's are therefore burdened with the difficult, and critical task to avoid biased designs, non-discriminatory harms, and other fairness-issues. This is a difficult task, as fairness is subjective, highly volatile, and context-sensitive.

Literature reveals that most often, fairness endeavors in ADM's are arguably being addressed, by technological approaches in isolation, using so called mathematical fairness-metrics on decision-outcomes. These approaches are said to be too narrow, as they abstract away the social environments these systems interact with. Therefore, several researchers suggest a holistic, socio-technical approach towards fairness in ADM's. Moreover, AI-developers, call for light-weighted, integrable tooling's to aid in designing for fairness, as they seem to struggle with fairness operationalizations in ADM's.

By using the Design Science Methodology, a design probe, i.e. artefact is built upon literature findings, which should serve practitioners' needs, by offering systemized guidance on a sociotechnical, Fairness-by-Design approach in ADM-development. Before the artefact is constructed, first a broad literature study is done, to reveal important concepts and their interrelations to conceptualize the overall framework that envisions a sociotechnical perspective. This framework is characterized to entail a top-down approach. As time, and resources are limited, a small, coherent set of constructs are derived from this broadly defined conceptual framework and cast in an artefact. This scoped artefact is tailored to the conceptual design stage of ADM's; the most common starting point of an ADM-development process. To meet the requirement of being light-weighted, integrable, and therefore method-agnostic, the artefact is shaped as a physical card-deck. For systemization, design-patterns of Value Sensitive Design, and Critical System Heuristics are incorporated. Fairness components, of the Organizational Justice Theory are also incorporated to provide a broader palette of fairness dimensions, so to promoting and illustrating a sociotechnical approach, and to customize the artefact on its testing domain of hiring.

In a first iteration, this artefact is introduced within an AI-developers (expert)organization. A workshop is conducted with a multidisciplinary expert-team of this organization, using a fictional Hiring-case as reflection material on which the artefact is demonstrated and evaluated. It is confirmed within this case-organization that practitioners find it challenging to embed fairness and therefore welcome a lightweight tooling in the shape of a guiding checklist to address this challenge. The research results unravel that there is a trade-off between simplicity, practicality, on the one hand, and detailed information on the other hand, as fairness is a complex phenomenon in which guidance has to be tailored to specific situations. Suggested points for improving the artefact are, amongst others: implementation of organizational roles, nuancing the fairness components more to dissolve ambiguities in terminology, enriching the artefact with situation-specific hints, and including (organizational) alignment considerations. The experts in the workshop also introduce a best practice; so called "low ethical impact, high value-initiatives". The "low hanging fruit": small, and simple objective tasks to be automated by an AI-decision component, which all-together bring big gains, with little ethical risks. This is a valuable add-on, which entails a pragmatic approach, in

contrast to the top-down approach envisioned in our framework, and to be found in the majority of academic articles.

To be mentioned is that the generalizability of these findings are highly limited, as there is only a single organization used for this first iteration in artefact-design. A weakness of the research design is that a lot of information had to be bridged, so that at certain points participants may not have fully understood the artefacts’ content, and intentions. A reiteration in design and development is recommended, wherein an improved version (taking the suggestions in consideration) is evaluated. To prevent information overload, one could consider to use a smaller set of cards in that iteration.

Moreover, ideally taken one would use an Exploratory Focus Group, and a Confirmatory Focus Group, as for valid determination of artefact-improvement. The structured, and gathered information in this study, by means of the overall framework around fairness, stakeholders, but also data, bias(-taxonomy), bias-mitigation techniques, and remedies could provide a stepping stone to continue this research. Future iterations, using one or more components of the overall framework, could by these means systemically be evaluated in small portions, to mature the artefact, so that the leftovers of these iterations define a coherent artefact, experienced to be of practical use.

Abbreviations used

The used abbreviations in this Master Thesis are presented, in alphabetical order, in the table below.

Abbreviation	**Term**
ABWG	Algorithmic Bias Working Group
ACM	Association for Computing Machinery
ADM	Algorithmic Decision-Making system
AI	Artificial Intelligence
AI:CEID	AI: Considering Ethics In Design (toolkit)
CFG	Confirmatory Focus Group
CRISP-DM	Cross Industry Process for Data Mining
CSH	Critical System Heuristics
DADM	Discrimination Aware Data Mining
DSR	Design Science Research
DSRM	Design Science Research Methodology
EFG	Exploratory Focus Group
FAT/ML	Fairness, Accountability, and Transparency in Machine Learning
FBD	Fairness-By-Design
IEEE	Institute of Electrical and Electronics Engineers
ML	Machine Learning
NLP	Natural Language Processing
OJT	Organizational Justice Theory
RAI	Responsible AI (Artificial Intelligence)
SLR	Systematic Literature Review
SVM	Support Vector Machines
VSD	Value Sensitive Design

Table of Contents

1. Introduction

Simultaneously with increased interweaving of - and dependency of society on – Machine Learning (ML) algorithms in different decision domains, we are already seeing our deep-rooted biases and prejudices mirrored in these algorithmic decision processes. Many times these algorithms and its outputs are considered to be neutral or objective, but the Algorithmic Decision-Making systems (hereafter: ADM's) we use are inescapably value-laden. Therefore, ADM's have the potential to amplify, perpetuate and systemise our own pre-existing human biases on an unprecedented scale (Rovatsos, Mittelstadt, & Koene, 2019), resulting in possible, systemic disadvantages to certain protected groups upon personal data (like e.g., race, gender, and age). Addressing fairness is therefore recognized to be of great importance in the design and application of contemporary data driven AI-systems, especially in the context of high impact decision-making. Harms like discriminatory outcomes and diminishing rights already occurred in real world settings of AI-decision making systems (e.g.: Ofqual's A-Level Algorithm, COMPAS, and Amazons' hiring algorithm (Angwin, Larson, Mattu, & Kirchner, 2016; Fu, Huang, & Singh, 2020; Smith, 2020)). These harms are to be avoided from both the perspective of affected parties, as well as from the perspective of involved companies and institutes with regard to regulatory compliance, costly fines and reputational risks. Arrieta et al. (2020) emphasize that fairness is one of the (Responsible) AI-specific components that must be operationalized within an ADM. Despite the recognized harmful potential of biased algorithms, so far there is neither an agreed upon solution nor an agreed upon terminology related to fairness and bias (Tal et al., 2019). Tools, methods, guidelines, and techniques are required for mitigating identified risks as discrimination and privacy-violations so that fair outcomes are warranted (Adadi & Berrada, 2018; Arrieta et al., 2020; Kroll et al., 2016; VSNU, 2017), but there still remain some challenges in this particularly field (Cramer, Garcia-Gathright, Reddy, Springer, & Takeo Bouyer, 2019; Lee & Singh, 2020).

By using the Design Science Research Methodology (DSRM) an attempt will be made to contribute to earlier challenges mentioned, by transforming the identified knowledge in the existing academic literature in a practical and useable form. An artefact in which we envision a sociotechnical, ex-ante approach, we call: Fairness-by-design (FBD).

In the next sections, §1.1 until §1.4, respectively the problem identification and motivation (based upon findings within the academic state-of-the art literature describing practical issues), solution- and research-objectives, and formulated research questions will be presented and discussed. In the last section (§1.5) of this Chapter an overview will be given upon the structure of this thesis.

1.1. Problem Identification and Motivation

Despite the significant literature on a large number of methods, new arising fair ML-techniques and different schools of thought around the pressing topic how to make fair models and mitigating biases, there are still no ready-made, industry-standard processes and tools for practitioners to assess and address unfair algorithmic and data biases (Caton & Haas, 2020; Cramer et al., 2019; Ntoutsi et al., 2020; Springer, Garcia-Gathright, & Cramer, 2018).

Results of recent research on open-source fairness toolkits show that industry practitioners are still struggling with approaches towards fairness, and mitigations of potential biases, in their ML-models and -systems (Cramer et al., 2019; Lee & Singh, 2020; Springer et al., 2018). Different communities are working on the topic, because of the needed interdisciplinary approach, such as the Algorithmic Bias Working Group (ABWG). This group is developing a standard for Algorithmic Bias Considerations (so called IEEE P7003[1]), intended to be a framework with guidelines, procedures and methods to identify and mitigate undesired biases in (the outcome of) algorithmic systems (Koene, Dowthwaite, & Seth, 2018). Unfortunately, this framework is at the present time of writing inaccessible due to the "work in progress" status of this research initiative.

The constant increasing of work, new perspectives on this mutual fairness and bias problem, and the relatively scattered literature around this topic (Springer et al., 2018) has created the need for "a unified framework for Fairness in AI in order to simplify the process of evangelization and implementation, especially in the industry." (Benjamins, Barbado, & Sierra, 2019). Some researchers state that a pure mathematical approach towards fairness is too narrow and insufficient (M. A. Madaio, Stark, Wortman Vaughan, & Wallach, 2020; Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019), hereby calling for a sociotechnical approach. Since law- and regulations regarding ethics in AI are still in development, and AI-experts are not ethicists, practical endeavours towards embedding fairness in the design of ADM's, bear the risks of ethics-washing, and ethical cherry-picking (Ashurst, Barocas, Campbell, & Raji, 2022; Bietti, 2020); Gogoll et al. (2021).

Researchers and practitioners urgently call for:

- ….a holistic, sociotechnical approach towards (un)fairness, taking the social context into account, instead of a focus on so called "techno-solutionism" solely (Draude, Klumbyte, Lücking, & Treusch, 2019; M. A. Madaio et al., 2020; Selbst et al., 2019);
- …..practical, lightweight tools for industry to asses and address (un)fairness in AI-solutions, that can be integrated into engineers' workflows by encountering the pragmatic challenges in translating the growing research literature into methods and pragmatic processes that are applicable across domains and easy to communicate, while still informative enough to be of help (Cramer et al., 2019; Lee & Singh, 2020; Springer et al., 2018);
- …..a systemic evaluation of the existing approaches and mitigation techniques for tackling bias and fairness in outcomes to understand their capabilities and limitations in real-world contexts, and how they can be used together to form effective mitigation strategies, so that a comprehensive solutions to all forms of unfair bias is provided (Ntoutsi et al., 2020; Rovatsos et al., 2019);

[1] Part of IEEE P7000: a process for addressing ethical concerns in system design. *Spiekermann, S. (2017). IEEE P7000—The first global standard process for addressing ethical concerns in system design. Multidisciplinary Digital Publishing Institute Proceedings, 1(3), 159.*

In this study we would like to follow a holistic, sociotechnical approach and make a first step towards an artefact[2] which is to be used for embedding FBD in AI. Holistic means that the solution is envisioned as an integrative, approach that collects, and aggregates the best components of current research and proposes an encompassing design. Mainly the artefact is intended to be a guidance and useful tool by illustrating which avenues one can take in addressing (un)fairness in the development of ADM's.

1.2. Solution Objectives and Scope

To address the above-mentioned issues and needs, this design study aspires to develop an artefact founded on findings in academic literature and experiences in the practical field, which entails a holistic angle of approach to ensure fairness in data driven ADM's. The research focus is limited to the development i.e., design-phase of data driven ADM's underlying supervised ML-algorithms for classification. This means that the deployment-phase of ADM's, as well as ADM's based upon unsupervised (e.g., clustering) and reinforcement learning in general are out of the research scope.

As mentioned before the DSR-Methodology is used to develop this artefact, which entails an incremental approach from global to detail with multiple iterations across the design, relevance, and rigor cycle. In accordance with this method a first step to be taken is to segment the problem conceptually so that the solution can capture the problem its complexity (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007).

When looking at the stated problems defined earlier (please see *§1.1: Problem identification and motivation*) we could formulate the following atomized parts of the problems stated:

P1: The literature around fairness and bias mitigation seems difficult to digest due to fragmentary, scattered and continuously growing research. A holistic overview approach fails (more transparency offered by systemisation is needed to keep track of the developments) (Ntoutsi et al., 2020);

P2: A Socio-technical approach is asked for instead of techno-solutionism solely (Draude et al., 2019);

P3: Translation process from literature findings towards practical field seems difficult. Tools for enhancing fairness, including assessing and addressing bias should be integrable in the engineers' workflow (Cramer et al., 2019; M. Madaio, Egede, Subramonyam, Wortman Vaughan, & Wallach, 2022);

P4: Practitioners need practical, light-weighted tools to assess and address (un)fairness in AI-solutions (Cramer et al., 2019; Lee & Singh, 2020; M. Madaio et al., 2022; Springer et al., 2018)

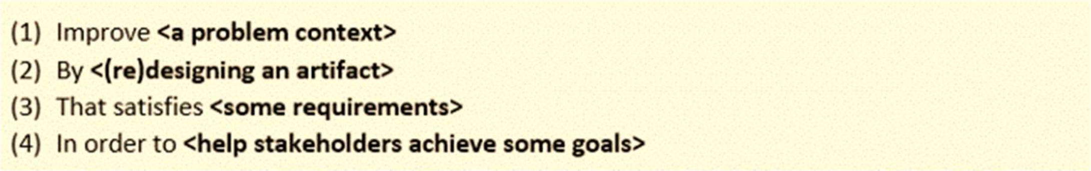
[2] Please note: what exactly is meant by an artefact, will be explained in Chapter 2 (Research Design).

1.3. Research Objectives

Based upon the above-mentioned solution objectives, the main research objective can be constructed as follows: designing a "proof-of-concept", that organizations can use to support fairness-enhancing in the development process of data driven ADM's, based upon findings and best-practices in the academic literature.

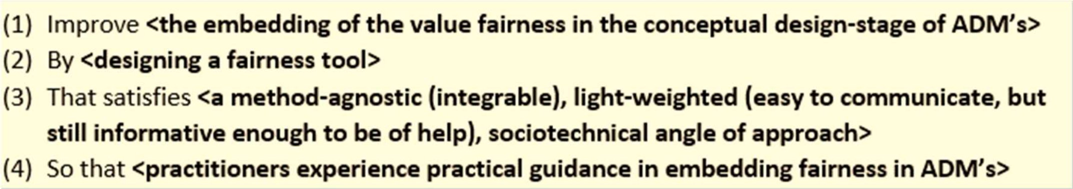
This design probe is used to test if a holistic, sociotechnical approach towards fairness-enhancement (where bias-mitigation forms only a part of) is feasible and which knowledge can be derived from this approach, when bridging the scientific and practical knowledge gap.

By following the template of Wieringa (2014):



(1) Improve **<a problem context>**
(2) By **<(re)designing an artifact>**
(3) That satisfies **<some requirements>**
(4) In order to **<help stakeholders achieve some goals>**

The research objective is stated as follows:



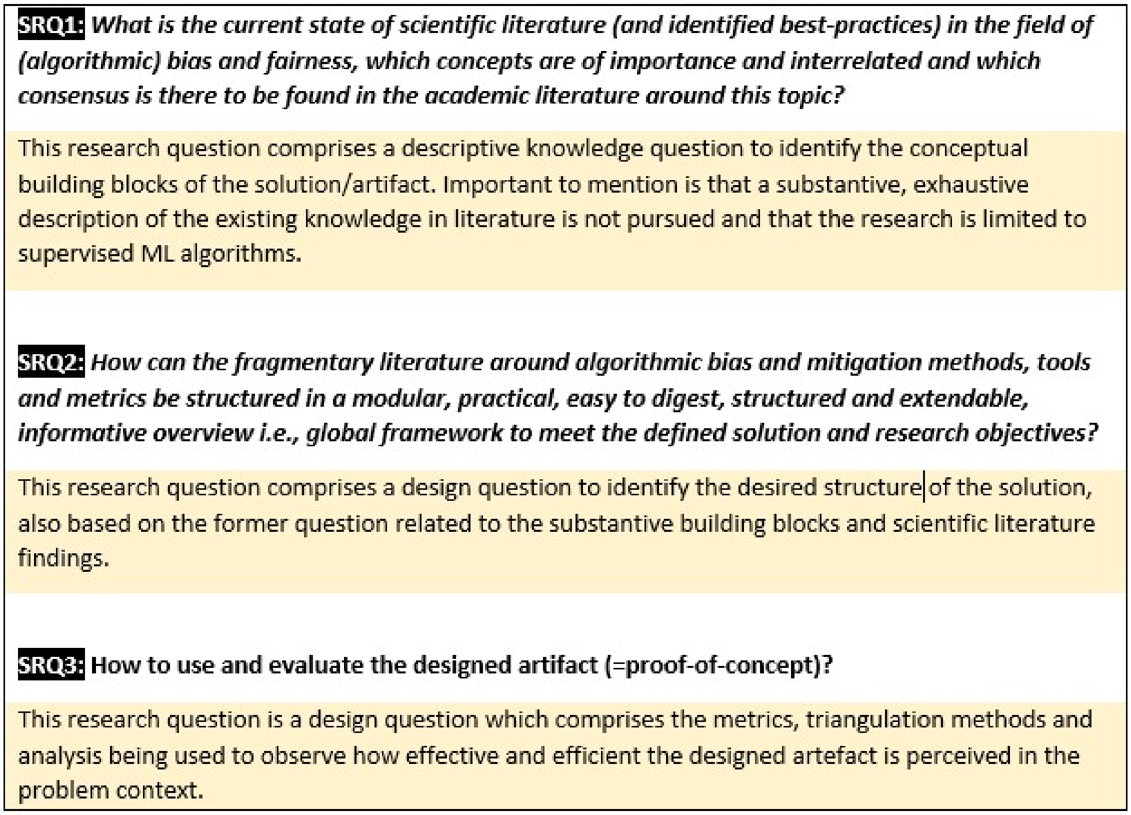
(1) Improve **<the embedding of the value fairness in the conceptual design-stage of ADM's>**
(2) By **<designing a fairness tool>**
(3) That satisfies **<a method-agnostic (integrable), light-weighted (easy to communicate, but still informative enough to be of help), sociotechnical angle of approach>**
(4) So that **<practitioners experience practical guidance in embedding fairness in ADM's>**

1.4. Research Questions

Given the stated solution- and research objectives in the former sections of this Chapter, the main research question (MRQ), in this design study, can be defined as follows:

> **MRQ*: What is the added practical value of a sociotechnical approach in the shape of an integrative, light-weighted, artefact/framework towards fairness-enhancement in the development-process of supervised, algorithmic decision-making systems?***

This leads to the following, investigative, sub-research questions (SRQ's), being knowledge and design-questions, where formulation, and construction is inspired by Thuan, Drechsler, and Antunes (2019); Wieringa (2014)[3]:



SRQ1: ***What is the current state of scientific literature (and identified best-practices) in the field of (algorithmic) bias and fairness, which concepts are of importance and interrelated and which consensus is there to be found in the academic literature around this topic?***

This research question comprises a descriptive knowledge question to identify the conceptual building blocks of the solution/artifact. Important to mention is that a substantive, exhaustive description of the existing knowledge in literature is not pursued and that the research is limited to supervised ML algorithms.

SRQ2: ***How can the fragmentary literature around algorithmic bias and mitigation methods, tools and metrics be structured in a modular, practical, easy to digest, structured and extendable, informative overview i.e., global framework to meet the defined solution and research objectives?***

This research question comprises a design question to identify the desired structure of the solution, also based on the former question related to the substantive building blocks and scientific literature findings.

SRQ3: **How to use and evaluate the designed artifact (=proof-of-concept)?**

This research question is a design question which comprises the metrics, triangulation methods and analysis being used to observe how effective and efficient the designed artefact is perceived in the problem context.

[3] Investigative questions: what prior knowledge is available(?), which requirements define the artefact(?), how can the artefact be evaluated(?), what are the models' essential components(?), what (new) knowledge does the artefact contribute(?)

1.5. Thesis Structure

The figure below provides the step by step approach which is followed, provided by Peffers et al. (2007). The several steps (activities 1 until 5) correspond each to a Chapter in this Master Thesis.

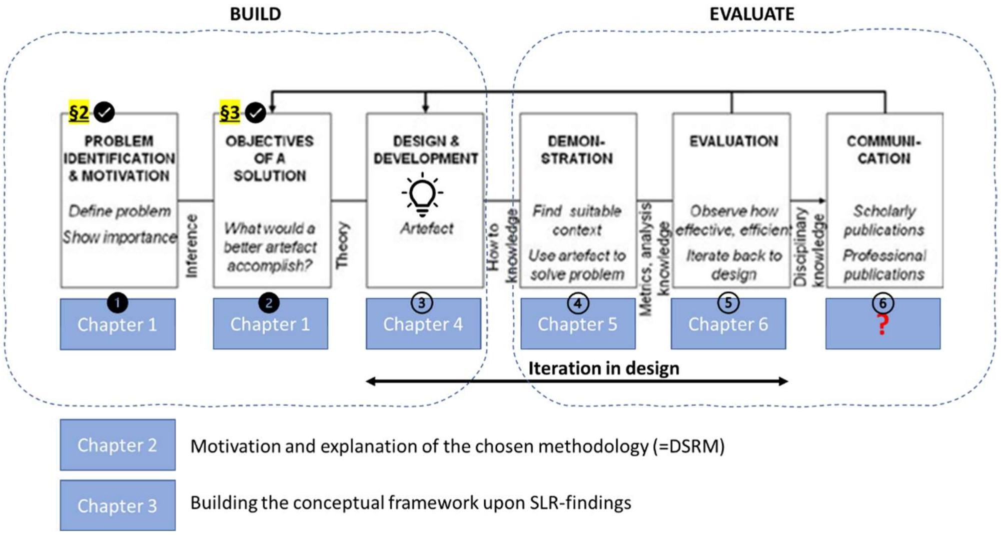


Figure 1-1: DSRM nominal process sequence adapted from Peffers et al. 2007

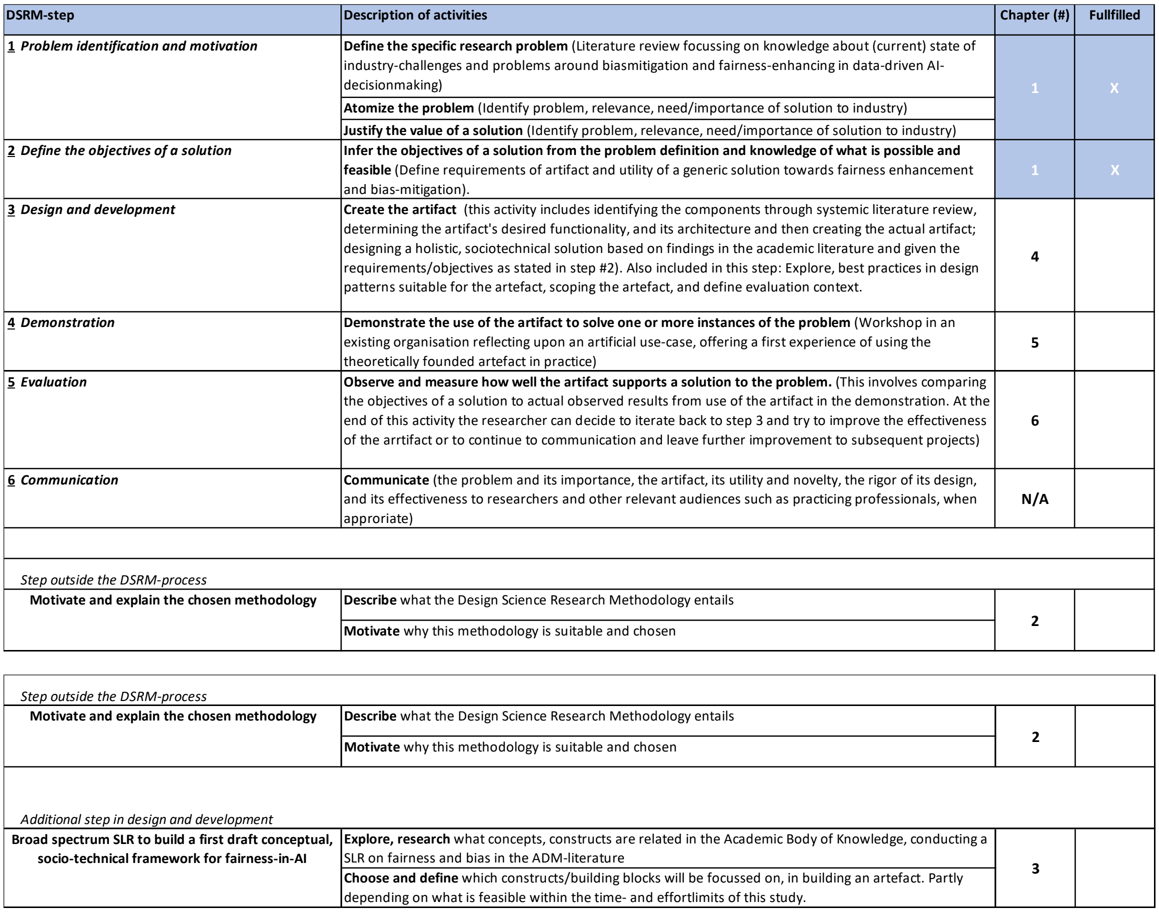
In Chapter 2 the chosen problem-centred research approach of DRSM will be explained and justified, since this is the thread trough this Master Thesis. In this methodology the research focus is twofold: utility and knowledge driven.

In Chapter 3 the Systematic Literature Review (SLR), discovering different building blocks of a first overall conceptual framework in relation to fairness within AI, is described. This conceptual draft framework forms the foundation for the design of a (scoped) artefact, wherein some building blocks are accentuated, and focussed on. One could consider this as a step part of Design & Development, but since the approach here is to broadly research the field of fairness in ADM's, discovering concepts which will not all be used, and evaluated by the developed artefact it is chosen to devote a separate Chapter to it.

The following Chapters 4 until 6 correspond to the different activities in the DSRM nominal process sequence (please see figure 1-1); respectively Design & Development, Demonstration, Evaluation of the artefact are being presented and discussed[4].

[4] DSRM-step #6 is displayed, as it is part of the whole cycle. However, this step is not applicable within this study, as communication is done at the end of artefact development after research is done, by going to a conference, or publishing an article about a potential successful artefact.

In table 1-1 these activities are detailed:



DSRM-step	Description of activities	Chapter (#)	Fullfilled
1 *Problem identification and motivation*	**Define the specific research problem** (Literature review focussing on knowledge about (current) state of industry-challenges and problems around biasmitigation and fairness-enhancing in data-driven AI-decisionmaking)	1	X
	Atomize the problem (Identify problem, relevance, need/importance of solution to industry)		
	Justify the value of a solution (Identify problem, relevance, need/importance of solution to industry)		
2 *Define the objectives of a solution*	**Infer the objectives of a solution from the problem definition and knowledge of what is possible and feasible** (Define requirements of artifact and utility of a generic solution towards fairness enhancement and bias-mitigation).	1	X
3 *Design and development*	**Create the artifact** (this activity includes identifying the components through systemic literature review, determining the artifact's desired functionality, and its architecture and then creating the actual artifact; designing a holistic, sociotechnical solution based on findings in the academic literature and given the requirements/objectives as stated in step #2). Also included in this step: Explore, best practices in design patterns suitable for the artefact, scoping the artefact, and define evaluation context.	4	
4 *Demonstration*	**Demonstrate the use of the artifact to solve one or more instances of the problem** (Workshop in an existing organisation reflecting upon an artificial use-case, offering a first experience of using the theoretically founded artefact in practice)	5	
5 *Evaluation*	**Observe and measure how well the artifact supports a solution to the problem.** (This involves comparing the objectives of a solution to actual observed results from use of the artifact in the demonstration. At the end of this activity the researcher can decide to iterate back to step 3 and try to improve the effectiveness of the arrtifact or to continue to communication and leave further improvement to subsequent projects)	6	
6 *Communication*	**Communicate** (the problem and its importance, the artifact, its utility and novelty, the rigor of its design, and its effectiveness to researchers and other relevant audiences such as practicing professionals, when approriate)	N/A	
Step outside the DSRM-process			
Motivate and explain the chosen methodology	**Describe** what the Design Science Research Methodology entails	2	
	Motivate why this methodology is suitable and chosen		

Step outside the DSRM-process			
Motivate and explain the chosen methodology	**Describe** what the Design Science Research Methodology entails	2	
	Motivate why this methodology is suitable and chosen		
Additional step in design and development			
Broad spectrum SLR to build a first draft conceptual, socio-technical framework for fairness-in-AI	**Explore, research** what concepts, constructs are related in the Academic Body of Knowledge, conducting a SLR on fairness and bias in the ADM-literature	3	
	Choose and define which constructs/building blocks will be focussed on, in building an artefact. Partly depending on what is feasible within the time- and effortlimits of this study.		

Table 1-1: DSRM-steps explained incl. roadmap

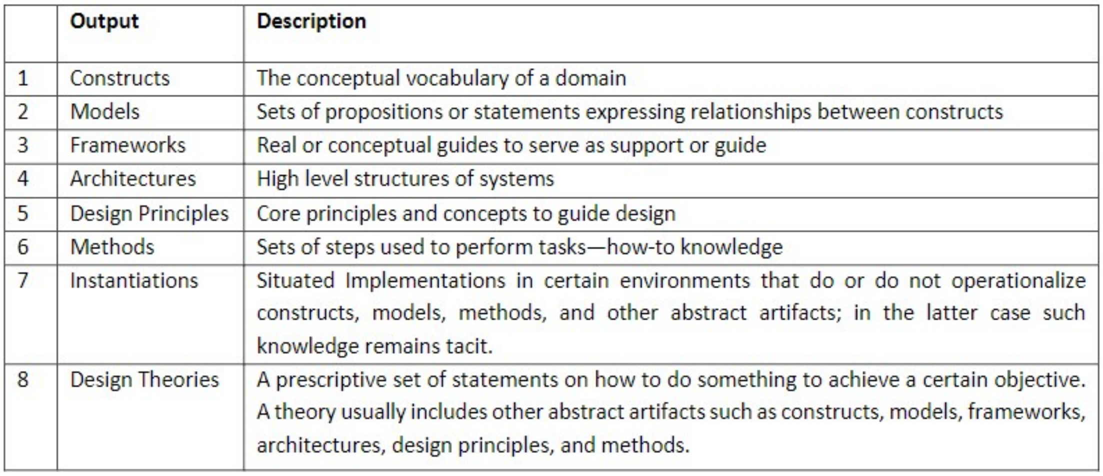
2. Research Design

Since the research goal is knowledge and utility driven, the overall research methodology which is followed comprises the Design Science Research Methodology (DSRM) of Peffers et al. (2007). The gap between industry and scientific body-of-knowledge is being bridged by developing an artefact, using the nominal process as presented earlier in Figure 1-1. In the first section of this Chapter (§2.1) Design Science Research (DSR) will be shortly introduced and some clarity will be given about what is being meant by an artefact, as well as which type of artefact will be developed. In §2.2 the methods and approaches for addressing the formulated research knowledge and design questions are being presented.

2.1. Design Science Research

The goal of behavioural science is truth, where the goal of design science is utility, where truth and utility are considered to be inseparable (Hevner, March, Park, & Ram, 2004). The design-science paradigm, is fundamentally a problem solving paradigm, in which new, scientific knowledge and understanding of a problem domain and its solution are achieved, by seeking to extend the boundaries of organizational and human capabilities, through the creation and application of innovative artifacts (Hevner et al., 2004).

Artifacts, considered as the output of DSR, are potentially, broadly defined constructs, models (abstractions and representations), frameworks, architectures, design principles, methods, instantiations (implemented and prototype systems), and/or design theories (Kuechler & Vaishnavi, 2008) (find table 2-1 below). Conceptually, a design research artefact can be any designed object in which a research contribution is embedded in the design[5].



	Output	Description
1	Constructs	The conceptual vocabulary of a domain
2	Models	Sets of propositions or statements expressing relationships between constructs
3	Frameworks	Real or conceptual guides to serve as support or guide
4	Architectures	High level structures of systems
5	Design Principles	Core principles and concepts to guide design
6	Methods	Sets of steps used to perform tasks—how-to knowledge
7	Instantiations	Situated Implementations in certain environments that do or do not operationalize constructs, models, methods, and other abstract artifacts; in the latter case such knowledge remains tacit.
8	Design Theories	A prescriptive set of statements on how to do something to achieve a certain objective. A theory usually includes other abstract artifacts such as constructs, models, frameworks, architectures, design principles, and methods.

Table 2-1: Outputs in DSR: Artefacts, source: Kuechler and Vaishnavi (2008)

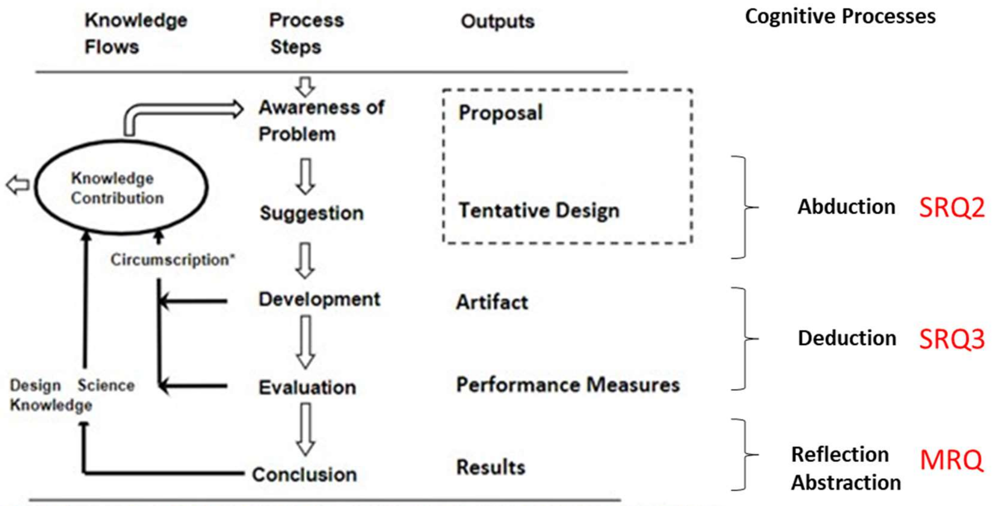
The suggestion is that a Framework, would be an appropriate artefact for this research, as the solution would be a conceptual guide to serve as support or a guide in fairness enhancement through, amongst others, identifying important design-questions and creating awareness around fairness, biasdetection and -mitigation.

[5] As Hevner et al. (2004) state: "*An artefact may have utility because of some yet undiscovered truth and a theory may yet to be developed to the point where its truth can be incorporated in the design*".

2.2. Research Methods versus Research Questions

In this section the different approaches and methods for answering the research questions are being highlighted. The different questions are constructed upon systemization suggested by Thuan et al. (2019) follow-up asking what knowledge is available in literature (SRQ1), how can this knowledge be framed/shaped (SRQ2), used and evaluated (SRQ3). Between the evaluation phase and the design-phase an iteration may be introduced to refine the artifact. At the end of this whole cycle, there will be reflected upon the knowledge and practical experiences which will be abstracted to the theory, answering the Main Research Question (MRQ).

The table below provides an overview of the different kinds of questions and used methods to address them (table 2-2), these are derived and justified from the model presented by Kuechler and Vaishnavi (2008) (figure 2-1). In DSR the tentative design is based upon abduction, as the evaluation and development are based upon deduction. In figure 2-1, below, these described knowledge flows and cognitive processes are presented, which occur in the Design-cycle and which identify the suitable approaches/methods for the different stages in the DSR-cycle, which correspond to the defined research questions (please find Table 2-2).



* Circumscription is discovery of constraint knowledge about theories gained through detection and analysis of contradictions when things do not work according to theory (McCarthy, 1980)

Figure 2-1: Knowledge and Cognitive Flows in DSR - Kuechler and Vaishnavi (2008)

Subresearch Questions	Type of Question	Method
SRQ1*	Descriptive knowledge question	Systematic Literature review
SRQ2	Design Question	(Analogous) Abduction
SRQ3	Design Question	Deduction / DSRM Case Study
MRQ	Knowledge contribution	Reflection and abstraction

*The findings around the topics are also used for the initial problem identification and motivation (section 1.4)

Table 2-2: Research Questions vs Methods / Approaches

Here the different approaches for addressing the investigative research questions are shortly motivated:

SRQ1: ***What is the current state of scientific literature (and identified best-practices) in the field of (algorithmic) bias and fairness, which concepts are of importance and interrelated and which consensus is there to be found in the academic literature around this topic?***

A theoretical, i.e. conceptual framework should provide insight into concepts, interrelations, scope, complexity, and content of the research area, in this study being approaches, and forms of fairness, bias, mitigation methods, metrics, coherence and applicability in context. The artefact designed upon this conceptual framework is not intended to be exhaustive in methods, scope, and so forth, but is a first "proof-of-concept" (design probe), which should be developed through iterative cycles. The primary focus is to identify and map most common types of fairness, bias, and used mitigation measures and approaches. A broad-spectrum literature review has been carried out for this. By keeping the given time for this research in mind, this review is limited to the state-of-the-art literature providing an overview of the field (please find Appendix 1 and Chapter 3 for details on how the literature review has been executed). The main goal of this first step is to build the contours of the framework, it is justified that a non-exhaustive approach is taken. Several used literature surveys provide an overview of findings in the literature, where peer-reviewed articles form the basis of.

In Chapter 3 (Conceptual Framework Design) this descriptive, knowledge (sub)research question is examined, and treated.

SRQ2: How can the fragmentary literature around fairness, algorithmic bias and mitigation methods, tools and metrics be structured in a modular, practical, easy to digest, structured and extendable, informative overview i.e., global framework to meet the defined solution and research objectives?

Requirements are derived from different parts of the problem formulation. To meet the earlier defined solution objective of integrability in the engineers workflow, and light-weighted, a format will be determined by using abductive reasoning[6], creativity, and searching for technical design possibilities in academic literature. Next to this, also searched for is design patterns; known systemizations of embedding ethical values in design.

A premature intuition is to pinpoint pausing points for fairness- and bias awareness, detection and mitigation at the model developed by Aysolmaz, Iren, and Dau (2020). This model is closely related and therefore quite similar to the well-known CRISP-DM model and has been developed within a Delphi-study with experts in which findings suggest that it fits the ADM-development-processes in an effective way (figure 2-2).

[6] This research question is a design question approached by analogous abduction, which also comprises creativity (please find **Appendix 8** for background information on abduction).

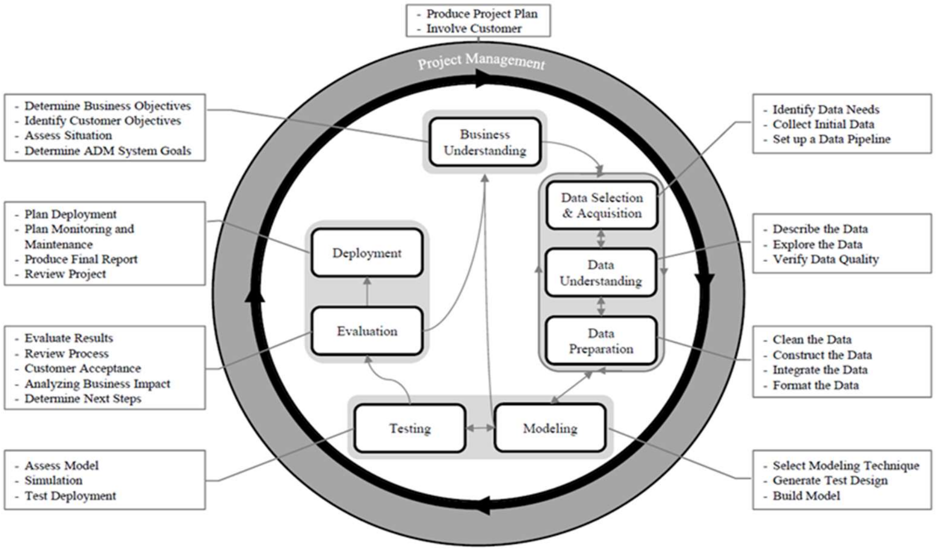


Figure 2-2: Final Algorithmic Decision-Making system (ADM) development process adapted from Aysolmaz et al. (2020)

However, this does not meet integrability in all kinds of workflows, so that also will be searched for design alternatives. A possibility could be to refer to CRISP-DM-stages within the framework, but not limit the framework towards this methodology, so that it is method-agnostic.

Also, it needs to be considered, and explored which design is concise enough to be evaluated within time limits of this study.

In Chapter 4 (Design- and Development) this design (sub)question is examined, and treated; here the technical artefact design on both format, and scoped content are considered.

SRQ3: How to use i.e., demonstrate and evaluate the designed artifact (= "proof-of-concept")?

The artefact could be demonstrated in an expert organization where data scientists and other stakeholders could be questioned about the perceived usefulness of the artefact. The evaluation strategy has to be investigated and determined, so that the artefact is testable.

This design (sub)question, is treated in Chapter 5 (Artefact Demonstration).

3. Conceptual Framework Design

In this Chapter the setup of a first draft, overall conceptual Framework is being elaborated, in which fairness alignment is pursued in a sociotechnical way from the outset. The first section describes the search process which has been conducted to answer SRQ1, and which is stated to identify relevant academic work of previous scholars, across various domains, providing the conceptual building blocks which form the foundation of the framework. Next in section 3.2 the overall literature findings are presented. Main finding hereof is that very often fairness is arguably, addressed by technical approaches solely. In section 3.3, and subsections, the different constructs, i.e. building blocks are explained and presented in a conceptual model, which displays the current (also challenging) technical fairness approach in ADM's. Finally in §3.4 this framed and isolated, technical framework will be positioned in a wider framework, which promotes a holistic, sociotechnical approach towards fairness embedding: the overall, conceptual "Fair-ML" framework.

3.1. SLR – Search Strategy

As for the first step in building a conceptual framework, a Systematic Literature Review (SLR) is conducted. The methodology of Wolfswinkel, Furtmueller, and Wilderom (2013) is used for this SLR to answer the broadly defined knowledge question (SRQ1) which is necessary to identify the most common, interrelated, important concepts - in the state-of-the-art scientific, (algorithmic) bias and fairness literature. A non-exhaustive approach is being pursued and the search is limited to articles centred around the "supervised ML – classification" domain, as regression, unsupervised ML and reinforcement learning are beyond the scope of this Master Thesis. Furthermore, it must be said that the primary focus is on possible interventions in the conceptual, i.e., building phase of ADM's.

In this SLR - as to the extent available - mostly surveys are being searched for in combination with deepened peer-reviewed fair-AI literature and conference papers (proceedings from FAT/ML, CHI, ICCS en NIPS), to make a "sketch" of the domain and possible building blocks to build the overall framework. Due to the broad domain, enormous number of papers and scattered research in combination with the limited time available for this research, non-exhaustiveness is pursued and justified.

In conducting the search for articles, two search-engines i.e., databases - EBSCO Host and Google Scholar - are being used, where the search is being performed by both search-terms (e.g., ALGORITHMIC BIAS AND SURVEY OR LANDSCAPE SUMMARY and synonyms) combined with Boolean operators (so called building-blocks) as well as using the back- and forward snowballing method. Primarily academic peer-reviewed literature from the last decade (2010-2020) restricted to papers in English have been included, and to a limited extend grey literature (conference papers and academic preprints from ArXiv).

The longlist of articles which were found upon the above-mentioned search terms, inclusion criteria were screened on title and abstract. The first rigor list of articles was thereafter assessed by additional selection criteria.

For more details on this SLR, please find Appendices 1 and 2, where amongst others the shortlist is presented, as well as the key-findings from used papers and reasons for in- and excluding certain articles.

3.2. SLR – Background and Overview of Findings on ADM's

AI has the potential to benefit the whole of society. However, as industries increasingly use data-driven, algorithmic, decision-processes, the urgency grows for mechanisms which secure that systemic, unfair (e.g., discriminatory) practices are being avoided. These ADM's, where data and ML algorithms form the basis of, have the potential to not only mirror, but also amplify our pre-existing, deep-rooted, human-biases on an unprecedented scale, or even introduce new forms of unwanted biases (Cofone, 2018; Commission, 2020; Rovatsos et al., 2019). Algorithms are often (mistakenly) interpreted to be objective, but since an algorithm is based upon reversed engineering this is unfortunately and undecided not the case. The applied decision-rules are not programmed by humans, but are rather inferred from the given data (Kroll et al., 2016). In other words: the algorithms mimic the historical (human) decision-making behaviour and since algorithms operate systematically and are deployed on a large scale, mirrored existing biases are potentially both amplified and systemized, justly accompanied by growing concerns (Koene, 2017) about ethical harms, whereby Fairness[7] in particular.

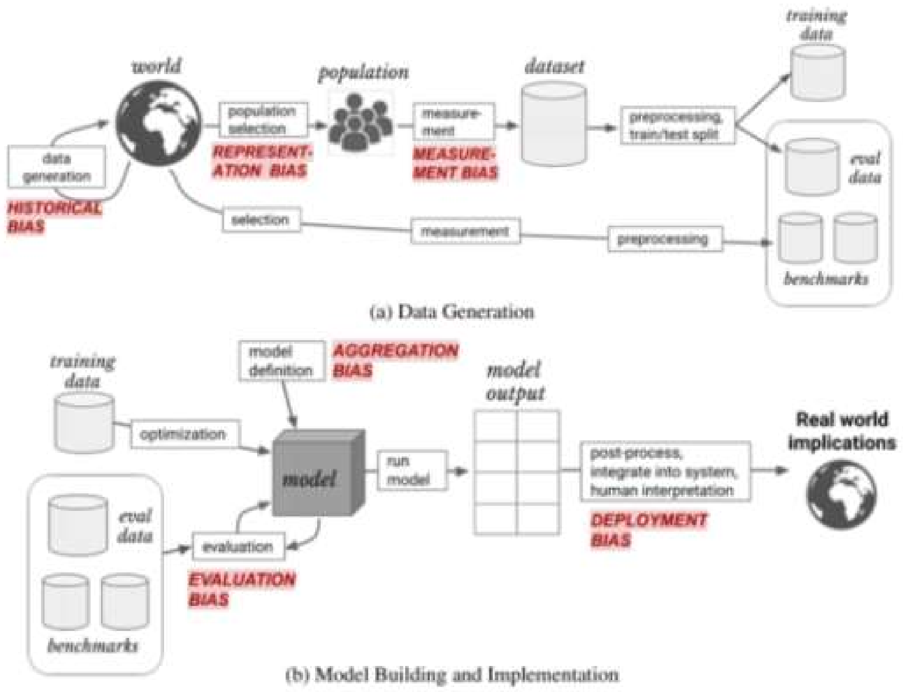
In practice, algorithms "in the wild" already showed some harmful biases in several sectors, like in financial services, local government, crime & justice and recruitment (Rovatsos et al., 2019). To safeguard against discrimination on protected and social attributes (like race, gender, colour, national origin, religion, sex, disability, or family status, to name a few), developers need to be aware how and where (harmful, problematic) bias can be introduced in the process and how to prevent or mitigate this unwanted bias[8]. However battling unfairness and mitigating bias appears to be challenging for practitioners for different reasons (scattered research (Manseau & Mbuko, 2020), systemic evaluation is failing around which approaches work best (Ntoutsi et al., 2020; Rovatsos et al., 2019), and translating literature findings in pragmatic processes (Lee & Singh, 2020)).

There are many potential causes (Arrieta et al., 2020) and types of biases and different kinds of unwanted bias can therefore enter the ML-systems in multiple ways (please see also Figure 3-1):

- the source (bias manifestation in data), the data which is most likely contaminated with human (historical, institutional, and social) bias;
- data collection: the selection and sampling of data;
- the design of the algorithm (which features, and constructs are used, who labelled the training-data);
- data generation;
- the machine learning algorithm and techniques itself; e.g.: NLP amplifying gender bias found in text corpora and word embedding models (Rozado, 2020).

[7] It is therefore not surprising that "justice and fairness" form one of the basic ethical principles in existing AI guidelines (Jobin, Ienca, & Vayena, 2019) and is also representing a building block of RAI and Trustworthy AI (Arrieta et al., 2020). Fairness is the past years grown into a major subfield of Machine Learning, complete with several dedicated archival conferences like ACM FAT (Chouldechova & Roth, 2020) and research communities like FATML, DADM (Olteanu, Castillo, Diaz, & Kiciman, 2019; Veale & Binns, 2017) and conferences like AIES (AAAI Conference on Artificial Intelligence, Ethics and Society) (Olteanu et al., 2019).

[8] While bias originates from a neutral term, referring to "deviation from a standard", in the ML-community the word "bias" has a negative connotation (Danks & London, 2017), and where mostly is being referred to an inclined or prejudiced decision for or against one individual or (sub)group considered to be unfair *(Aysolmaz et al., 2020; Ntoutsi et al., 2020).* These types of discrimination are predominantly divided into two kinds of discrimination: discrimination upon protected and unprotected characteristics.



(a) Data Generation

(b) Model Building and Implementation

Figure 3-1: Presentation of the ML Pipeline and Possible entry points of biases adapted from Suresh and Guttag (2019)
(a) The data generation process begins with data collection from the world. This process involves both sampling from a population and identifying which features and labels to use. This dataset is split into training and evaluation sets, which are used to develop and evaluate a particular model. Data is also collected (perhaps by a different process) into benchmark datasets. (b) Benchmark data is used to evaluate, compare, and motivate the development of better models. A final model then generates its output, which has some real-world manifestation. This process is naturally cyclic, and decisions influenced by models affect the world that exists the next time data is collected, or decisions are applied. Red indicates where in this pipeline different sources of downstream harm can arise.

The visualisation of biases presented in figure 3-1, is only used as an example, and are only one of the non-exhaustive bias-taxonomies provided by literature (find Appendix 3 for the explanation of these and other identified forms of biases and taxonomies). The different types and sources of bias are an identified building block in our holistic angled framework as being aware how bias can sneak in, which risks are associated with it and detecting bias are preconditions for properly mitigating it. A taxonomy of bias and its sources will be part of the framework and is substantiated in §3.3. As some supervised ML techniques (such as word embeddings; (Rozado, 2020)) are potential sources of additional bias, a separate section is devoted to a taxonomy of supervised ML-techniques.

An important insight remains that fairness and bias are interconnected (which also can be deduced form the earlier definition); as fairness defines indirectly what is ought to be bias (bias is the derivative) and how to approach bias-mitigation, so that the desired state of fairness is achieved. As fairness is the overall goal to be achieved as well as the fact that there are several approaches towards achieving and formulating it, fairness is a distinct (even the most important) building block in our framework (§3.6).

The challenging part is that most papers bound the system of interest narrowly by focussing on technical approaches on bias and fairness; best approximations or fairness guarantees are being used based upon hard constraints, considering the ML-pipeline with its inputs and outputs. These hard constraints are being used in efforts to mitigate bias, entailing formalized, mathematical fairness metrics, which quantify deviations from various statistical parities that are related to what ought to be fair and (since it is quantified) can be used to make trade-offs between utility and fairness. (Fazelpour & Lipton, 2020). These technical interventions are commonly categorized within pre-, in-, and postprocessing techniques, where respectively these methods i.e., statistical interventions focus on data, ML-algorithm (optimization at training-time) and ML-model (Choraś, Pawlicki, Puchalski, & Kozik, 2020; Ntoutsi et al., 2020; Rovatsos et al., 2019).

However, despite the volume and velocity of published work - grown into an enormous amount of algorithmic fairness literature (mainly focussing on supervised learning) - "*..the understanding of fundamental questions related to fairness and ML remains in its infancy.*" as Chouldechova and Roth (2020) describe. Real-world fairness challenges are not abstract, constrained optimisation problems Veale and Binns (2017) underpin. Therefore, a technical approach in isolation is not a panacea for all sorts of bias and unfairness; a holistic, sociotechnical view and approach is needed (Fazelpour & Lipton, 2020; Grasso, Russell, Matthews, Matthews, & Record, 2020; Ntoutsi et al., 2020; Selbst et al., 2019). The issue with addressing fairness by using a technical approach solely, using computational tools and mathematics, is that the problem can look like it is solved cleanly, but only because it has been defined so narrowly by abstracting away any contexts surrounding these systems (Selbst et al., 2019). These problems are institutionally and contextually grounded, and therefore it requires embracing a sociotechnical view and understanding by also taking (reference to) context and stakeholders into account (Fazelpour & Lipton, 2020; Selbst et al., 2019; Veale & Binns, 2017).

Researchers amongst Kroll et al. (2016) and Aysolmaz et al. (2020) stress that the ethics principle should be incorporated in the design and emphasize that, while there are many ex-post interventions, it is important to discover and eliminate bias (ex-ante) during the development of these ADM's. Unintended biases are harder to eliminate if ignored at the beginning (Cramer, Garcia-Gathright, Springer, & Reddy, 2018; Kroll et al., 2016).

Since the research aim is to take a holistic, sociotechnical approach, literature upon the technical side of fair ML is included as well as the literature discussing other approaches towards contextual fairness, and fairness in ML that emphasize the importance of questioning the design, used data, trade-offs and so forth (such as Fairness Checklists (M. A. Madaio et al., 2020), Model-cards (Mitchell et al., 2019), Datasheets/statements (Bender & Friedman, 2018; Gebru et al., 2018), and explanation building around the ML-model (Kusner, Loftus, Russell, & Silva, 2017). The combination of important design questions, checklists, questioning around the data origin and usability (source) and the available technical tooling will eventually shape the framework which will be created for guiding the development of fair ADM's given its context.

3.3. Conceptual Model: Technical Fairness Approach

As a starting point for the incremental, iterative development of a first draft framework, particularly literature surveys are being used which provide an overview of the developments regarding fairness promotion and bias mitigation in ML. To mention some: studies like Mehrabi, Morstatter, Saxena, Lerman, and Galstyan (2019), Arrieta et al. (2020), Caton and Haas (2020), and Ntoutsi et al. (2020) provide an overview of the area of fairness, bias, intervention strategies, and mitigation-techniques classified according to the different intervention types that correspond to the common classification of pre-, in-, and post-processing methods.

This overview can be used as a guideline for handling it in a systematic manner of the complex issues and considerations surrounding the use of fair ML. Additional added value will depend strongly on the feasible level of detail of this framework and the possible rules of thumb that are included in it.

To subdivide the problem, we will highlight the different aspects in this domain, focussing on different kinds and sources of biases, different common (sometimes mutually exclusive) fairness definitions, forms of discrimination, ML-techniques, and bias-mitigation methods, tools, and techniques. All these different aspects to be unravelled from literature form the building blocks of the this conceptual framework. For the substantive findings in literature about bias and all its important interrelated component's a state-of-the art, broad spectrum, literature review has been carried out (§3.2). Details on the SLR-output, and explanation of the SLR-process are included in Appendix 1. A sketch of the concepts and their interrelationships identified in literature are presented in figure 3-2. In the next (sub)sections the different concepts (numbered from 1 until 5) will be explained shortly. For more details and literature groundings on fairness, bias and mitigation we refer to the appendices (respectively Appendix 9, 4, 5, and 10).

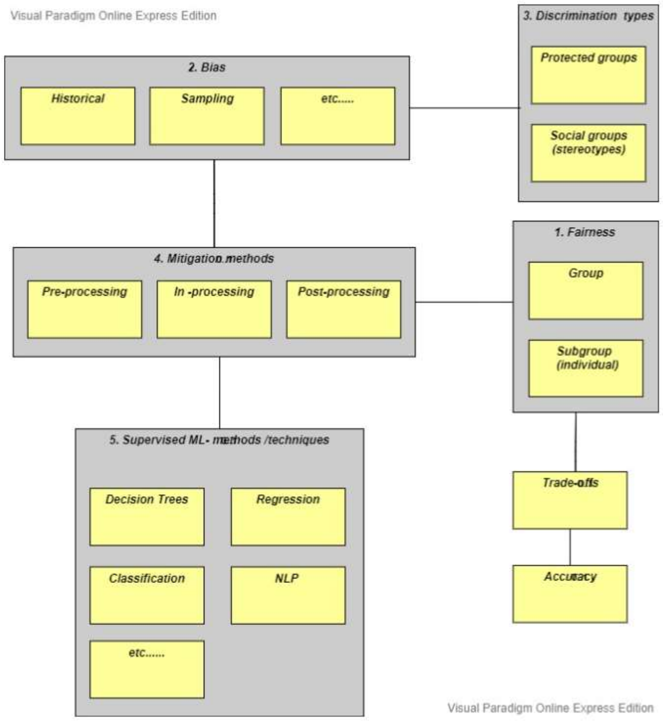


Figure 3-2: Sketch of conceptual model around bias, fairness, discrimination, and bias-mitigation in supervised ML

3.3.1. Concepts – Fairness [1]

In ML-literature (consensus on) a universal definition of fairness fails, as it is context-dependent and volatile, therefore fairness is often approached by metrics. The statistical fairness definitions provide an unambiguous terminology, as it provides a conscious choice around the pursued outcome fairness in a certain context provides clarity what is bias and how to deal with this matter.

The different fairness-metrics posed and used in literature have different underlying principles, and are to be divided in two common mappings, that of individual and (sub)group fairness (Chouldechova & Roth, 2020; Mehrabi et al., 2019; Tal et al., 2019):

Individual Fairness: similar individuals should be treated similarly.

Group Fairness: the proportion of members in a protected group receiving positive classification is identical to the proportion in the population.

In table 3-1 the most common metrics adopted from (Mehrabi et al., 2019) are presented (for details please find Appendix 9):

Name	Group	Individual
Demographic parity	X	
Conditional statistical parity	X	
Equalized odds	X	
Equal opportunity	X	
Fairness through unawareness		X
Fairness trough awareness		X
Counterfactual fairness		X

Table 3-1: Categorizing different fairness notions into group vs individual types. Source: Mehrabi et al (2019)

These metrics are mainly based upon the confusion-matrixes' Type I and Type 2 Errors (examples given in Appendix 3), which correspond to falsely designating a true or false outcome.

However, different scholars argue that a technical approach, by using these fairness metrics, are not sufficient for dealing with fair-ML. Contextualizing and situating fairness is at least as important a step (Draude et al., 2019; Selbst et al., 2019). Therefore, in the framework a sociotechnical, contextual approach will be encouraged by considering import questions and checklists in the design of the ADM wherein stakeholders' concerns are considered, and the context of the algorithm and its decision are being taken into account. This wider form of (contextual) fairness will be elaborated upon later in this thesis, in light of a particular use-case chosen for evaluation purposes.

3.3.2. Concepts – Bias [2]

In the context of data driven AI decision-making the term bias is mostly referred to Algorithmic Bias: systemic harms and unfairness in terms of discriminatory outcomes (Aysolmaz et al., 2020; Koene et al., 2018; Ntoutsi et al., 2020; Olteanu et al., 2019; Rovatsos et al., 2019).

In this research with respect to (algorithmic) bias we consider bias to be negative bias, potentially leading to unfairness, and follow the definition from (Aysolmaz et al., 2020; Friedman & Nissenbaum, 1996); Ntoutsi et al. (2020), and Aysolmaz et al. (2020) which states: *"Inclination or prejudice of a decision made by an AI system*[9] *which is for or against one person or group especially in a way considered to be unfair".*

Sources and causes of bias

There are several sources of bias identified in literature, as indicated by (Arrieta et al., 2020; Barocas & Selbst, 2016; Datta, Fredrikson, Ko, Mardziel, & Sen, 2017; Zhong, 2018). Mostly the types of biases refer to their origin, the process by which they enter the model or system or the types of affected parties they discriminate against. As for the first 2 types of biases mentioned (origin and process), we discuss them here, where we also look at the different entry points of bias in the development process. As for the latter bias, we refer to the different kinds of fairness notions (§ 3.3.1).

Non-exhaustive bias-sources identified are summed up here:

***Skewed sample/datasets*:** bias within the data acquisition process, amongst others resulting from reporting, sampling, and selecting data, so that some (protected) classes are underrepresented (leading to unintentional discrimination).

Tainted data/examples: errors in the data modelling definition and wrong feature labelling (mostly rooted in practices, institutions, and attitudes); e.g.: ML-algorithms replicating the HR-managers preference for males (=bias) in selecting job-applicants.

Limited features: using too few, unreliable and/or less informative features from minority classes leading into lower accuracy in the prediction of minority classes.

Proxy features: there may be correlated features with sensitive ones that can induce bias even when the sensitive features are not present in the dataset.

Taxonomy of bias

In literature there are many, different taxonomies on bias found. This is not surprising, as in the psychological literature alone, there are over 100 identified types of cognitive biases in human decision-making (Baer, 2019) (where algorithms mimic human behaviour). Olteanu et al. (2019) point out that there are both general types of bias, as well as domain-specific types of bias, referring to a whole pallet of specific biases which may occur in using social data.

As an agreement on a vocabulary or taxonomy of bias is missing, for here the broadly defined bias-classification (within algorithms) of Tal et al. (2019) is being used, which differentiates between the main sources of bias in ML-algorithms: Data Bias, Human Bias and Algorithmic Processing Bias as shown in figure 3-3:

[9] Ntoutsi et al. (2020) added the underlined text "made by an AI-systems" to the original widely cited definition of Friedman and Nissenbaum (1996).

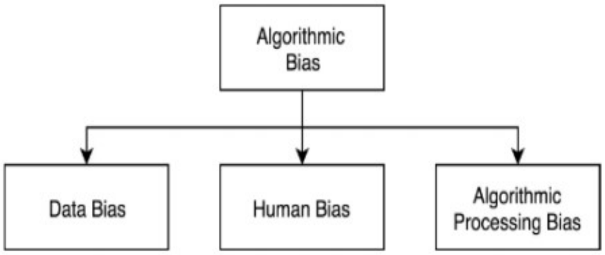


Figure 3-3: Classification of algorithmic biases adapted from Tal et al. (2019) Where Data Bias, Human Bias and Algorithmic Processing Bias, respectively refer to: bias caused by the training i.e., input data, bias caused by humans (e.g., inappropriate system use) and bias added during algorithmic processing.

In Appendix 4 there is a more granular classification of different biases to be related to this broader taxonomy, and which are not exhaustive and/or mutually exclusive. However, the taxonomy covers various kinds i.e., sources of bias, so that it provides guidance in mitigation by creating a richer space within which to assess whether a particular bias is potentially present, merits a response, and what mitigation approach or corrective measure is appropriate. This taxonomy could be used as a starting point for a checklist to provide awareness around biases which may occur.

As can be seen later (§3.3.4), the different bias-sources (data, humans, processes) are partially covered by diverse approaches i.e., strategies for bias-detection and intervention, such as checklists, datasheets, and so forth. Also it must be mentioned that both data pre-processing's (e.g., word embeddings/NLP) and the type of ML-model chosen, affect both "explainability" of outcomes and the risk of bias (please also find Appendix 11).

3.3.3. Concepts - Discrimination types [3]

In literature the most common used distinction is been made between direct (explicit) discrimination and indirect (implicit) discrimination.

Direct Discrimination

Direct discrimination (a.k.a. disparate treatment) consists of rules or procedures that explicitly impose disproportionate burdens on minority or disadvantaged groups, based on sensitive, by law protected attributes such as gender, race, age, religion, etc. (Ruggieri, Pedreschi, & Turini, 2010).

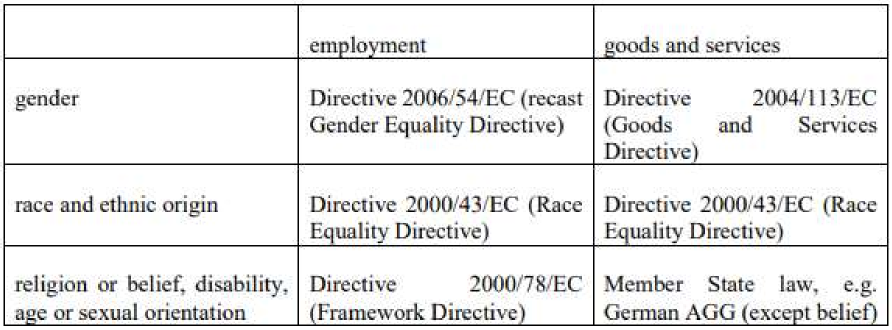
Indirect Discrimination

Indirect discrimination (a.k.a. disparate impact) consists of (apparently neutral) rules or procedures that, while not explicitly mentioning discriminatory attributes, intentionally or not impose the same disproportionate burdens (Ntoutsi et al., 2020; Ruggieri et al., 2010). This is also known as proxies or redlining. A well-known example of redlining is when a residential zip-code of an individual, apparently a neutral attribute, is an influencer in an algorithmic decision outcome, but highly correlated with the sensitive attribute (like race or age) due to the composition of residential areas.

The GDPR brings an extra challenge here, because in certain contexts companies are not allowed to store and use data on protected characteristics (please also see Protected groups in EU-law), so that proxies are difficult to detect in these circumstances.

Protected groups in EU-law

For the control of discriminatory decisions, the principle of equality and the prohibition of discrimination (Art. 20, 21 EU Chapter of Fundamental Rights, Art.4 Directive 2004/113 and other directives) applies. As can be seen in table 3-2, the protected groups by EU-law are dividable in: gender, race (including characteristics as colour), ethnic origin, religion / belief (political or other opinions), disability, age, and sexual orientation.



	employment	goods and services
gender	Directive 2006/54/EC (recast Gender Equality Directive)	Directive 2004/113/EC (Goods and Services Directive)
race and ethnic origin	Directive 2000/43/EC (Race Equality Directive)	Directive 2000/43/EC (Race Equality Directive)
religion or belief, disability, age or sexual orientation	Directive 2000/78/EC (Framework Directive)	Member State law, e.g. German AGG (except belief)

Table 3-2: Overview of Secondary EU Anti-Discrimination Law adapted from Hacker (2018)

In the framework primary attention will be given to the attributes gender, race, and ethnic origin, as these are the most protected attributes in EU-law and legislation. Besides that, most papers deal with mitigation methods for discrimination on race and gender.

Social Groups (stereotypes)

Social groups are not protected by law but evolve over time. These are individuals who identify themselves to particular groups (e.g., transgenders). Even though these groups are not protected by law, discrimination on these grey lines, can lead to harm and reputational business risks (Baer, 2019).

3.3.4. Concepts – Mitigation Methods & Remedies [4]

Both Kroll et al. (2016), Ntoutsi et al. (2020), as Aysolmaz et al. (2020) underline that an important strategy and step towards fair algorithms is to be taken in the design of the ADM-system. A "better safe than sorry" mentality is a prerequisite for preventing harm, especially in high-stake decisions affecting people; fair-washing by reactionary technical solutions and deploying algorithms which are biased in their design is not advisable.

The desired approach would not be to develop an algorithm and looking ex-post how to correct for undesired bias and making things fair, but first looking at the context: who is affected by the algorithm, engaging with stakeholders, taking in consideration both risks and impact of bias, asking "what is ought to be fair in the particular context".

When risks of unfairness, and bias are high and the possible impact of the (automated) decision on affected people as well, then it justifies resources, time, and costs in a very thorough analysis on possible fairness and bias issues. Even the question "to build or not to build" should (then) be a consideration in the design and development of an ADM-system (Crawford & Calo, 2016).

Before mitigating bias, one should be aware how bias may creep into the ADM from the beginning. The provided taxonomies (Appendix 4) create awareness of potentially biases that can occur through the entire process-cycle resulting in different forms of unfairness.

Beyond the computational pre-, in, and postprocessing techniques for binary classification used for bias-mitigation as presented in Appendix 5, several strategies, remedies, awareness tools and methods are identified for intervening on (unwanted) biases. These on literature grounded remedies are summarized below, and detailed in Appendix 10 with reference to related literature:

- Impact and Risk Assessments (AIA's);
- AI-Fairness Checklists;
- Using questions surrounding 4P's (People, Place, Power and Participation), for a systemic, contextualizing-sociotechnical-approach;
- Datasheets or statements for datasets;
- Common pitfalls in developing ADM´s;
- Questions on ethical algorithms (please see Appendix 6);
- Model cards (please see Appendix 7).

Some of these will later be included in a first design probe, i.e. artefact for evaluation.

3.3.5. Concepts – Supervised ML Taxonomy [5]

As most papers, this research is focussed on classification algorithms (so called classifiers) which are part of the Supervised Machine Learning (SVL) family. SVL differs from Unsupervised Machine Learning (UML) in that SVL learns on previous examples, given an outcome (so called label). Within unsupervised Machine Learning one attempts to make a natural grouping without a specific target/outcome.

An inexhaustive taxonomy of Classification methods is to be listed as follows: Logistic Regression, Support Vector Machines (SVM), Naïve Bayes, k-Nearest Neighbours, Neural Networks, Decision Trees (Random Forest and Boosted Trees). Depending on the ML Model chosen, the outcomes are less or more interpretable. When decisions are more interpretable, it can be verified upon which constructs, independent variables the output is generated and if problematic bias is present. Black-boxes like Neural Networks are not interpretable, so that more attention should be given to bias-issues, although it is more challenging as well.

Special attention needs to be given to the ML-pipeline in which in some settings additional biases are lurking. A general ML Pipeline can be presented as follows:

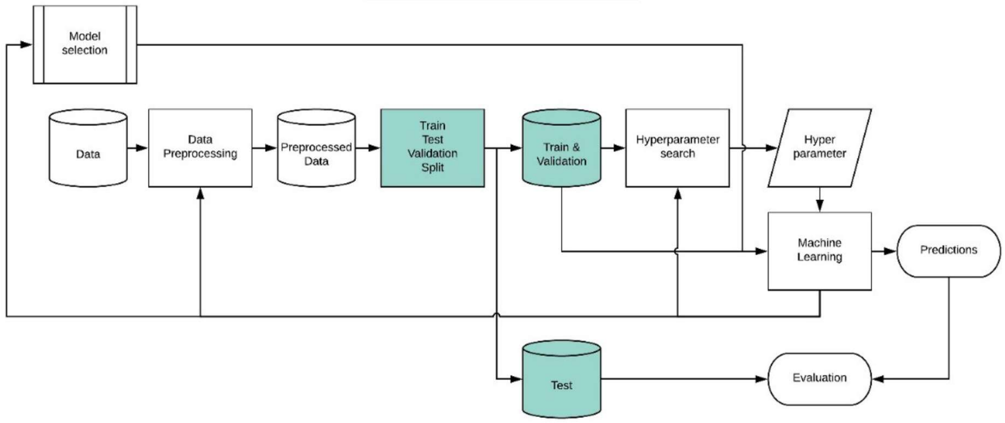


Figure 3-4: ML Pipeline, source: https://medium.com/datadriveninvestor/my-machine-learning-workflow-7576f7dbcef3

In some pre-processing techniques additional bias-risks are present. One of which is NLP; where word embeddings are being used, which empirically exhibit and potentially can amplify cultural stereotypes and (demographic) biases, because of the strong statistical effects the underlying methodology can induce. So, considering the ML-model used and the ML-pipeline also creates awareness on possible bias-issues and triggers certain additional techniques to be used.

3.4. Conceptual Model: Sociotechnical Fairness Approach

Combining the knowledge, insights, as well as previous building blocks, a draft design of a conceptual, so called “Fair-ML” framework, promoting a holistic, socio-technical approach towards fairness in supervised ADM’s is made, which is presented in figure 3-5:

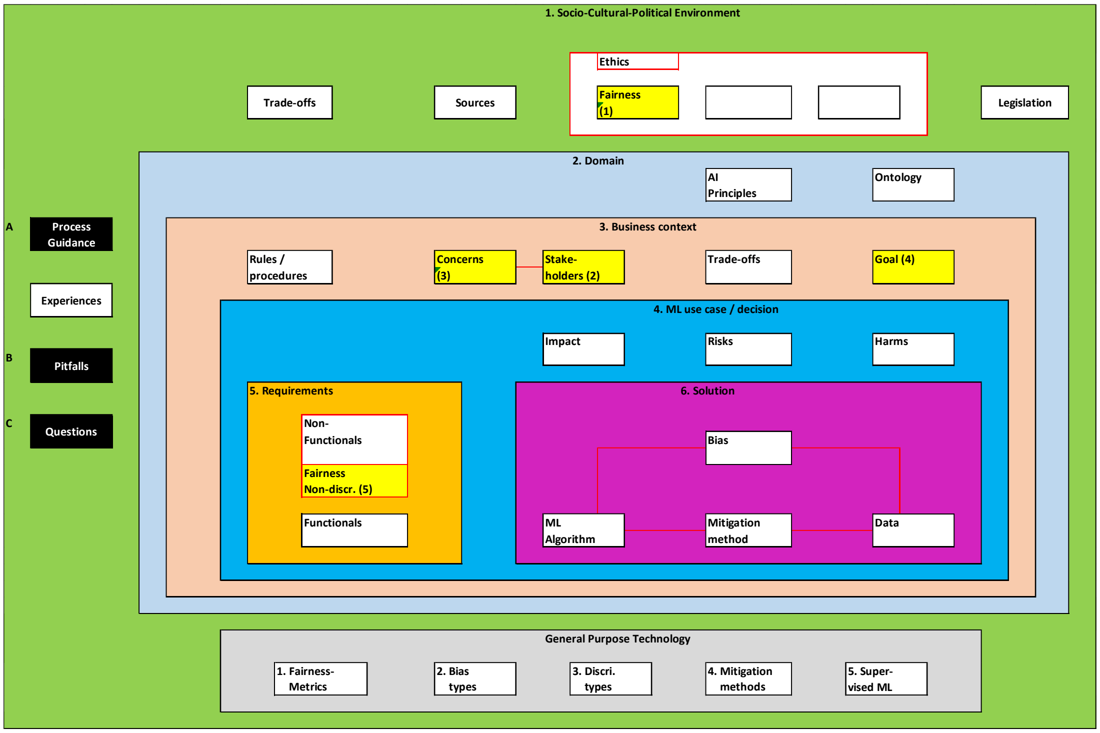


Figure 3-5: Framework in rigor form (first draft of the “Fair-ML”-Framework)

The big difference between the earlier presented conceptual model, is that fairness is presented in the wider context of the socio-, cultural environment, and not limited to discrimination, i.e., unbiased fairness metrics. The earlier presented conceptual model with a technical fairness approach is here presented as “6. Solution” and positioned within the higher level fields, of ML use-case, Business Context, Domain and ultimately the Socio-Cultural-Political Environment (in ascending order from micro- to macrolevel). In Appendix 14 this conceptual framework and the holistic view it envisions, is explained in its fullest form. As for this study the most important insight in building the artefact is that the framework entails a contextual, top-down approach in finding answers to a fair design.

The conceptual model “Fair-ML Framework”, is used to decide what the artefact should include, and gives the big picture, i.e. overview of the subject matter of developing fair ADM’s. Building the framework is a huge project, requires detailed design and specification, which is out of scope in this study. Also, by means of the given research-goal, this is justified as the artefact is meant to be a “growth framework” and will be tested in its incompleteness. After a first test, a “continuation approval” in the field the artefact can be further expanded and refined in multiple iterations, in future research. If the artefact has proven its right to exist, it can mature, by adding unused concepts of the above framework, and insights which appear by using the artefact in practice.

In the next Chapter, the design and development of the scoped artefact will be elaborated upon, in reference to the overall conceptual framework. A coherent set of constructs (yellow marked, and numbered 1 until 5, in figure 3-5) is chosen to construct a design probe, i.e. artefact.

4. Design and Development

Now, in the first cycle of the DSR-methodology a part of the conceptual model `Fair ML framework` described in the previous Chapter will be realized in an artefact, named the AI:CEID toolkit. In §4.1 the scope of this artefact is demarcated and tailored, wherein §4.2 the actual artefact is built, and highlighted in terms of format and content. Also, this section contains explanations around the intended use of the artefact, by providing an example.

4.1. Scoping and Tailoring the Content of the AI:CEID Toolkit

In downsizing the research scope to what is feasible within time, and effort given for this study, the artefacts' prototype focusses on guiding the embedding of fairness within the conceptual stage[10] ADM-design. Higher level principle of the framework, is that it is essential that fairness is proactively, addressed, and embedded from the start. The conceptual stage, is a crucial phase, in which the context is explored. Here a similarity is drawn with CRISP-DM's Business Understanding-phase where Goals, Stakeholders, -Concerns, and Goals (referring to respectively, (2), (3), and (4) in the Fair-ML framework, Figure 3-5) are examined. Therefore these concepts are treated as a coherent set of constructs for the scoped artefact. The building blocks of the fair-ML framework, Process Guidance (A), Pitfalls (B), and Questions (C) can then be substantively focused on this phase of design, promoting a sociotechnical approach in design.

Therefore it makes sense, to first accentuate the artefact on embedding Fairness (1) at the level of this stage, as this is the starting phase of building an ADM. Higher level fairness norms should guide the non-functional, fairness-requirements (5) to be embedded in the AI-solution.

The concepts Data, Risks, Harms, Bias, ML-Algorithms, Mitigation Methods, Trade-Offs, and Rules/Procedures are not operationalized within this premature artefact. Data, Bias, and Mitigation Methods are considerations which are essential, but are focused upon later down the ML-Pipeline.

Second way of scoping this research is to choose a context, in which commonly ADM-solutions are deployed/designed and in which unfairness is a relevant, risky, impactful issue. At this point, hiring is chosen as a suitable, and illustrative, problem context (which is substantiated in §4.1.1).

Before crafting the prototype (artefact 0.1 version), an additional literature study will be done on methods of embedding fairness in ADM-design, and fairness in relation to the domain of hiring. On top of the latter a HR-expert, descriptive, multiple case study is conducted to unravel important concepts, and reveal the visions of HR-experts on the topic of fairness.

Herewith the artefact v0.1 can be illustrated, and tailored to both specific hiring context, and conceptual stage of ADM-design.

In the first two sections of this Chapter, the literature review and findings are presented of respectively research on perceived fairness in Hiring, and methods for embedding fairness in ADM-design. In §4.1.3 the main insights of the sideways research on HR are presented, as in §4.1.4 the actual artefact is crafted.

[10] Corresponding to the problem formulation in the Business Understanding phase of CRISP-DM

4.1.1. Perceived Fairness in Hiring

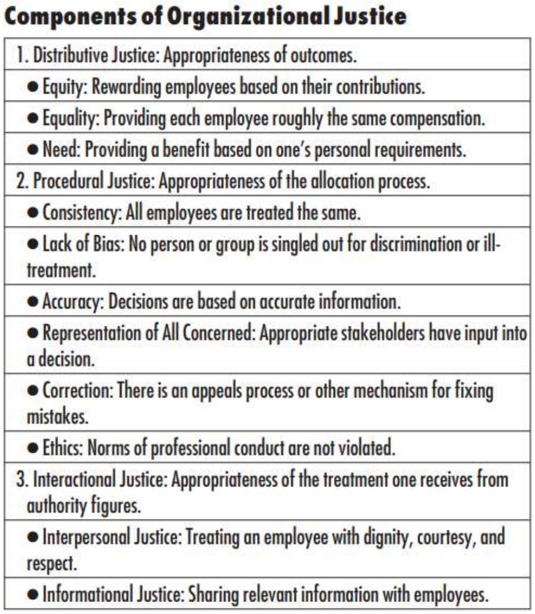
Hiring is considered a suitable, relevant, social context to be used for demonstrative- and evaluation-purpose of the artifacts' prototype as it is a domain where:

(1) AI is currently, and increasingly being used;
(2) High-impact decisions are made (as it determines who is allocated a job or not);
(3) (Un)Fairness is a relevant, ethically problematic issue.

Here referring to the articles of Leicht-Deobald et al. (2019), and Köchling and Wehner (2020) as understatements, of academically pronounced relevance of fairness-embedding endeavors in the development of ADM's within this domain.

As earlier mentioned, mostly in practice fairness-approaches within ADM's are restricted to technical, fairness-metrics, but contextualizing and situating fairness (Draude et al., 2019; Selbst et al., 2019) from the start is considered at least equal important. The conceptual framework recognizes (in its top-down approach) that fairness is subjective and highly context-related (domain being only one element). Therefore fairness is examined, and conceptualized, within relation to this specific context of hiring.

As for time-limitations, the strategy justified here is to search for widely cited articles which reveal the fundamental theories on localized (perceptions of) fairness within the domain of hiring (please find Appendix 13 for detailed setup). The Organizational Justice Theory (OJT) is found to be a widely accepted, appraised, and researched theory on fairness[11] in an organizational context[12], in which 4 dimensions are centralized: procedural, distributional, informational, and interpersonal fairness (Colquitt, Greenberg, & Zapata-Phelan, 2013; Cropanzano, Bowen, & Gilliland, 2007; Gilliland, 1993; Greenberg, 1987, 1990). All these dimensions consist of several constructs to be operationalized, which are presented in below figure (4-1):



Components of Organizational Justice

1. Distributive Justice: Appropriateness of outcomes.
● Equity: Rewarding employees based on their contributions.
● Equality: Providing each employee roughly the same compensation.
● Need: Providing a benefit based on one's personal requirements.
2. Procedural Justice: Appropriateness of the allocation process.
● Consistency: All employees are treated the same.
● Lack of Bias: No person or group is singled out for discrimination or ill-treatment.
● Accuracy: Decisions are based on accurate information.
● Representation of All Concerned: Appropriate stakeholders have input into a decision.
● Correction: There is an appeals process or other mechanism for fixing mistakes.
● Ethics: Norms of professional conduct are not violated.
3. Interactional Justice: Appropriateness of the treatment one receives from authority figures.
● Interpersonal Justice: Treating an employee with dignity, courtesy, and respect.
● Informational Justice: Sharing relevant information with employees.

Figure 4-1: Components of the Organizational Justice Theory, bron: Cropanzano et al. (2007)

[11] The terms fairness and justice are used interchangeably here.
[12] The organizational justice theory is therefore suitable, but not limited to a hiring-context

The addressment of bias is only one of the considerations on fairness, forming a separate component (called: "Lack of Bias") within this theory, mapped in the dimension of procedural fairness. The bias-taxonomy is therefore related, and connected to this component and by that positioned within the wider frame of other fairness dimensions, which are also recognized to be important. The four dimensions mentioned are grouped within the artefact as one (card-)quartet, named: "Fairness-by-design", and each dimension is shaped as a card with the above components, where informational fairness is divided in the well-defined, components: *adequacy*, *timeliness*, *truthfulness*, and *specificity*, upon additional literature of Colquitt (2001).

4.1.2. Embedding Values in AI-design

The other refinement which we consider necessary is the mechanism of embedding fairness, as fairness is recognized to be a multi-dimensional construct (instead of only referring to unbiased AI-solutions). Therefore, a small literature search is conducted on methodologies of embedding values in technology design, whereby ADM's in particular.

So, next to investigating localized fairness in a hiring context, also a quick and dirty, additional literature research is done on mechanisms for embedding ethical values in technology design. When searching for this, *the Handbook of Ethics, Values, and Technological Design* (Van den Hoven, Vermaas, & Van de Poel, 2015) is found to be a bundled, and therefore handy reference book. Five different methodologies for embedding values (e.g., Values at Play, and Participatory Design) are introduced in the handbook, which all take context, and stakeholders in account, when searching for answers around operationalizing ethical values in a solution. Of all these, Value Sensitive Design (VSD) is posited as the most extensive, pioneering endeavor till date, in proactive addressment of human values throughout technology design (Van den Hoven et al., 2015). Therefore this method is chosen here for escorting the embedding of the ethical value of fairness in (AI-)design .

Poel (2013); Van den Hoven et al. (2015) introduce the concept "Value Hierarchy", for translating values into norms, and ultimately design requirements, which is a crucial step in VSD. By reconstructing a value hierarchy, the translation of values into requirements is systemized, making the value judgments involved explicit, and therefore more transparent, and debatable among the parties involved:

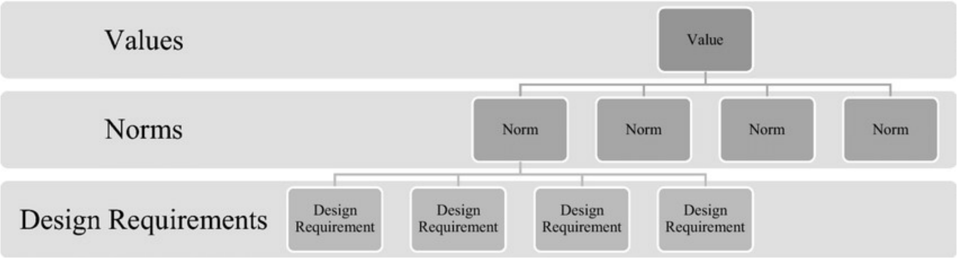
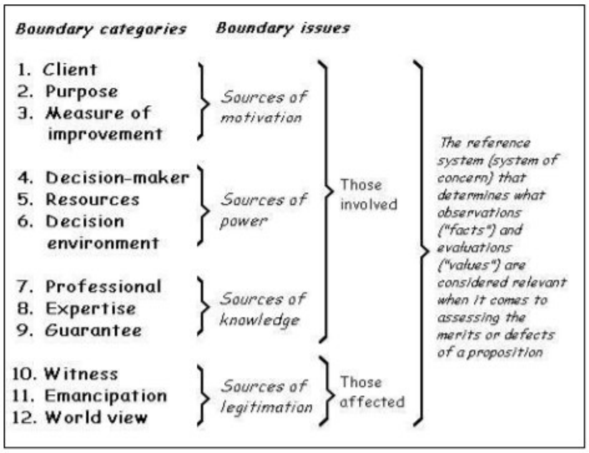


Figure 4-2: Poel, I. V. D. (2013). Translating values into design requirements. In Philosophy and engineering: Reflections on practice, principles and process (pp. 253-266). Springer, Dordrecht.

By using these hierarchies during (ADM-)design, conflicting values (a.k.a. value tensions) are surfaced, so that an explicit prioritization in norms is enforced. The frameworks' intention is not to provide the right answers, but to guide conscious choices, and making them explicit. Likewise the conceptual framework, the Value Hierarchy embraces a top-down approach, as the higher level value is translated ultimately in non-functional design requirements with regard to the solution. The earlier identified fairness dimensions of the OJT (in §4.1.1) can act as illustrative (domain-specific) norms in this systemization. Of course these substantive fairness-norms can be expanded in future iterations, when the artefact matures.

Critical System Heuristics (CSH)

When scanning earlier handbooks' Chapter (Van den Hoven et al., 2015) on this VSD-methodology, it is found that Yetim (2011) problematizes that it is infeasible to include all stakeholders in discourse about values, and that interpretations of values and tools may change over time. For advancing the application of VSD, Yetim (2011) recommends the use of Ulrich and Reynolds (2010) "Critical Heuristic boundary questions" in identifying stakeholders, as a heuristic is a pragmatic approach i.e., mental shortcut, speeding up the process of finding a satisfactory solution (Wikipedia, 2022a). In the artefact, it is searched for a practical form so this shortcut is very welcomed. The questions, which concern 4 boundary categories (namely: sources of motivation, power, knowledge, and legitimation), intend to reveal unresolved boundary issues. Therefore these questions are ideally addressed in the descriptive " is-" and prescriptive "ought-"mode (please find figure 4-3 below):



Boundary categories	Boundary issues		
1. Client 2. Purpose 3. Measure of improvement	Sources of motivation	Those involved	The reference system (system of concern) that determines what observations ("facts") and evaluations ("values") are considered relevant when it comes to assessing the merits or defects of a proposition
4. Decision-maker 5. Resources 6. Decision environment	Sources of power		
7. Professional 8. Expertise 9. Guarantee	Sources of knowledge		
10. Witness 11. Emancipation 12. World view	Sources of legitimation	Those affected	

Figure 4-3: Table of Boundary Categories, source: W. Ulrich 1983, p.258; 1996, p.43; and 2000, p.256

Here we introduce CSH as a practical (both efficient and quick), communicative, and emancipatory instrument for facilitating a democratic and transparent design process, which we pose as: "Fairness-in-Design". The 4 boundary categories is therefore formed a separate quartet in the artefact, grouped by the name "Context Analysis". This seems to be appropriate as a naming, as boundaries around goals of the ADM, stakeholder(concern)s to be involved, power of the decision-makers, and so on, are considered when designing a system. More background information on CSH and the substantive, corresponding boundary questions are found in Appendix 16.

4.1.3. Findings HR-expert descriptive case-study

To gain more insights on the HR-selection process, becoming more familiar with the HR-domain as a researcher, and to increase the reality level of the Princeton use-case (fictional case for evaluating the artefact, please find §5.3), a qualitative, sideways research on AI in hiring (wherein resume screening in particular), is conducted. Whereas earlier mentioned OJT is mainly highlighting fairness from the viewpoint of perceived fairness of applicants, here the perception of HR-recruiters is centralized. Since this research is not the main research, the research design of this multiple-case study is included in Appendix 20.

The questions asked in interviewing 4 HR-experts follow some of the major themes of the theoretical framework in light of the particular HR-domain. These are amongst others: potential biases, and fairness, next to findings related to the illustrative use-case of automated resume screening, where opportunities, risks, and appropriateness of resume screening in hiring are explored from the viewpoint of 4 experienced HR-managers. The goal of this oriental case-study is to see if there are any important constructs which should be introduced on top of the construct in the framework based upon literature findings.

In this section the results of this case-study are presented. Since the interviews generate quite a lot of data, findings are coded, and presented in a more concise, transparent way. The coding's, and complete interviews, are filed in a separate, confidential Appendix.

Examining the solution space of an AI-solution

In the chosen use-case[13] for evaluation purposes, the solution is already given: a resume screening algorithm decides upon incoming resumes if a candidate is hired or not. However, some of the first important design choices, occur in the so called problem-formulation phase a.k.a. the conceptional (design)stage. Contrary to research-findings that resume screening algorithms were designed in a fair manner offering a candidate a "hire" or "no-hire " in the HR process (e.g.; *AI in talent acquisition: a review of AI-applications used in recruitment and selection* (Albert, 2019)), the general tendency arising from the interviews is that the HR managers indicate that a resume is incomprehensive in the assessment of job-suitability of an applicant. Therefore, it can be strongly disputed if would be a fair practice (even when the output would be fair in terms of output, i.e., fairness metrics). When considering the role of a resume in the traditional hiring process, it is said that:

1) a resume is considered only the half part of the selection process and tells something about the strict demands, which is only the first filtering;
2) although resumes are used in traditional application processes the predictive ability is limited towards job-suitability. Partly because resumes are results from the past, candidates put on paper about themselves, which are not necessarily a guarantee for the future;
3) in the domain of technicians, for example, mostly practitioner-types, there are good candidates who not always have the competences to write a decent resume;
4) soft criteria (skills, personal characteristics, competencies), are often equally important, or in some cases even more important, and which in most cases cannot be derived from a resume;
5) a personal conversation is very relevant, to get more in-depth information which is not directly revealed in a resume itself. Therefore considering a Human-in-the-Loop, would be advisable.

[13] Shortened version of the Princeton use-case: Hiring-by-Machine (https://aiethics.princeton.edu/wp-content/uploads/sites/587/2018/12/Princeton-AI-Ethics-Case-Study-5.pdf).

When asking the HR-managers' expert-opinion on the suitability of using an automated resume screening algorithm, the opinions differ mead in regard to the objective. As a prefilter maybe, but to determine an affirmative hire seems to be inappropriate, the HR managers mutually agree, here presenting one of the responses as an example:

> *"[..] if you have people come up with their own CV and have all kinds of smart scans come over it, with all kinds of algorithms, and use that to deduce something about suitability, no luckily I don't see that happening any time soon." SN*

In the light of fairness, mostly referred to, by the HR-experts, is: non-discrimination, inclusiveness, and diversity. However, it is underlined that diversity, and inclusiveness is created in the mix of attributes, and that groupings like "men", "Dutch people" are not monoliths, but creates stereotyping.

Influence on the artefact

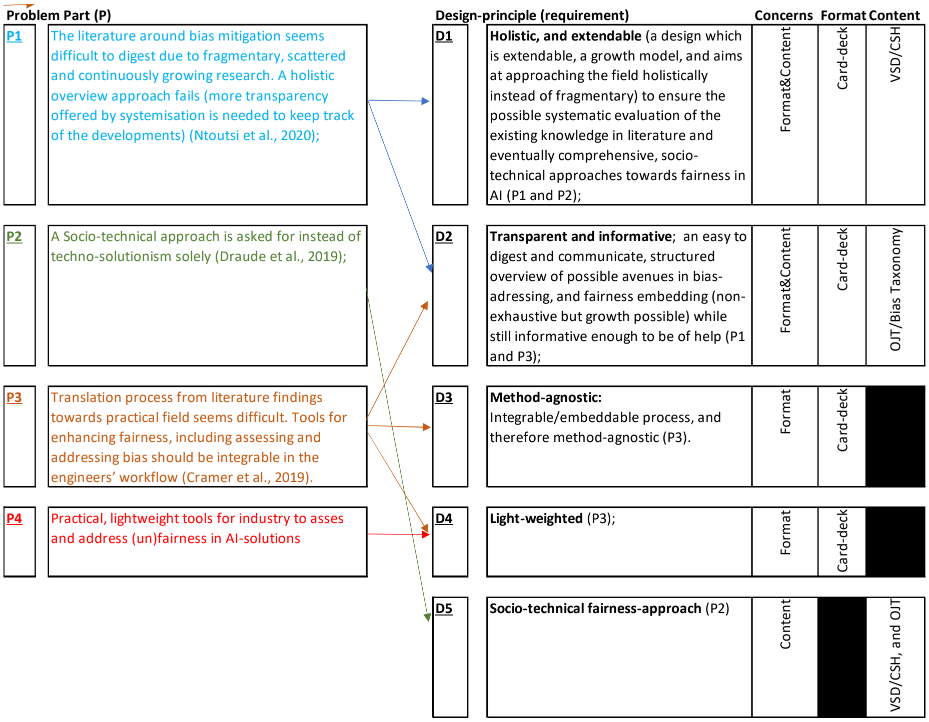
When interviewing the recruiters, it appears that a sociotechnical approach towards fairness is justified within the Hiring domain, and more specifically in designing a resume screening algorithm. Considering the "solution space" is found to be of importance, when designing, and provides an additional add-on, integrated within the artefact, as we will see in the next Chapter.

4.2. Responsible AI:CEID Toolkit

As the goal of design-science is utility and effectiveness next to truth (Hevner et al., 2004), we define here the requirements, so called design principles[14] (D1 until D5), and properties which are derived from the mentioned atomized (problem)parts (P1 until P4 in Chapter 1.2) and should be embedded in the artefact:

- D1: Holistic and extendable;
- D2: Transparent and informative;
- D3: Method-agnostic;
- D4: Light-weighted;
- D5: Socio-technical in approach.

In figure 4-4 the match between problem parts en design principles, i.e. requirements is visualised, and explained:



Problem Part (P)

P1	The literature around bias mitigation seems difficult to digest due to fragmentary, scattered and continuously growing research. A holistic overview approach fails (more transparency offered by systemisation is needed to keep track of the developments) (Ntoutsi et al., 2020);
P2	A Socio-technical approach is asked for instead of techno-solutionism solely (Draude et al., 2019);
P3	Translation process from literature findings towards practical field seems difficult. Tools for enhancing fairness, including assessing and addressing bias should be integrable in the engineers' workflow (Cramer et al., 2019).
P4	Practical, lightweight tools for industry to asses and address (un)fairness in AI-solutions

	Design-principle (requirement)	Concerns	Format	Content
D1	**Holistic, and extendable** (a design which is extendable, a growth model, and aims at approaching the field holistically instead of fragmentary) to ensure the possible systematic evaluation of the existing knowledge in literature and eventually comprehensive, socio-technical approaches towards fairness in AI (P1 and P2);	Format&Content	Card-deck	VSD/CSH
D2	**Transparent and informative**; an easy to digest and communicate, structured overview of possible avenues in bias-adressing, and fairness embedding (non-exhaustive but growth possible) while still informative enough to be of help (P1 and P3);	Format&Content	Card-deck	OJT/Bias Taxonomy
D3	**Method-agnostic:** Integrable/embeddable process, and therefore method-agnostic (P3).	Format	Card-deck	■
D4	**Light-weighted** (P3);	Format	Card-deck	■
D5	**Socio-technical fairness-approach** (P2)	Content	■	VSD/CSH, and OJT

Figure 4-4: Problem Parts translated in Design Requirements

4.2.1. Format/Shape

For structuring, and systemization purposes the artefact is shaped as a physical card deck, in which content is grouped and instantiated outside-in along the line of the holistic approach based upon what is gathered and learned in our SLR. This format satisfies the design principles. Literature confirms, that card-based tools aid in the design process and provide information, methods, or good practices in a handy form (D. Urquhart & J. Craigon, 2021; Roy & Warren, 2019).

[14] Design principles provide a comprehensive view of how to construct the artifact to deliver its objectives. They establish the criteria for building the artifact and setting the requirements and expectations of stakeholders regarding its use and maintenance.

Card decks are assigned to have many strong properties as a design tool, as they:

- facilitate creative combinations of information and ideas;
- provide common basis for understanding and communication in a team;
- provide tangible external representations of design elements or information;
- provide convenient summaries of useful information and/or methods;
- are semi-structured tools between blank Post-it notes and detailed instruction manuals.

4.2.2. Realization of the Design

The card deck, so called design-quartet for Responsible AI (please find Appendix 11 for an example-set) is introduced here, as a tentative design, and is considered to be an easy to use, tangible, method-agnostic way of presenting, and offering the knowledge of the theoretical framework to be used as a tool, meeting the requirements, i.e. design principles: D1 until D4. Design-principle 5 (D5) is addressed by using VSD, CSH, and OJT as guiding content.

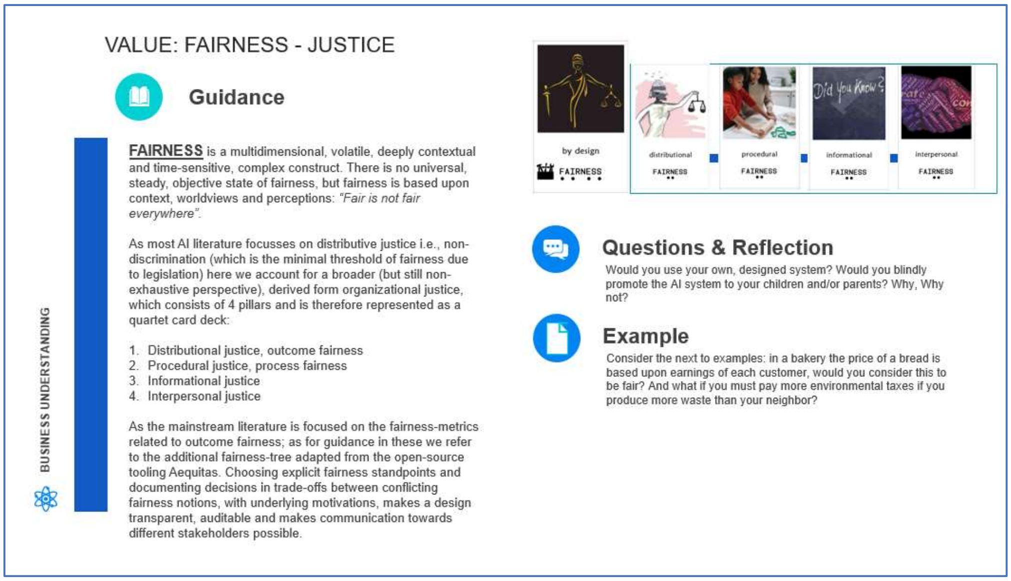
These are physical cards, in which important self-reflective questions, and guidance in the shape of hints are processed. The cards are derivatives of the earlier presented theoretical fair-ML framework. Guidance and questions are structured, following a top-down approach, organized in different main-topics (as included in the theoretical framework), being, amongst others: Fairness, Context Analysis in which Stakeholders and their concerns are explored, and Pitfalls. Also, different stages of CRISP-DM are referred to on these cards, for companies using this or similar processes for Data Mining.

As for this study different design quartets are crafted, being:

- Quartet 1: “Fairness-by-design” in which embedding fairness is centralized, using the four dimensions of the OJT (fairness-dimensions in hiring) and where systemization of the Value Hierarchy, as a design pattern, should be followed to operationalize fairness-norms;
- Quartet 2: “Context-analysis”, in which context is analysed and embedding fairness within the process of AI-design is facilitated by following the methodology of CSH, as it intends to give Stakeholders affected, and involved a voice in the process, by including their concerns, referring to “Fairness-in-Design”;
- Quartet 3: “Common Pitfalls”: a quartet with 4 traps grafted on ADM-design. These pitfalls are founded upon the traps mentioned by Selbst et al. (2019) (please find Appendix 10 for details) being:
    - Solutionism Trap (which is expanded with the additional concept of solution space, initiated by the HR-expert case-study);
    - Portability Trap;
    - Formalism Trap, and;
    - Ripple-effect Trap.

All these quartets have in common, that they are intended to be used from the very beginning: in the inception, i.e. conceptual stage of ADM-design in which ideas on ADM-solutions are explored, and initiated. Additionally a Quartet called “Data” is crafted, however not used for evaluation within this study. This Quartet introduces self-reflective question on Data Quality (timeliness, recency, sufficiency and accuracy), with the aim of reducing bias in the design of ADM’s. As this is considered to be a next step in ADM-design-process this is out of scope for testing. The “Data”-quartet is intended to be used in a future evaluative iteration together with the constructed bias-taxonomy (Appendix 4).

As for example-purposes, here below a full quartet (called Fairness-by-Design), and thereafter one of the included physical cards is presented. This fairness-quartet is based upon earlier findings of the OJT, providing different norms where designers can reflect upon, formulating design requirements towards the ADM to be designed. One card is the main-card on which guidance, examples for illustration, and self-reflective questions are presented, introducing the four cards which form a quartet. Also the CRISP-DM stage which best connects to the issues addressed is implemented on this main-card, called "Fairness-by-Design":


by design
FAIRNESS
distributional
FAIRNESS
procedural
FAIRNESS
informational
FAIRNESS
interpersonal
FAIRNESS
Did you Know?
BUSINESS UNDERSTANDING

VALUE: FAIRNESS - JUSTICE

Guidance

FAIRNESS is a multidimensional, volatile, deeply contextual and time-sensitive, complex construct. There is no universal, steady, objective state of fairness, but fairness is based upon context, worldviews and perceptions: *"Fair is not fair everywhere"*.

As most AI literature focusses on distributive justice i.e., non-discrimination (which is the minimal threshold of fairness due to legislation) here we account for a broader (but still non-exhaustive perspective), derived form organizational justice, which consists of 4 pillars and is therefore represented as a quartet card deck:

1. Distributional justice, outcome fairness
2. Procedural justice, process fairness
3. Informational justice
4. Interpersonal justice

As the mainstream literature is focused on the fairness-metrics related to outcome fairness; as for guidance in these we refer to the additional fairness-tree adapted from the open-source tooling Aequitas. Choosing explicit fairness standpoints and documenting decisions in trade-offs between conflicting fairness notions, with underlying motivations, makes a design transparent, auditable and makes communication towards different stakeholders possible.

Questions & Reflection

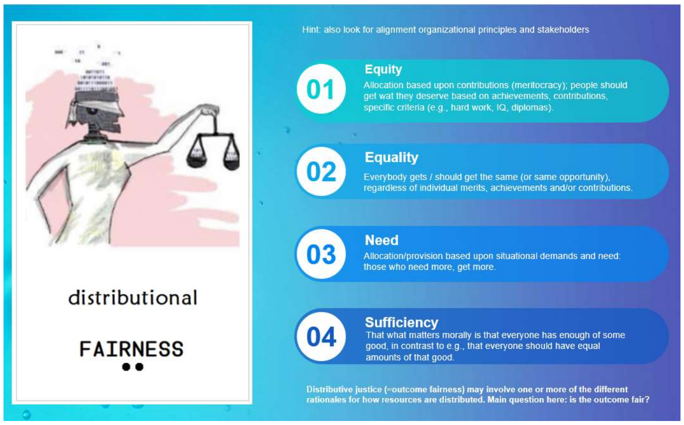
Would you use your own, designed system? Would you blindly promote the AI system to your children and/or parents? Why, Why not?

Example

Consider the next to examples: in a bakery the price of a bread is based upon earnings of each customer, would you consider this to be fair? And what if you must pay more environmental taxes if you produce more waste than your neighbor?

Figure 4-5: Example of Fairness Main-card for instructive purposes

The four cards, corresponding to the 4 dimensions of organisational justice are the different cards on which different constructs related to these dimensions are presented. Here, as an example, one of the four cards of the quartet is presented:


distributional
FAIRNESS

Hint: also look for alignment organizational principles and stakeholders

01 Equity
Allocation based upon contributions (meritocracy); people should get wat they deserve based on achievements, contributions, specific criteria (e.g., hard work, IQ, diplomas).

02 Equality
Everybody gets / should get the same (or same opportunity), regardless of individual merits, achievements and/or contributions.

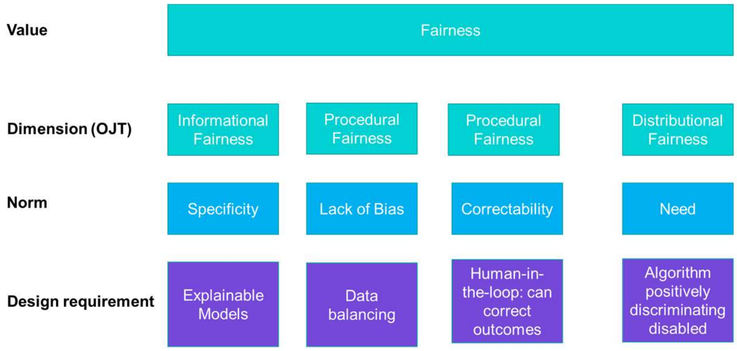
03 Need
Allocation/provision based upon situational demands and need: those who need more, get more.

04 Sufficiency
That what matters morally is that everyone has enough of some good, in contrast to e.g., that everyone should have equal amounts of that good.

Distributive justice (=outcome fairness) may involve one or more of the different rationales for how resources are distributed. Main question here: is the outcome fair?

Figure 4-6: Example of a card that consists of the Fairness-by-Design quartet

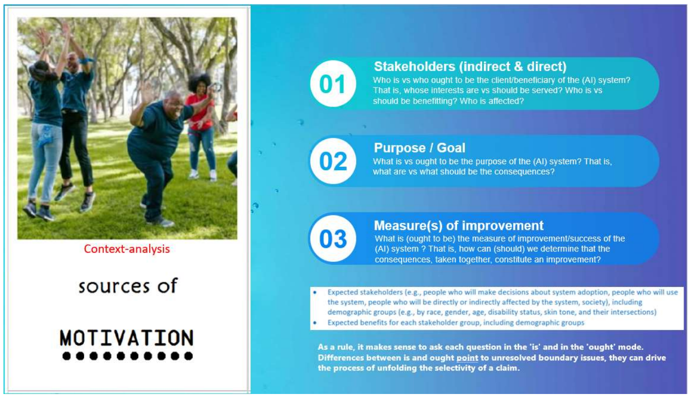
When using these cards, the Value Hierarchy should be followed, to formulate the design requirements. In figure 4-7 an example is provided of how it intends to work:



Value	Fairness			
Dimension (OJT)	Informational Fairness	Procedural Fairness	Procedural Fairness	Distributional Fairness
Norm	Specificity	Lack of Bias	Correctability	Need
Design requirement	Explainable Models	Data balancing	Human-in-the-loop: can correct outcomes	Algorithm positively discriminating disabled

Figure 4-7: Illustration of using the Fairness-by-Design Quartet following the Value Hierarchy systemization.

Likewise a (bundled context exploration-)quartet is made as for the earlier described, 4 boundary categories of CSH, where each card represents one category representing the category-related, self-reflective boundary-questions for fairness-in-design (please find Appendix 16 for details on these):



Context-analysis

sources of

MOTIVATION

●●●●●●●●●●

01 **Stakeholders (indirect & direct)**
Who is vs who ought to be the client/beneficiary of the (AI) system? That is, whose interests are vs should be served? Who is vs should be benefitting? Who is affected?

02 **Purpose / Goal**
What is vs ought to be the purpose of the (AI) system? That is, what are vs what should be the consequences?

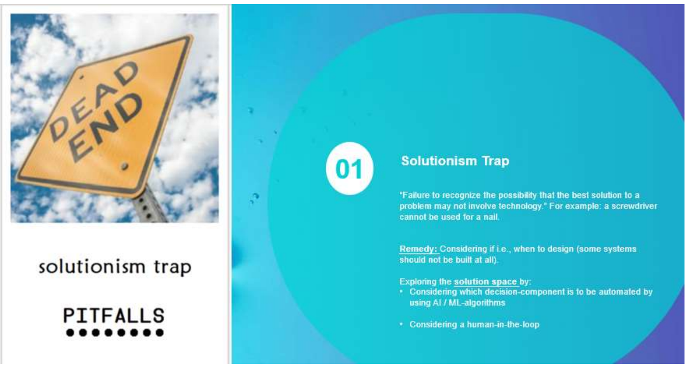
03 **Measure(s) of improvement**
What is (ought to be) the measure of improvement/success of the (AI) system ? That is, how can (should) we determine that the consequences, taken together, constitute an improvement?

- Expected stakeholders (e.g., people who will make decisions about system adoption, people who will use the system, people who will be directly or indirectly affected by the system, society), including demographic groups (e.g., by race, gender, age, disability status, skin tone, and their intersections)
- Expected benefits for each stakeholder group, including demographic groups

As a rule, it makes sense to ask each question in the 'is' and in the 'ought' mode. Differences between is and ought point to unresolved boundary issues, they can drive the process of unfolding the selectivity of a claim.

Figure 4-8: Example card of the Context Analysis quartet

And, finally here an example of a card in the "common pitfalls"-quartet, which relates to the blind spot, to not take in consideration the option of not building a solution, when searching for fair-AI solutions. As mentioned additional insights of the HR-expert-study are incorporated in this card-deck.



solutionism trap

PITFALLS

●●●●●●●●

01 **Solutionism Trap**

"Failure to recognize the possibility that the best solution to a problem may not involve technology." For example: a screwdriver cannot be used for a nail.

Remedy: Considering if i.e., when to design (some systems should not be built at all).

Exploring the solution space by:

- Considering which decision-component is to be automated by using AI / ML-algorithms
- Considering a human-in-the-loop

Figure 4-9: Example card of the Pitfalls Quartet

In the first, conceptual stage of building an ADM, corresponding to the Business Understanding Phase of Crisp-DM, all 3 design quartets are considered to be important.

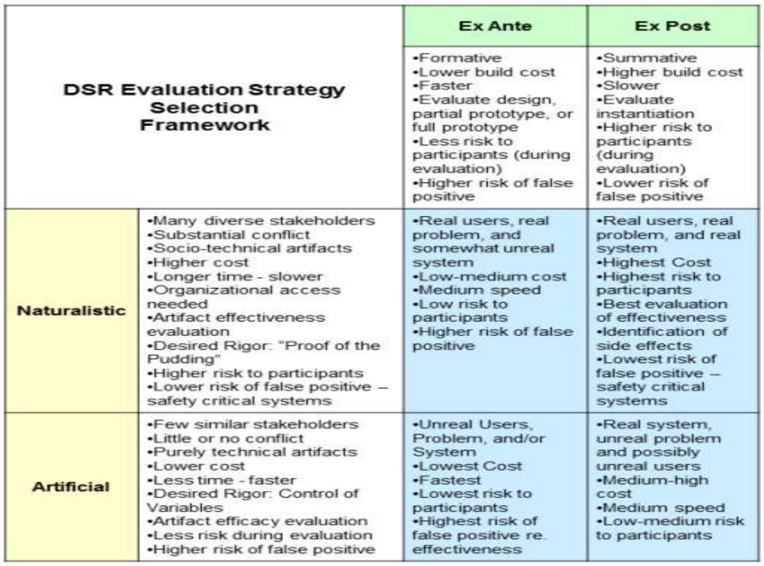
5. Artefact Demonstration

Referring to the overall DSRM-cycle (please find Figure 1-1, Chapter 2), demonstration is part of the evaluation-cycle and consists of two major parts: (1) finding a suitable context, and (2) using the artefact to solve a problem. As Pries-Heje, Baskerville, and Venable (2008); Venable, Pries-Heje, and Baskerville (2012) underline, it is preferred to evaluate a sociotechnical artefact, ex-post in a naturalistic setting . Hereby referring to a real-life environment, solving a real problem, as this is the desired rigor: “the real proof of the pudding”. Therefore, a suitable context would be an organization where a high-impact ADM is being developed in which the value fairness should be embedded (=the problem).

5.1. Defining the demonstration strategy

Although earlier mentioned preferences of using a naturalistic environment (for testing), it appeared impracticable to gain access to an organizational setting with access to a real-life-situation where designing an ADM for fairness is applicable. Due to limited time, available resources, cost constraints, and (data-)privacy reasons, an alternative DSR-evaluation-strategy is followed. Known validity limitations of having a higher risk of false positives, and negatives within this chosen strategy are taken for granted.

Here an artificial, ex-ante strategy (Venable et al., 2012) is identified to be the most feasible evaluation approach within this study (please find figure 5-1; in which the rows represent the paradigm, and the columns represent the functional purpose of the evaluation). This means testing a prototype (the artefacts’ 0.1 version) on perceived usefulness in a formative way, before the artefacts’ 1.0 version is constructed.



DSR Evaluation Strategy Selection Framework		Ex Ante	Ex Post
		•Formative •Lower build cost •Faster •Evaluate design, partial prototype, or full prototype •Less risk to participants (during evaluation) •Higher risk of false positive	•Summative •Higher build cost •Slower •Evaluate instantiation •Higher risk to participants (during evaluation) •Lower risk of false positive
Naturalistic	•Many diverse stakeholders •Substantial conflict •Socio-technical artifacts •Higher cost •Longer time - slower •Organizational access needed •Artifact effectiveness evaluation •Desired Rigor: "Proof of the Pudding" •Higher risk to participants •Lower risk of false positive – safety critical systems	•Real users, real problem, and somewhat unreal system •Low-medium cost •Medium speed •Low risk to participants •Higher risk of false positive	•Real users, real problem, and real system •Highest Cost •Highest risk to participants •Best evaluation of effectiveness •Identification of side effects •Lowest risk of false positive – safety critical systems
Artificial	•Few similar stakeholders •Little or no conflict •Purely technical artifacts •Lower cost •Less time - faster •Desired Rigor: Control of Variables •Artifact efficacy evaluation •Less risk during evaluation •Higher risk of false positive	•Unreal Users, Problem, and/or System •Lowest Cost •Fastest •Lowest risk to participants •Highest risk of false positive re. effectiveness	•Real system, unreal problem and possibly unreal users •Medium-high cost •Medium speed •Low-medium risk to participants

Figure 5-1: A DSR Evaluation Strategy Selection Framework. (Source: Venable et al, 2012)

Taken concretely, as for the first iteration in the DSR-cycle a fictional use case (§5.3) will be used for the artefacts’ demonstration purpose, by conducting a workshop (§5.2) within an existing (AI-expert-)organization (§5.4). This is the scenario of a real business setting, addressing a simulated context.

In the next sections, respectively, (1) the workshop set-up, (2) artificial use-case (used for demonstration purposes), and (3) case-organization where demonstration takes place will be outlined.

5.2. Workshop setup: "Responsible AI"

Given the size in terms of effort and time, a first iteration, demonstrating the artefact is conducted within a single case organization (§5.4) only, although this choice has obviously limits on the generalization of the research findings (please find §5.5.1.).

For the workshop a multidisciplinary group is used, as a multi-disciplinary approach is promoted by the card-deck.[15] This group acts as an Exploratory Focus Group (EFG), used to propose subjective improvements in a first iteration of the "build-and-evaluate" cycle of artefact-design. A focus group brings together a small group of people to answer questions in a moderated setting, in this particular research shedding light on the artificial use-case on resume screening (as described in §5.3). EFG's can provide more nuanced and natural feedback than individual interviews and are easier to organize than experiments or large-scale surveys (Tremblay, Hevner, & Berndt, 2010).

Some preparations were done in order to increase the chance of a successful workshop-session. Before executing the "real" workshop a pilot is held with two professors of the Data Science Management Course for dry swimming purpose, and to reveal suggestions for change and improvement. The pilot resulted in various improvement points which were taken to heart, and implemented, for the final workshop.

Suggested improvement points were to reduce the amount of material within the workshop, using more laymen's' terminology, and providing a better explanation what the exact artefact is about. Next to that, is was advised to do less talking, putting the participants more to work, and preparing a script with a time schedule and escalation plan. In terms of content, the pilot-participants seemed to need more guidance in using the cards, in terms of a roadmap or "how-to"-manual.
As for the PowerPoint intro it was advised to include a shorter, concise version of the pitch as for repetition purposes, preventing information overload. All the above suggested improvements, and additions were taken to heart, and implemented, for the final workshop.

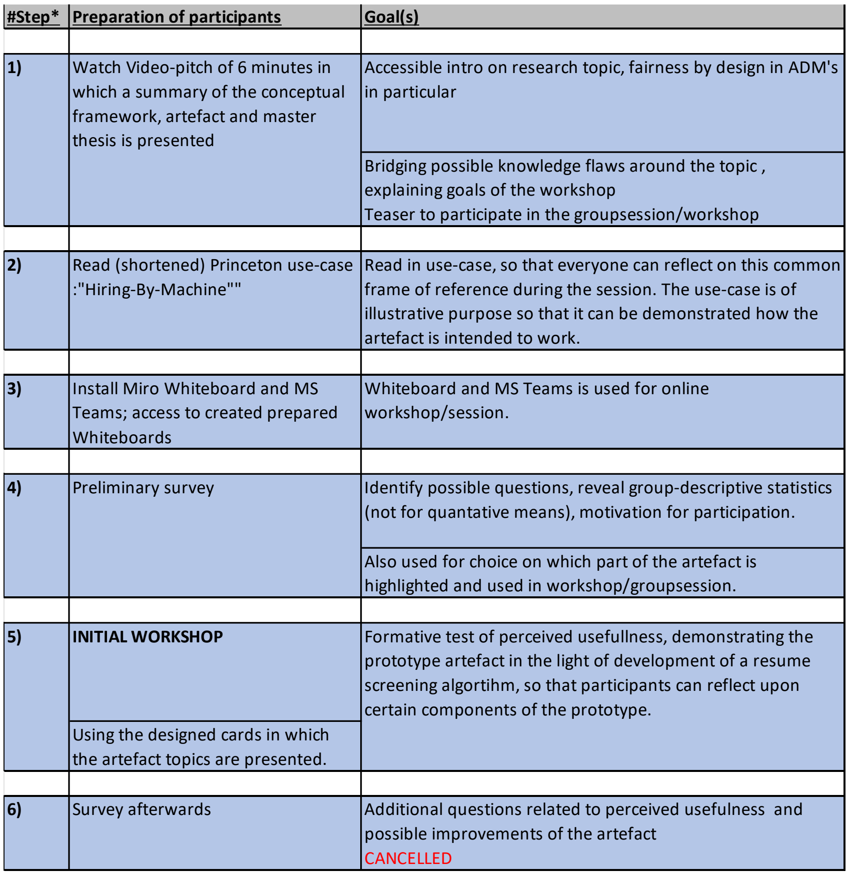
To summarize the set-up of the workshop, these are the different steps, and measures taken (in chronological order) in preparation of the workshop:

- A video pitch of approximately 6 minutes, is being made for informational, introductive and teaser purposes;
- The Princeton use-case is translated in Dutch, and reduced in size, focusing on the essential points needed for the purpose of testing the artefact;
- A survey is set-up to question the participants before the workshop on self-reported experience, interests, unclarities of the video-pitch and/or research topic;
- A PowerPoint presentation is made as for intro and clarification purposes, during the actual workshop, including a shorter, 3 minutes repetitive version, of the pitch;
- For systemization, providing more guidance in the use of cards, and to escort the workshop 5 online white-boards are prepared;
- A script with the content to be handled, including time-table, and escalation plan is made;
- A survey is set-up to question the participants after the workshop mainly for evaluation purposes of the artefact.

For more details on these different steps we refer to Appendix 12.

[15] Based upon literature findings that one of the important remedies against bias and unfairness, is that diverse, multi-disciplinary teams are being used when designing ADM's.

A week, prior to conducting the online workshop, all participants are requested to prepare for participation by following the “game plan” provided by e-mail (please find Appendix 17). Here the different steps all participants are encouraged to take before attending the workshop are visually presented in chronological order:



#Step*	Preparation of participants	Goal(s)
1)	Watch Video-pitch of 6 minutes in which a summary of the conceptual framework, artefact and master thesis is presented	Accessible intro on research topic, fairness by design in ADM's in particular
		Bridging possible knowledge flaws around the topic , explaining goals of the workshop Teaser to participate in the groupsession/workshop
2)	Read (shortened) Princeton use-case :"Hiring-By-Machine""	Read in use-case, so that everyone can reflect on this common frame of reference during the session. The use-case is of illustrative purpose so that it can be demonstrated how the artefact is intended to work.
3)	Install Miro Whiteboard and MS Teams; access to created prepared Whiteboards	Whiteboard and MS Teams is used for online workshop/session.
4)	Preliminary survey	Identify possible questions, reveal group-descriptive statistics (not for quantative means), motivation for participation.
		Also used for choice on which part of the artefact is highlighted and used in workshop/groupsession.
5)	**INITIAL WORKSHOP**	Formative test of perceived usefullness, demonstrating the prototype artefact in the light of development of a resume screening algortihm, so that participants can reflect upon certain components of the prototype.
	Using the designed cards in which the artefact topics are presented.	
6)	Survey afterwards	Additional questions related to perceived usefulness and possible improvements of the artefact CANCELLED

(chronological order)

Table 5-1: Workshop instructions for participants

The first part of the workshop (approximately 20 minutes) is planned for introduction purposes by using a PowerPoint presentation, repeating a shorter version of the pitch, question-handling, and clarifying the goals in terms of background information, intentions, expectations, and output (please find Appendix 15). Then one hour and 15 minutes is reserved for the activities, using the cards, when addressing the use-case, and illustrations. The cards should provide some guidance for the use-case. The users, i.e. participants can experience the added value, in solving a problem in a relevant professional context using the prototype of the artifact.

The workshop will be concluded with a 15-20 minute-part for evaluation purposes. As for the evaluation a whiteboard is prepared, using the retrospective 4L’s (Liked, Learned, Longed for, and Lacked) in which participants can reflect upon the perceived usefulness of the prototype, expressing their good, and bad experiences on a first acquaintance with the artefact.

5.3. Princeton Use-case: “Hiring-by-Machine”

An openly, online-accessible, artificial HR use-case on automated resume screening using NLP called “Hiring-by-Machine” (please find Appendix 18) is found to act as a suitable case, to reflect upon during a group session with (AI-)experts in an existing AI-developers organization. This case is set up by a multidisciplinary team at Princeton, and although fictitious in design, it has an empirical basis. It represents a rich, vibrant, in-depth described example of an existing, real-life AI-application, with identical accompanied issues in both design and implementation.

The case is designed interactively, with the aim of stimulating group discussion. Furthermore, the case is widely accessible in both, academic and professional disciplines (Ethics, Computer Science, Philosophy, Social Sciences, Laws and Regulations). Above facts given, and the fact that the particular themes of fairness, diversity and contextual integrity in an ADM-setting are included in this case, it is considered appropriate for testing-purposes of the artifact. There are multiple stakeholders involved in the offered material, with different concerns and where fairness is a centralized challenge which the developers in this case have to deal with.

During a (focus)group session i.e. workshop the prototype of the artefact is demonstrated on this use case, so all workshop-participants have the same frame of reference to evaluate the prototypes’ problem-solving ability, and in experiencing its intended use.

The original Princeton use-case is shortened and translated in Dutch by the researcher, to reduce the barrier for participation, as time and effort for participants to prepare for the workshop is shortened. Also this measure is taken to increase understandability (by trying to avoid misunderstandings due to English-deficiencies, and by omitting unnecessary parts so that participants only have to focus on the essential parts).

During the workshop this use-case will be used as context, whereby it is required that the participants look at the usefulness of the artefact in relation to this fictional case from both the perspective of one’s own expertise, and from a practical point of view.

5.4. Background info: Case Organization

The developed framework will be demonstrated, in an expert organisation which develops ML and data-driven (business)decision-making solutions as one of their core activities. The facilitator chosen for the iterative development and testing of the artefact is an SME located in the South of the Netherlands (Limburg), to be characterized as an independent, international IT-company. This IT-organization focuses mainly on delivering consultancy services, and has a multidisciplinary team on board with digital, business, and organizational expertise. The team consists of approximately 75 experts in the field of amongst others, Software Engineering and Data Science. The services they deliver are correspondingly their expertise around: AI & Machine Learning, Artificial Intelligence for IT Operations (AIOPS), Software Engineering, Big Data Warehousing, and Business Intelligence.

Since this research is focused on the development of ML Algorithms in the context of decision-making, this organization is considered to be an appropriate case organisation for this research, as (1) designing AI-solutions is one of its repetitive (daily-based) core activities and therefore (2) this organisation has specialists on board, experienced enough for leveraging feedback on the developed artefact. Notable as well is that customers are diverse, in the sense that they are operating in several domains, within various industries, and are sized ranging from SME's to large multinationals.

The roles which are considered to be relevant for this workshop are presented in the table below, including the self-reported, surveyed, main-incentive(s) to participate:

Abbrevation	Role / Function	Special Interest / Motive to participate
MDI	Manager Data & Intelligence	(1) The organization intends to focus more on this in the future (2) Bias addressing
HRE	HR-Employee	AI applications within HR
EA	Enterprise Architect (and co-owner)	Resonsible AI, fairness in AI
SDS	Senior Data-Scientist	Bias adressing

As for the workshop 4 experts are included in the group session, upholding a diverse, multidisciplinary group (in consultation with the company supervisor):

1. FS – Senior Data Scientist (SDS);
2. TP - HR Expert (HRE), and specialist in labour-law;
3. MK – Manager Data & Intelligence (MDI) with education in Econometry;
4. JS – Enterprise IT-Architect (EA), and managing partner (co-owner).

1. The role of a SDS is seen relevant, as senior expertise in AI/ML, and bias-addressing on an operational level is wanted and covered by including this role;
2. A multi-disciplinary approach in the artefact is promoted, by including domain-expertise in the design. The role of the HRE is chosen as for specific domain-expertise within HR, as this is the particular domain of the use-case;
3. The MDI is chosen in a double role; this role has both Data Science Management (tactical, and management) expertise, as well as operational experience, and knowledge (daily design choices in algorithmic design);
4. The EA, also Co-owner and managing partner also has a double role. In the role of EA this is a key-informant on the topic of designing a sociotechnical solution using a helicopter view, by incorporating an organization's strategy, processes, and IT assets, pursuing business alignment by including relevant stakeholders. From the perspective and role of a managing partner, expertise in strategic, and tactical decision-making, probably risk-management, as well as organizational policymaking is potentially covered.

5.5. Reflection on validity, reliability and ethical aspects

This section is devoted to the aspects of validity, reliability and ethics that relate to the Workshop, which are discussed in respectively §5.5.1, §5.5.2, and §5.5.3.

5.5.1. Validity

In guaranteeing validity of research findings, Saunders, Lewis, and Thornhill (2009) recognize three types of validity: internal, external and construct validity.

Internal validity

Internal validity is the degree to which one can state with certainty that findings of an established causal relationship can be attributed to the interventions instead of any flaws in the research design (Saunders et al., 2009). As for improving internal validity following measures are taken:

- Interpretation issues: data is recorded, transcribed, and coded;
- Interpretation issues: clarifying, additional questions are answered when needed, in order to create an unambiguous holistic picture. A write-up of the transcribed interviews is not sent to the interviewees, to confirm that the findings, and insights are well reproduced, and interpreted. This measure is not taken, as the participants already put a lot of effort in preparations of, and attending the workshop;
- Interpretation issues: a video pitch, and introductory PowerPoint presentation is provided to the workshop participants, to bridge knowledge gaps, and create common understanding;
- Prior to the workshop a questionnaire is sent (for qualitative means only), to determine if knowledge, and expertise on the different subjects is present, and to verify the preconditions of having troubles with addressing fairness in design, and their need of aid in this. Also this pre-survey questions if the materials (pitch, use-case, and e-mail) provided are understood (please find Appendix 17, and 19);
- Socially desirable answers: as certain topics are sensitive, at the start of the workshop, and in the survey it is mentioned that privacy, confidentially, and anonymity is respected. Some of the more sensitive topics are, amongst others: discrimination, current (non-)practices of fairness-addressing in AI-design, and labor law;
- Subjective judgement: as a researcher, it is tried to avoid influencing attendees, as much as possible in the assessment, and collection of data. Amongst others, by avoiding to include own experiences, and opinions, not expressing personal feelings towards, and taking a stance on certain topics covered;
- Influence exercised by the researcher himself: the case-organization is not related in any way to the researcher, other than it is the case for demonstrating the artefact. So there is no incentive to answer differently based upon that relationship-type. Therefore with regard to this item, no additional validity-measures are taken;
- Interpretation issues: the Princeton-case which each participant had to read trough was shortened and translated in Dutch, as seemed that all participants were Dutch natives. The workshop itself was limited to 2 hours, so that it would not be too time-consuming and participants can stay focused;
- Non-response (pre-survey): a pitch is sent, to provide an easy to digest summary of the research topic, and also for teasing participants, making them curious to participate, read the use-case, and fill-in the survey, which were all preconditions for attending the workshop;

External validity

External validity questions the generalizability of findings; the extent to which you can generalize results to other conditions or target groups (ADM-developing organizations).

The workshop can be seen as a single case-study. However, it doesn't present a very straightforward, typical or extreme case. Therefore sayings like: "if it works here, it should work everywhere", or "if it doesn't work here, it will work nowhere" are not applicable. It is just a first, gentle touch of a theoretically build prototype, i.e. proof of concept, with daily practice in a randomly chosen, easy accessible, AI-expert organization. Generalizability of findings are therefore highly limited.

Construct validity

Construct validity is about measuring, what is intended to be measured (Saunders et al., 2009). Within this study, construct validity is safeguarded by properly defining and operationalizing the definitions of terms and concepts to be researched (e.g., fairness). As for the workshop a complete set-up is done, partly to avoid ambiguities, and give a thorough explanation of the research topic, as detailed in Appendix 12, and Chapter 5.2.

5.5.2. Reliability

Reliability is about the ability of repeating the same research, and where it is to be expected that consistent results will be found (Saunders et al., 2009).

This also minimizes errors and biases in this study. The following measures have been taken to minimize risks, like bias and errors:

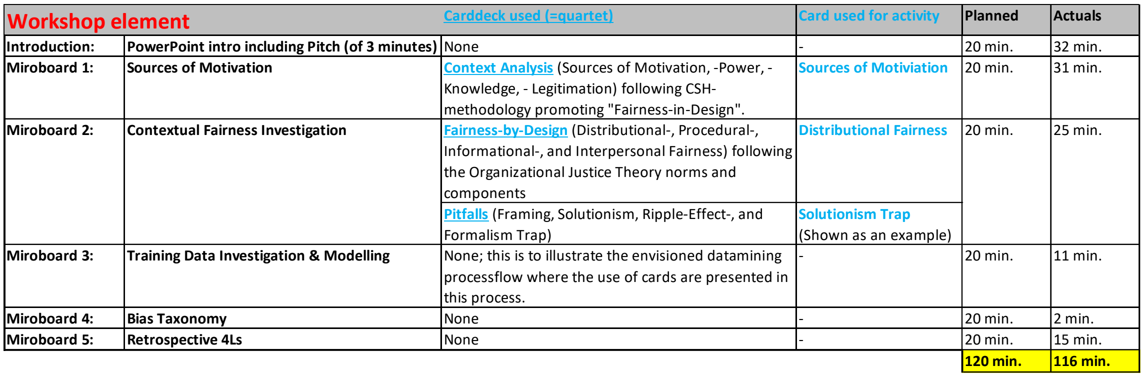
- Semi-structured interview protocols are used, and available;
- The workshop is coded independently, and notes are made during the workshop;
- A chain of evidence is created, both recordings, and notes on the whiteboard are stored, and can be consulted afterwards;
- As said, transcripts are not validated by participants, as for correct presentation, and interpretation, as the workload for this case-organization is already high in preparing for, and participating in the workshop.

5.5.3. Ethics

Participants are not mislead for participating by e.g. promises, gifts, and so on. Also before recording audio and/or video, informed consent is obtained. As earlier said: privacy, confidentially, and anonymity is respected. Each participant, participates voluntarily, and every participant is free of withdrawal. No harms, delusions, or whatsoever apply, as far is known.

6. Artefact Evaluation

In this Chapter the activities are described, and the results of the workshop are presented and discussed. The workshop was structured as follows, including timetable (planned versus actuals):



Workshop element		Carddeck used (=quartet)	Card used for activity	Planned	Actuals
Introduction:	PowerPoint intro including Pitch (of 3 minutes)	None	-	20 min.	32 min.
Miroboard 1:	Sources of Motivation	Context Analysis (Sources of Motivation, -Power, -Knowledge, - Legitimation) following CSH-methodology promoting "Fairness-in-Design".	Sources of Motiviation	20 min.	31 min.
Miroboard 2:	Contextual Fairness Investigation	Fairness-by-Design (Distributional-, Procedural-, Informational-, and Interpersonal Fairness) following the Organizational Justice Theory norms and components	Distributional Fairness	20 min.	25 min.
		Pitfalls (Framing, Solutionism, Ripple-Effect-, and Formalism Trap)	Solutionism Trap (Shown as an example)		
Miroboard 3:	Training Data Investigation & Modelling	None; this is to illustrate the envisioned datamining processflow where the use of cards are presented in this process.	-	20 min.	11 min.
Miroboard 4:	Bias Taxonomy	None	-	20 min.	2 min.
Miroboard 5:	Retrospective 4Ls	None	-	20 min.	15 min.
				120 min.	116 min.

Figure 6-1: Workshop sections planned vs actuals

As it was the intention to discuss all Miro White Boards, it seemed a little short in time, so that at a certain point the pace was increased. Therefore the whiteboard presenting the bias-taxonomy was only shortly explained, and shown. Next to that, also a change as compared to the initial plan is, that the "after survey" is not sent, as participants already invested a lot of their time in preparing, and attending the workshop. In the next section the Workshop Findings are presented.

6.1. Workshop Findings per Element/Activity

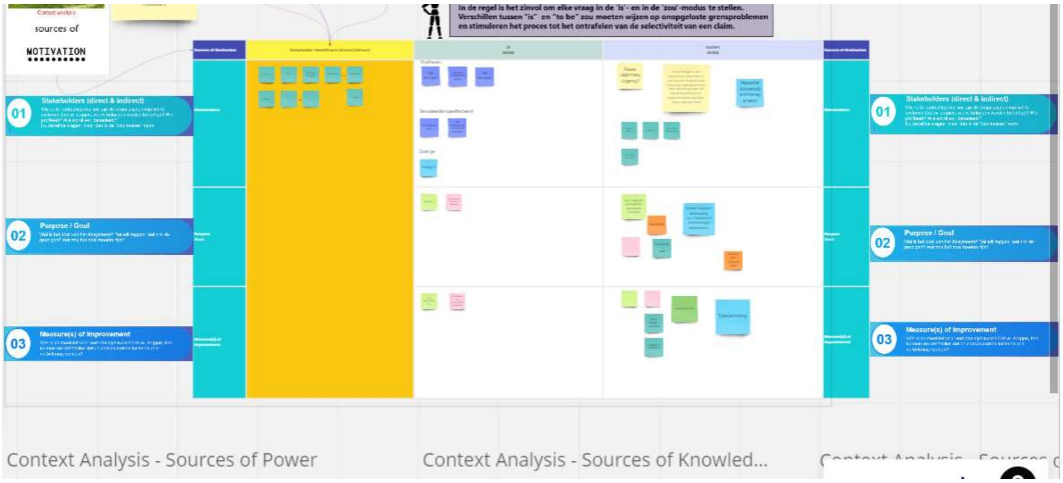
Here the different workshop sections are discussed in chronological order. Every subsection represents one of the several Miro Online Whiteboards and associated activity treated within that particular part of the Workshop.

6.1.1. Activity 1: Context-Analysis

As for this first activity in the workshop, a single card (="Sources of Motivation") of the so called "Context-Analysis" quartet was being used on the Princeton use-case. Participants filled in the descriptive (is-)mode, and prescriptive (ought-)mode; they identified several stakeholders, goals, and measures of improvement in both modes by reflecting upon the fictional use-case (please find figure 6-2). The systematic, and purpose of this activity seemed to be clear. The data science manager mentions here, that awareness around the importance of treated themes (stakeholders, goals, measures of improvement) is already present within the organization, but that predominantly the added value in this, would be the overall, standardized list with clear guidelines to be used company-wide, as in the current mode these considerations are left to the individual:

"I think we are aware of this sort of thing. What I think we can do even more is if we do projects for clients, then, now we're very dependent on the people doing the project if they think about this enough and I think something is what I like would like to have is that everyone can give something to this [check]list, it must all meet this. So just clear guidelines." MK

The Enterprise Architect adds that it is important to consult and involve the identified stakeholders (in step 1) in determining the Purpose/Goal (step 2), and Measures of Improvement (step 3) of the envisioned ADM. This is also how it is done at the present time within the case-organization. This is also the way this "Fairness-in-Design"-approach intends to work.



In de regel is het zinvol om elke vraag in de 'is'- en in de 'zou'-modus te stellen. Verschillen tussen "is" en "to be" zou moeten wijzen op onopgeloste grensproblemen en stimuleren het proces tot het ontrafelen van de selectiviteit van een claim.

Context analysis
sources of
MOTIVATION
01
Stakeholders (direct & indirect)
02
Purpose / Goal
03
Measure(s) of Improvement
01
Stakeholders (direct & indirect)
02
Purpose / Goal
03
Measure(s) of Improvement
Context Analysis - Sources of Power
Context Analysis - Sources of Knowled...

Figure 6-2: Miro Board with Activity #1 - Context Analysis

6.1.2. Activity 2: Contextual Fairness Investigation

The second activity, is named Contextual Fairness Investigation. Here the systemization of Value Hierarchies is outlined. A single card (=distributional fairness) of the "Fairness-by-Design" quartet was introduced, in which several rationales of (decision-outcome)distributions are provided. The Miro Board (figure 6-3) provided some additional instructions. The participants seemed to have difficulties with the terminology of defined fairness-components within the Distributional Fairness card (consisting of Equality, Equity, Need, and Sufficiency):

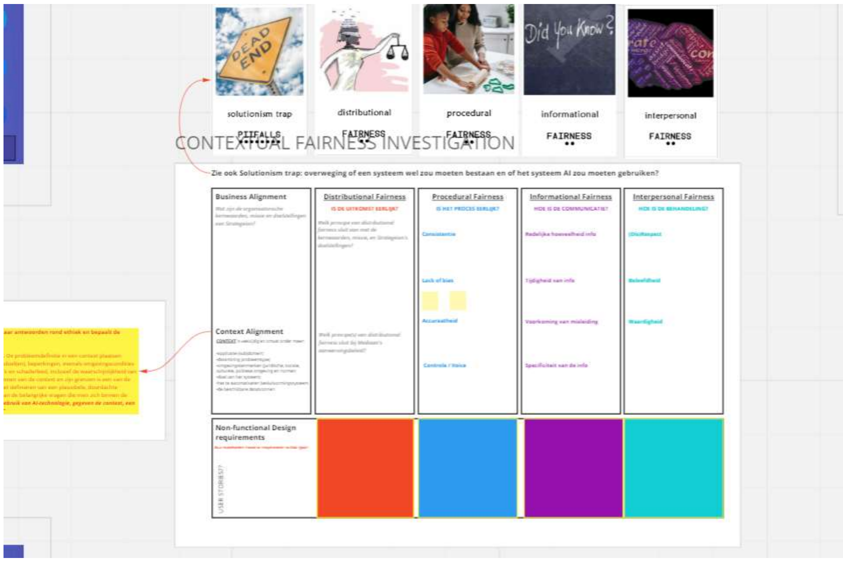
"[..] as for these four things I'm a bit sceptical. For example, just the nuance: everyone should get the same or everyone should get the same chance. Yes, something completely different and you can interpret that in very different ways, so that's why I find this a difficult one." MK

With regard to Contextual Fairness Investigation, the Enterprise Architect underlines the importance of organizational alignment in the quest for addressing fairness:

"[..] the approach is good, to think about it beforehand [=top-down approach]. But I say for us in [organizational] culture, in a company, [..] when all has to be ethical [..] responsible, 100% and you name it, then maybe you shouldn't risk this sort of thing. Saying I'll leave it at this, because there is still a lot of uncertainty and risk that I disadvantage a party after all, or [..] well, I accept it to a certain extent. I am willing to take that risk, I can do the image damage, I can absorb it somewhere, and then it is something else." JS

Herewith, the importance of the Solutionism Trap is confirmed as well, as this pitfall is to consider if a system should exist at all, given the ethical risks. Albeit that alignment with organizational culture is one of the considerations in making this decision.

The activity itself seemed difficult for the participants, so that at a certain point this activity was interrupted, by presenting the ideas behind the systemic; providing some illustrations of how the approach towards fairness-by-design is envisioned.



solutionism trap PITFALLS

distributional FAIRNESS

procedural FAIRNESS

informational FAIRNESS

interpersonal FAIRNESS

CONTEXTUAL FAIRNESS INVESTIGATION

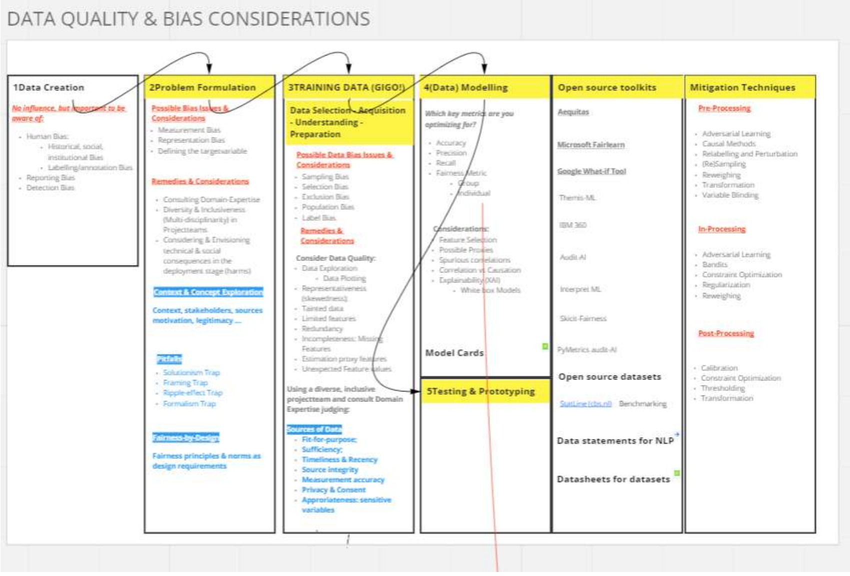
Zie ook Solutionism trap: overweging of een systeem wel zou moeten bestaan en of het systeem AI zou moeten gebruiken?

Business Alignment	Distributional Fairness	Procedural Fairness	Informational Fairness	Interpersonal Fairness
	IS DE UITKOMST EERLIJK?	IS HET PROCES EERLIJK?	HOE IS DE COMMUNICATIE?	HOE IS DE BEHANDELING?
		Consistentie	Redelijke hoeveelheid info	(Dis)Respect
		Lack of bias	Tijdigheid van info	Beleefdheid
Context Alignment		Accuraatheid	Voorkoming van misleiding	Waardigheid
		Controle / Voice	Specificiteit van de info	
Non-functional Design requirements USER STORIES?				

Figure 6-3: Miro Board Activity #2 - Contextual Fairness Investigation

6.1.3. Activity 3: Showing the envisioned Process

To provide the participants a helicopter view of the envisioned process, where and when to use the cards within a traditional ML/Data-Science process, like CRISP-DM, this Miro Board was created:



DATA QUALITY & BIAS CONSIDERATIONS

1Data Creation

No influence, but important to be aware of:
- Human Bias:
  - Historical, social, institutional Bias
  - Labelling/annotation Bias
- Reporting Bias
- Detection Bias

2Problem Formulation

Possible Bias Issues & Considerations
- Measurement Bias
- Representation Bias
- Defining the targetvariable

Remedies & Considerations
- Consulting Domain-Expertise
- Diversity & Inclusiveness (Multi-disciplinarity) in Projectteams
- Considering & Envisioning technical & social consequences in the deployment stage (harms)

Context & Concept Exploration

Context, stakeholders, sources motivation, legitimacy ...

Pitfalls
- Solutionism Trap
- Framing Trap
- Ripple-effect Trap
- Formalism Trap

Fairness-by-Design

Fairness principles & norms as design requirements

3TRAINING DATA (GIGO!)

Data Selection - Acquisition - Understanding - Preparation

Possible Data Bias Issues & Considerations
- Sampling Bias
- Selection Bias
- Exclusion Bias
- Population Bias
- Label Bias

Remedies & Considerations

Consider Data Quality:
- Data Exploration
  - Data Plotting
- Representativeness (skewedness)
- Tainted data
- Limited features
- Redundancy
- Incompleteness: Missing Features
- Estimation proxy features
- Unexpected Feature values

Using a diverse, inclusive projectteam and consult Domain Expertise judging:

Sources of Data
- Fit-for-purpose;
- Sufficiency;
- Timeliness & Recency
- Source integrity
- Measurement accuracy
- Privacy & Consent
- Approriateness: sensitive variables

4(Data) Modelling

Which key metrics are you optimizing for?
- Accuracy
- Precision
- Recall
- Fairness Metric
  - Group
  - Individual

Considerations:
- Feature Selection
- Possible Proxies
- Spurious correlations
- Correlation vs Causation
- Explainability (XAI)
  - White box Models

Model Cards

5Testing & Prototyping

Open source toolkits

Aequitas

Microsoft Fairlearn

Google What-if Tool

Themis-ML

IBM 360

Audit AI

Interpret ML

Skicit-Fairness

PyMetrics audit-AI

Open source datasets

StatLine (cbs.nl) Benchmarking

Data statements for NLP

Datasheets for datasets

Mitigation Techniques

Pre-Processing
- Adversarial Learning
- Causal Methods
- Relabelling and Perturbation
- (Re)Sampling
- Reweighing
- Transformation
- Variable Blinding

In-Processing
- Adversarial Learning
- Bandits
- Constraint Optimization
- Regularization
- Reweighing

Post-Processing
- Calibration
- Constraint Optimization
- Thresholding
- Transformation

Figure 6-4: Miro Board Activity #3 - Demonstrating the Envisioned Overall Process when using the Cards

The blue labels, and blue words respectively, represent the Quartets, and Cards to be used. As time was limited, it only staid with explaining the envisioned overall process, at what moments, which quartets, could be how, of use. Additionally in this overall process also the Bias-taxonomy, Data-cards, Bias-Mitigation Techniques, and Open Source Toolkits were presented in a quick overview.

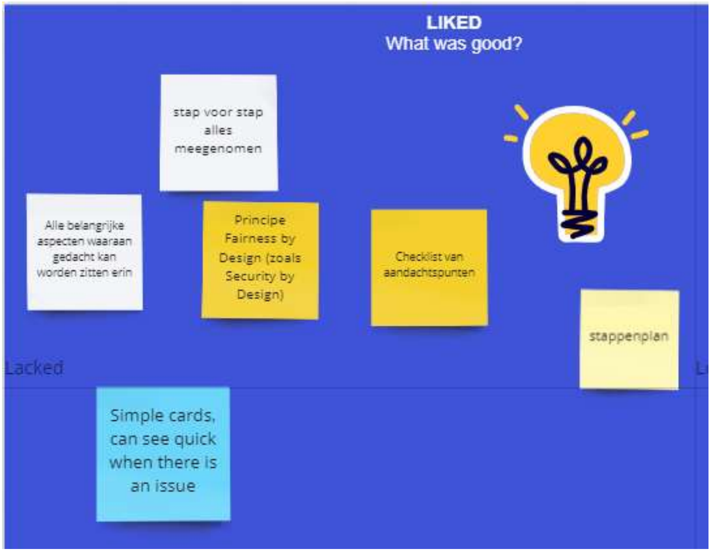
6.2. General Overall Impression using the 4L’s

At the end of the workshop the last 15 minutes were used, to do a holistic review i.e. retrospective analyses in the form of the 4L’s: “Liked”, “Learned”, “Lacked”, and “Longed for”. These different aspects are theme-accordingly discussed in this section.

6.2.1. Liked – “What was good?”

To start with the good parts of the artefact, the workshop participants liked the idea of a checklist, overall guidance in the design-process, in which fairness is implemented by default. The following “liked”-experiences of participants were mentioned:

- Step-by-step everything included;
- All important aspects that can be considered are in it;
- Principle of Fairness by design (like security-by-design);
- A roadmap and checklist of attention-points;
- Simple cards, so that you can see quickly when there is an issue.



LIKED
What was good?

stap voor stap alles meegenomen

Alle belangrijke aspecten waaraan gedacht kan worden zitten erin

Principe Fairness by Design (zoals Security by Design)

Checklist van aandachtspunten

stappenplan

Simple cards, can see quick when there is an issue

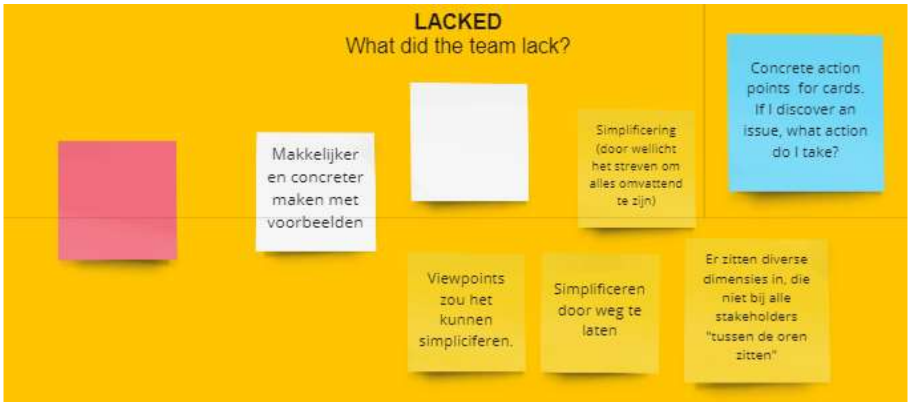
Figure 6-5: Retrospectively Liked

This underlines earlier mentioned finding, that a clear checklist is welcomed:

“And the good thing about it is that you put it all together, or at least you have quite the checklist of all kinds of things [..]” (JS)

6.2.2. Lacked – "What did the team lack?"

Here the remarks during the workshop, and the retrospective analysis is combined on the topic of what is missing in the artefact, according to the participants:



LACKED
What did the team lack?

Makkelijker en concreter maken met voorbeelden

Simplificering (door wellicht het streven om alles omvattend te zijn)

Concrete action points for cards. If I discover an issue, what action do I take?

Viewpoints zou het kunnen simpliciferen.

Simplificeren door weg te laten

Er zitten diverse dimensies in, die niet bij alle stakeholders "tussen de oren zitten"

Figure 6-6: Retrospectively Lacked

All participants agreed that it comes down to simplifying the artefact by dropping content and keeping the artefact concise in its form, next to detailing specific action points, i.e. hints in the content of cards regarding fairness and bias in the design. Suggestions for simplifying things are:

1. By making the artefact more specific by use of examples;
2. By omitting i.e. dropping content, although suggestions fail which content should fall;
3. And/or by allocating different parts of the framework to different stakeholders, i.e. domain experts, for example by using viewpoints (this is also related to the topic of defining roles and responsibilities, referring to section 6.3.3);
4. There are elements in the card-deck which not all stakeholders are familiar with.

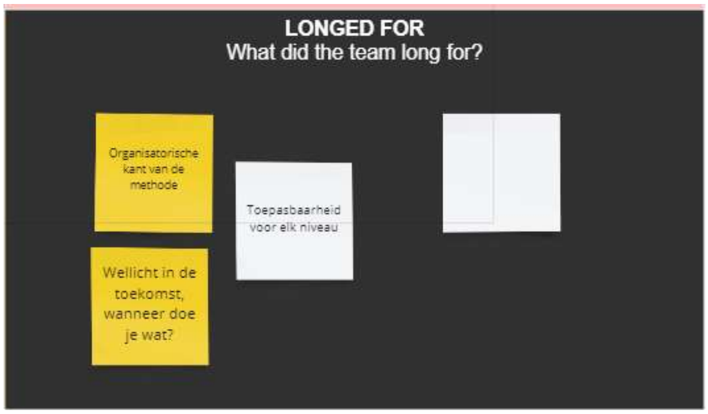
Additional remarks during the workshop underline this:

> *"…I think, when you see the most coming back on this board, simplifying and making it more specific, [..] in practice it just works that way. People don't feel like reading very much, or so if you really want it to be used, it just has to be simple." (MK)*

But at the same time, some detailed information is missing. More specifically, there is a need for specific action point for cards, as one of the notes on the board reveal: "*if I discover an issue, what action do I take?"* The suggestion is to put some inspirations, and hints on the cards, not super detailed, but pointing in the rough direction:

> *"It doesn't have to be super detailed, .. but you do need something of a hint. [..] Being pointed in the rough direction, from here you have to do about this and that. That could be the outcome of It [integrated in the artefact]." (FS)*

6.2.3. Longed for – “What did the team long for?”



LONGED FOR
What did the team long for?

Organisatorische kant van de methode

Toepasbaarheid voor elk niveau

Wellicht in de toekomst, wanneer doe je wat?

Figure 6-7: Retrospectively Longed For

Quite similar to earlier mentioned action points at the “lacked” sheet, one denotes on a yellow: Maybe in the future: which actions do I take at what stage?

The team longed for an organizational setup of the method in terms of roles, functions to assign who does what, as one of the notes on the board reveal. Next to that, pinpointing action-points to different design-stages:

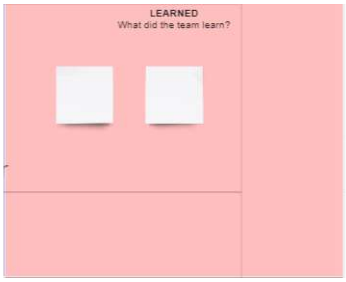
> *“Look Frank, I don't know what you have described with regard to this method, this framework, but it is important that you indicate who [=which role/function] should do this [pointing at sources of motivation, amongst others executing a stakeholder analysis].” (JS)*

Also the Data Science Manager, argues that different roles, and responsibilities are to be assigned to the different activities the card-deck pursues to undertake. High-level decisions are linked to board-level roles, as the Data Scientist is said to be in need for a much more concise, smaller, and framed toolkit. For example, considering if the amount of data suffices. At the moment these data-considerations are left to the individual Data Scientists, out of necessity, as no one is responsible for that.

Also written on the board: Applicability for every level; referring to roles, and functions again. The senior data scientist (FS) even suggests making a Scrum+ version, in which these cards are used, added up with defining roles and responsibilities in this Scrum+ version.

6.2.4. Learned – “What did the team learn?”

With regard to this perspective no “yellows” were attached.



LEARNED
What did the team learn?

Figure 6-8: Retrospectively Learned

6.3. Best-Practices and Other Insights

During the workshop, best-practices regarding the problem-formulation, i.e. conceptual design-stage, in daily AI-design is discussed.

6.3.1. Best Practice

One best practice is to make the datamining problem smaller, using the AI-component for a simpler, smaller, objective decision-part, creating less uncertainties:

> *"Yes, you see now that AI is much more popular, you see people trying to solve everything with AI, but, the bigger you make the problem, the more complicated things like this get. And often it is precisely the smaller solutions that AI is good at. So that's one thing I keep in mind."* (MK)

This entails considering which datamining problem is going to be solved, which part is going to be solved with AI. Aiming at "low hanging fruit", "low impact-high value-stuff", simple automation tasks using AI, taken together creating considerable added value, but with little (ethical) risks.

6.3.2. Other Insights

During the workshop several upcoming themes were discussed, which are collected in this section.

Importance of Fairness in ADM's

As for the introductory part of the Workshop there was a discussion between the participants, why this "fairness-theme" is particularly under a magnifying glass at current times. Reference is made to systems of the "old world", with "if" and "else" statements, in which automated decisions were being made as well. Systems where it occurred as well that things good go (and even went) wrong, resulting in discriminatory harms. The participants agree that this has mainly to do with transparency. In the "old world" the process leading to outcomes can be followed, humans make the assumptions, they code it, and therefore systems can be tested much better to prevent abuses. When things go wrong, you can point at someone/something. The "new world" of AI/ML is said to be less transparent, in which unforeseen things happen. Bias, which is not explicitly appointed, although having an effect because not all factors are taken into account. In the "new world" the assumptions are made by the algorithms, causing uncertainties, and unpredictable behavior, so that additional risk management needs to address this.

Alignment

An emerging theme during the evaluation of the artefact is that alignment should be searched for between fairness to be embedded in the ADM and the organizational culture, and organizational mission, and goals.

> *"I think what's also important is, whatever you rightly say in the surrounding field of that context, that [organizational] culture is very important."*

Experiences

The Enterprise Architect identifies that it is important to think ahead, but that in practice there are always unforeseen issues, creating a learning effect. So that you actually get a top-down vision, and a feedback spiral from practice, which lets you start again from the top.

7. Discussion, Conclusion and Recommendations

Within this Chapter research findings are interpreted, and compared with the existing Body of Knowledge, answering the research question by means of a discussion. In §7.1 the narrow discussion, and reflection is being treated, hereafter in §7.2 until §7.5, the main conclusions, limitations, recommendations for respectively practice, and further research are discussed.

7.1. Discussion – Reflection

Within this design study, it is confirmed, that practitioners find it challenging, and therefore welcome additional guidance in embedding fairness in their ADM-development processes, by informative, light-weighted tooling's (e.g., Lee & Singh, 2020). When combining insights of the artefacts' evaluation session, and feedback during the demonstrative workshop, it is found that ADM-developers seem to experience added value, in the shape of checklists. Preferably these checklists are being used company-wide, so that design-choices around fairness are not left to the individual.

The holistic, sociotechnical, top-down (fairness)approach the artefact envisions, is confirmed as a plausible, well-founded approach, in that respect, that not all things can be thought of in advance. It is said that experiences in the field can create a spiralizing feedback-loop, which can mature the artefacts' content. The eponymous building block "Experiences" positioned (left) in our conceptional "Fair-ML"-framework (figure 3-5), intends to accommodate these practical findings. It can be concluded that findings within this study underline the relevance of this conceptual building block and its function.

In the multiple HR case-study it is found, that fairness is not guaranteed by fairness-metrics solely (e.g., risks of stereotyping), but also relates to determining a justified solution space, "human-in-the-loop"-considerations, and deliberation on complex phenomenon's as diversity and inclusion. Latter HR-findings are therefore interpreted as an acknowledgment of the promising step forward, when using a holistic, sociotechnical ADM-design-approach.

Findings suggest, that in this first iteration, the card-deck is experienced as a welcomed (extensive, partly overwhelming) checklist, catalyzing this sociotechnical, top-down approach, although several important improvement points are suggested, as well as add-ons of best practices.

An unanimous improvement suggestion, is to simplify the artefact by dropping content. This underlines the developers' quest for a pragmatic, light-weight tooling. However, there seems a tension point between detailed information on addressing fairness in the design, and keeping the artefact concise. On the one hand findings reveal that it has to be simple, and pragmatic; practitioners[16] don't want to read a lot. On the other hand, practitioners long for hints, and tips, by making the artefact more concrete in its use in specific situations. Practitioners recognize that a lot of issues are to be taken into account, when designing an impactful ADM for fairness. Therefore, determining what should be left out, and simplifying the artefact is (said to be) a difficult task.

A hereto related improvement point, coming forward, is to provide an organizational setup by defining roles and responsibilities. Aligning the artefacts' content to these organizational role(s) is requested by the participants (as not all topics covered within the artefact are equally covered by their expertise/ roles). The adjustment of role-aligned content in the artefact, could also be beneficial, as to aid in simplification, i.e. reducing information overload for its users. However, the original intended approach of the artefact is to use a multi-disciplinary design-team in developing "ethical-risky" ADM's from the start. This approach is preached, and grounded in academic RAI-

[16] In mentioning practitioners here reference is made to the participants of the Workshop.

literature (Draude et al., 2019; Mikalef & Gupta, 2021). Therefore, in the initial design of this artefact every card is intended to be used by, and discussed within this cross-disciplinary team, and not to be assigned to (task)specific roles. Roles are therefore not (yet) provided.

Moreover, the remaining cards in the CSH-based card-quartet, called "Context Analysis" should partially compensate for this matter. These cards are not yet evaluated within this study due to time-framing, but the thematic content of each of these cards, are divided in (CSH-theme accordingly): sources of power, -knowledge, and - legitimacy. Each referring to respectively defining decision-making authority, expertise to be consulted, and including the voice of affected parties (=co-design). These cards envision a democratic, participatory, "fairness-in-design" approach, intending to offer a dynamic, project-dependent setup of essential roles to be included, on which the involved parties should agree upon.

When scanning literature it is found that defining roles and responsibilities, and allocation of decision-making authority is necessary, but substantively this literature is sparsely, when compared with matured fields, like e.g., Data Governance (Schneider, Abraham, Meske, & Vom Brocke, 2022). Roles, responsibilities, and accountabilities related to AI and its impacts are often ambiguous. It is indicated that more research is needed (e.g., for Model and System governance), as there is a considerable variance in these regarding AI-ethics (Schneider et al., 2022).

Establishing an interdisciplinary AI council for AI, executive sponsorship, and more specific roles related to model aspects are proposed, but still subject to investigation. Also participatory design, i.e. co-design initiatives, to elicit user values, are mentioned. Assignment of an owner to each design feature is another suggested approach (Serban, Poll, & Visser, 2020). It is even indicated that an interorganizational setup is required (Schneider et al., 2022), as designs of ADM's often navigate across company boundaries. This is not surprising, as a one-size-fits all approach is unlikely to be feasible, when an organizational setup is interdependent on organizational characteristics (organization maturity, size, principles, culture, policy, and so on). Numerous scholars are found to propose to equip technology teams with a "value advocate" (Shilton & Anderson, 2017), or similar "value lead"(Spiekermann, 2017); a technical member with a strong ethics background, explicitly in charge of ethical, value-oriented design.

In terms of content, one suggested improvement point specifically towards the "Fairness-by-Design"-quartet is to nuance the components of the OJT – distributional fairness dimension - in the card-deck more, leaving less space for multi-interpretability. Next to a more nuanced fairness-palette, business culture alignment considerations in searching for fairness are also an suggested add-on in the artefact.

An interesting, revealed best practice, towards situations where risks of impactful, ethical harm in design is lurking, is to downsize the AI-solution. Using AI where it is good at, the low hanging fruit: small, simple, and objective decision(part)(s). Taken together many of these so called "low ethical impact, high value" improvements together are said to bring big gains, with low ethical risks. This is an important alternate consideration, when designing, which can be integrated as an additional guidance point in the artefact. This is an add-on in scientific literature as well, in which the preached main approach is top-down, by e.g. consulting stakeholders, or using narrowed fairness-metrics. Either way, in both cases leaving the overall decision to AI, instead of downsizing the AI-decision-component. The best practice reveals a kind of bottom-up approach, as it is tried to downsize, and simplify the datamining problem, so that ethical risks are lowered, and therefore several (top-down) considerations around e.g., stakeholders affected, law- and regulations, are in all probability made superfluous.

7.2. Conclusions

In this design study an overall conceptual framework envisioning a socio-technical angle of approach is shaped, in response to industries' call for light-weight, integrable tooling, (while informative enough to be of help) in addressing the embedding of fairness in ADM-design.

A coherent set of concepts derived from this framework is cast in a generative, card-based tool, and evaluated in a first iteration by means of a workshop within a single case-organization, to answer the main research question:

MRQ*: What is the added practical value of a sociotechnical approach in the shape of an integrative, light-weighted, artefact/framework towards fairness-enhancement in the development-process of supervised, algorithmic decision-making systems?*

It can be stated that this question can only be partly answered:

- It is found, that the embedding of fairness is challenging, and that an extensive, standardized checklist in a light-weight form is welcomed;
- Therefore an important, but prudent conclusion remains that the continuation of the development of such an artefact is found to be worthwhile from a practical point of view, as practitioners from the evaluation-setting seem to experience the artefact partly as a checklist, integrated in a card-deck;
- The "light-weightedness" of the artefact in its current form, can be disputed, as unanimously the practitioners within the evaluative AI-expert organization call for simplification, by omitting content, and/or tailoring content to specific expertise roles.
- Also training, and education in related topics, e.g. AI-ethics could aid in this, and may form a necessary requisite;
- An organizational setup is requested, but literature reveals ambiguous, and sparse material regarding roles and responsibilities in AI/ADM-development.

Further, it can be concluded that a design research approach in terms of complexity, and multidimensionality of the research topic is justified, as developing ADM's is a multidisciplinary endeavour, in which complex interrelations between regulatory, business requirements, ethics, training data, and model outputs data occur. More iterations are needed, for both confirming, and evaluating the artefacts' viability, and to deal with this multifaceted, complex, field of research.

7.3. Limitations of the Research

By reflecting at the workshop, several weaknesses in this empirical research, resulting in research limitations are to be mentioned:

- First, the participants seemed to lose track at certain times during the workshop. At several moments they indicated an information-overload, but at other moments more details were asked for (e.g., situational action-points). The latter is a well-known potential weakness of card-based design tools, as Roy and Warren (2019) indicate. Card decks may overload users with too much information, or oversimplify information due to space limitations; balancing the right level of information is a difficult task. This is also confirmed in this study.
  Also cards bear the risk of being too complicated for users to understand and apply. Which at some moments seemed to be the case, as one of the workshop-activities was discontinued, due to the fact that participants didn't understand the assignment;

- Second, it seemed that the materials covered in relation to a 2-hour, single workshop were too extensive. The theory covered within the session, requires a lot of knowledge bridging. Despite the offered pitch, and PowerPoint intro, it seemed a lot to digest for the participants. Participants were seemingly overwhelmed, and intimidated by the amount of information they were confronted with. This also suggests, that a reiteration should consider more than one workshop for testing, and/or less material to be treated within each workshop. The RAI-maturity-level, and ethical awareness of the organization is also important in that consideration;
- Third, conducting a workshop requires expertise. As a researcher, a first, and only experience in conducting a workshop, was the pilot-session being held within this study. At several times it appeared difficult to do less talking, and energize, and guide the group to active participation in performing the tasks prepared;
- Fourth, the focus within the early stages of this study was too divergent, as this topic covers and touches multiple professional disciplines, narrowing the research scope is extremely necessary. The artefact evaluation, i.e. perceived usefulness could be improved in terms of measurability/operationalization.

These weaknesses taken together, have very likely clouded the intended use of the artefact. Participants, have reflected upon a version of the artefact, that they could understand. However, it is very doubtful that this was the artefact in its complete form.

Also it must be mentioned, that within this study the generalizability of findings is highly limited. This is only a single, explorative iteration, within one case organization. Ideally there are more iterations conducted in the design-cycle. However, taking the workload, and the research time-frame in consideration, this would be very unrealistic. Organizing, preparing and attending a workshop requires a lot of time, and commitment from both participants, and researcher.

7.4. Recommendations for Practice

A recommendation for practitioners (developers, and AI-developing companies) is to consider the solution space, when designing ADM's for fairness. By this, it is meant that it is advisable to include considerations of preventing ethical harms by downsizing the datamining-problem, and fine-tuning the objectives. So called "low-impact, high-value" initiatives can be appealing when using an AI-component, certainly when mechanisms for addressing fairness in design fail.

Another recommendation would be to consider a sociotechnical approach towards fairness-embedding in AI. An ADM can produce fair outcomes in terms of distributional fairness towards groups, or individuals, depending on the fairness-metrics chosen. However, defined groups on sensitive attributes (e.g., men, Caucasians) are not monoliths. Diversity, and inclusiveness is created in the mix. Therefore, procedural, informational, and interpersonal mechanisms for fairness are least as important for ADM-development within an organizational setting. Evoking ethical awareness, and encouraging ethics education, and training are highly recommendable, and enablers for a sociotechnical fairness approach in ADM-design.

7.5. Recommendations for Further Research

A suggestion for follow-up research would be to reiterate, with an improved version of a smaller set of cards (after implementing action-points/hints, solution space considerations, nuanced fairness-norms, and organizational roles/responsibilities). More ideally one would use an EFG, and a Confirmatory Focus Group (CFG), as for evaluation, and valid determination of artefact-improvement (Tremblay et al., 2010), with only a limited set of cards. Also, in future iterations, the bias taxonomy, and other cards developed, e.g., “Data Quality”, corresponding to different stages of ADM-development (e.g., data understanding) could be evaluated. A precondition for successful evaluation when choosing a case-organization is a certain level of knowledge, and experience in ethical design activities, i.e. Responsible AI-practice, as otherwise too much knowledge has to be bridged.

When searching for a more granular form of OJT-fairness components, a recent, scientific article is found, named *AI Fairness in Organizational Decisions: Conceptualization, Measurement, and Attainment* (Rai, Tian, & Xue, 2022), in which AI-fairness literature and OJT is bridged, providing a fairness-taxonomy which could serve as a richer, more nuanced foundation for disambiguation of distributional fairness-components. Another source of inspiration for adjusting the fairness components might be Moorman’s parsimonious representation of justice, as research findings indicate that these measures may dominate Colquitt’s version for explaining the nuances of perceptual differences regarding fairness and justice (B. K. Miller, Konopaske, & Byrne, 2012).

Identifying, and reflecting upon fairness requirements is one challenge, but choosing between conflicting values is even more challenging. A mechanism to aid in these trade-offs (also a building block in our conceptual framework, figure 3-5), by resolving value tensions, is considered as a very important next step in artefact development. Perhaps shortcuts in the shape of e.g., “Value Dams & Flows”, which consider impact, versus magnitude, could be used to guide these choices determine the “heaviness” of the ethical design process upfront (J. K. Miller, Friedman, Jancke, & Gill, 2007). Follow-up research could address this important step in maturing the artefact.

Last but not least, the organizational part, as to refer to roles and responsibilities could be investigated. A first evaluation could be performed with the card-deck of “Context Exploration”, in which the cards regarding expertise, decision-making authority, and knowledge can be used to reveal the added value of a project-dependent setup. As earlier mentioned research on this topic is sparsely, and ambiguous. Several AI-Governance initiatives, and for example the IEEE P7000 norm, could provide some handles for a starting point to spark experimentation, when this project-dependent setup fails. The artefact, i.e. toolkit will lead to ad-hoc design-processes, which should eventually be embedded in a more steady state of affairs. It is expected that the added value of the artefact will be reinforced by supplementary governance mechanisms, as AI-principles, policies, executive sponsorship, and defining organizational roles and responsibilities.

8. References used (including the Appendices)

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6*, 52138-52160.

Albert, E. T. (2019). AI in talent acquisition: a review of AI-applications used in recruitment and selection. *Strategic HR Review*.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica, 23*.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., . . . Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82-115.

Ashurst, C., Barocas, S., Campbell, R., & Raji, D. (2022). *Disentangling the Components of Ethical Research in Machine Learning.* Paper presented at the 2022 ACM Conference on Fairness, Accountability, and Transparency.

Aysolmaz, B., Iren, D., & Dau, N. (2020). *Preventing Algorithmic Bias in the Development of Algorithmic Decision-Making Systems: A Delphi Study.* Paper presented at the Proceedings of the 53rd Hawaii International Conference on System Sciences.

Baer, T. (2019). *Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists*: Apress.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev., 104*, 671.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587-604.

Benjamins, R., Barbado, A., & Sierra, D. (2019). Responsible AI by Design. *arXiv preprint arXiv:1909.12838*.

Bietti, E. (2020). *From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy.* Paper presented at the Proceedings of the 2020 conference on fairness, accountability, and transparency.

Caton, S., & Haas, C. (2020). Fairness in Machine Learning: A Survey. *arXiv preprint arXiv:2010.04053*.

Choraś, M., Pawlicki, M., Puchalski, D., & Kozik, R. (2020). *Machine Learning–the results are not the only thing that matters! What about security, explainability and fairness?* Paper presented at the International Conference on Computational Science.

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM, 63*(5), 82-89.

Cofone, I. N. (2018). Algorithmic Discrimination Is an Information Problem. *Hastings LJ, 70*, 1389.

Colquitt, J. A. (2001). On the dimensionality of organizational justice: a construct validation of a measure. *Journal of applied psychology, 86*(3), 386.

Colquitt, J. A., Greenberg, J., & Zapata-Phelan, C. P. (2013). What is organizational justice? A historical overview. In *Handbook of organizational justice* (pp. 3-56): Psychology Press.

Commission, E. (2020). White paper on artificial intelligence–a European approach to excellence and trust.

Cramer, H., Garcia-Gathright, J., Reddy, S., Springer, A., & Takeo Bouyer, R. (2019). *Translation, tracks & data: an algorithmic bias effort in practice.* Paper presented at the Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems.

Cramer, H., Garcia-Gathright, J., Springer, A., & Reddy, S. (2018). Assessing and addressing algorithmic bias in practice. *interactions, 25*(6), 58-63.

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature, 538*(7625), 311-313.

Cropanzano, R., Bowen, D. E., & Gilliland, S. W. (2007). The management of organizational justice. *Academy of management perspectives, 21*(4), 34-48.

D. Urquhart, L., & J. Craigon, P. (2021). The Moral-IT Deck: a tool for ethics by design. *Journal of Responsible Innovation, 8*(1), 94-126.

Danks, D., & London, A. J. (2017). *Algorithmic Bias in Autonomous Systems.* Paper presented at the IJCAI.

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., & Sen, S. (2017). Proxy non-discrimination in data-driven systems. *arXiv preprint arXiv:1707.08120*.

Dobbe, R., Dean, S., Gilbert, T., & Kohli, N. (2018). A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. *arXiv preprint arXiv:1807.00553*.

Draude, C., Klumbyte, G., Lücking, P., & Treusch, P. (2019). Situated algorithms: a sociotechnical systemic approach to bias. *Online Information Review*.

Fazelpour, S., & Lipton, Z. C. (2020). *Algorithmic Fairness from a Non-ideal Perspective.* Paper presented at the Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS), 14*(3), 330-347.

Fu, R., Huang, Y., & Singh, P. V. (2020). Artificial Intelligence and Algorithmic Bias: Source, Detection, Mitigation, and Implications. In *Pushing the Boundaries: Frontiers in Impactful OR/OM Research* (pp. 39-63): INFORMS.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of management review, 18*(4), 694-734.

Gogoll, J., Zuber, N., Kacianka, S., Greger, T., Pretschner, A., & Nida-Rümelin, J. (2021). Ethics in the software development process: from codes of conduct to ethical deliberation. *Philosophy & Technology, 34*(4), 1085-1108.

Grasso, I., Russell, D., Matthews, A., Matthews, J., & Record, N. R. (2020). *Applying Algorithmic Accountability Frameworks with Domain-specific Codes of Ethics: A Case Study in Ecosystem Forecasting for Shellfish Toxicity in the Gulf of Maine.* Paper presented at the Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference.

Greenberg, J. (1987). A taxonomy of organizational justice theories. *Academy of management review, 12*(1), 9-22.

Greenberg, J. (1990). Organizational justice: Yesterday, today, and tomorrow. *Journal of management, 16*(2), 399-432.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 75-105.

Jacobs, A. Z., & Wallach, H. (2019). Measurement and fairness. *arXiv preprint arXiv:1912.05511*.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389-399.

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research, 13*(3), 795-848.

Koene, A. (2017). Algorithmic Bias: Addressing Growing Concerns [Leading Edge]. *IEEE Technology and Society Magazine, 36*(2), 31-32.

Koene, A., Dowthwaite, L., & Seth, S. (2018). *IEEE P7003™ standard for algorithmic bias considerations: work in progress paper.* Paper presented at the Proceedings of the International Workshop on Software Fairness.

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable algorithms. *U. Pa. L. Rev., 165*, 633.

Kuechler, B., & Vaishnavi, V. (2008). On theory development in design science research: anatomy of a research project. *European journal of information systems, 17*(5), 489-504.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). *Counterfactual fairness.* Paper presented at the Advances in neural information processing systems.

Lee, M. S. A., & Singh, J. (2020). The Landscape and Gaps in Open Source Fairness Toolkits. *Available at SSRN*.

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. (2019). The challenges of algorithm-based HR decision-making for personal integrity. *Journal of Business Ethics, 160*(2), 377-392.

Madaio, M., Egede, L., Subramonyam, H., Wortman Vaughan, J., & Wallach, H. (2022). Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW1), 1-26.

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). *Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI.* Paper presented at the Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.

Magnani, L., Casadio, C., & Magnani. (2016). *Model-based reasoning in science and technology*: Springer.

Manseau, J., & Mbuko, I. (2020). A Design Science Research to Correct Inherent Biases in Natural Language Applications.

Martin, K. (2019). Designing Ethical Algorithms. *MIS Quarterly Executive, 18*(2).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management, 58*(3), 103434.

Miller, B. K., Konopaske, R., & Byrne, Z. S. (2012). Dominance analysis of two measures of organizational justice. *Journal of Managerial Psychology*.

Miller, J. K., Friedman, B., Jancke, G., & Gill, B. (2007). *Value tensions in design: the value sensitive design, development, and appropriation of a corporation's groupware system.* Paper presented at the Proceedings of the 2007 international ACM conference on Supporting group work.

Minnameier, G. (2005). *Wissen und inferentielles Denken: Zur Analyse und Gestaltung von Lehr-Lern-Prozessen*: Lang.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., . . . Gebru, T. (2019). *Model cards for model reporting.* Paper presented at the Proceedings of the conference on fairness, accountability, and transparency.

Narayanan, A. (2018). *Translation tutorial: 21 fairness definitions and their politics.* Paper presented at the Proc. Conf. Fairness Accountability Transp., New York, USA.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., . . . Krasanakis, E. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1356.

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data, 2*, 13.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45-77.

Poel, I. v. d. (2013). Translating values into design requirements. In *Philosophy and engineering: Reflections on practice, principles and process* (pp. 253-266): Springer.

Pries-Heje, J., Baskerville, R., & Venable, J. R. (2008). Strategies for design science research evaluation.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., . . . Barnes, P. (2020). *Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing.* Paper presented at the Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now Institute*, 1-22.

Rovatsos, M., Mittelstadt, B., & Koene, A. (2019). Landscape Summary: Bias In Algorithmic Decision-Making: what is bias in algorithmic decision-making, how can we identify it, and how can we mitigate it?

Roy, R., & Warren, J. P. (2019). Card-based design tools: A review and analysis of 155 card decks for designers and designing. *Design Studies, 63*, 125-154.

Rozado, D. (2020). Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. *PloS one, 15*(4), e0231189.

Ruggieri, S., Pedreschi, D., & Turini, F. (2010). Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD), 4*(2), 1-40.

Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research methods for business students*: Pearson education.

Schäfer, M., Haun, D. B., & Tomasello, M. (2015). Fair is not fair everywhere. *Psychological science, 26*(8), 1252-1260.

Schneider, J., Abraham, R., Meske, C., & Vom Brocke, J. (2022). Artificial intelligence governance for businesses. *Information Systems Management*, 1-21.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). *Fairness and abstraction in sociotechnical systems.* Paper presented at the Proceedings of the Conference on Fairness, Accountability, and Transparency.

Serban, A., Poll, E., & Visser, J. (2020). *Towards using probabilistic models to design software systems with inherent uncertainty.* Paper presented at the European Conference on Software Architecture.

Shilton, K., & Anderson, S. (2017). Blended, not bossy: Ethics roles, responsibilities and expertise in design. *Interacting with computers, 29*(1), 71-79.

Silva, S., & Kenney, M. (2018). Algorithms, platforms, and ethnic bias: An integrative essay. *Phylon (1960-), 55*(1 & 2), 9-37.

Smith, H. (2020). Algorithmic bias: should students pay the price? *AI & society, 35*(4), 1077-1078.

Spiekermann, S. (2017). IEEE P7000—The first global standard process for addressing ethical concerns in system design. *Multidisciplinary Digital Publishing Institute Proceedings, 1*(3), 159.

Springer, A., Garcia-Gathright, J., & Cramer, H. (2018). *Assessing and Addressing Algorithmic Bias-But Before We Get There.* Paper presented at the AAAI Spring Symposia.

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.

Tal, A. S., Batsuren, K., Bogina, V., Giunchiglia, F., Hartman, A., Loizou, S. K., . . . Otterbacher, J. (2019). *"End to End" Towards a Framework for Reducing Biases and Promoting Transparency of Algorithmic Systems.* Paper presented at the 2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP).

Thuan, N. H., Drechsler, A., & Antunes, P. (2019). Construction of design science research questions. *Communications of the Association for Information Systems, 44*(1), 20.

Tremblay, M. C., Hevner, A. R., & Berndt, D. J. (2010). The use of focus groups in design science research. In *Design research in information systems* (pp. 121-143): Springer.

Ulrich, W., & Reynolds, M. (2010). Critical systems heuristics. In *Systems approaches to managing change: A practical guide* (pp. 243-292): Springer.

Van den Hoven, J., Vermaas, P. E., & Van de Poel, I. (2015). *Handbook of ethics, values, and technological design: Sources, theory, values and application domains*: Springer.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society, 4*(2), 2053951717743530.

Venable, J., Pries-Heje, J., & Baskerville, R. (2012). *A comprehensive framework for evaluation in design science research.* Paper presented at the International conference on design science research in information systems.

VSNU. (2017). *Digital Society Research Agenda - Leading the way through cooperation in a Digital Society.* Retrieved from Netherlands: https://www.vsnu.nl/files/documenten/Domeinen/Onderzoek/DigitaleSamenleving/VSNU%20Digital%20Society%20Research%20Agenda.pdf

Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*: Springer.

Wikipedia. (2020). Abductive Reasoning. Retrieved from https://en.wikipedia.org/wiki/Abductive_reasoning

Wikipedia. (2022a). Heuristic. Retrieved from https://en.wikipedia.org/wiki/Heuristic

Wikipedia. (2022b). Organizational justice. Retrieved from https://en.wikipedia.org/wiki/Organizational_justice

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. (2013). Using grounded theory as a method for rigorously reviewing literature. *European journal of information systems, 22*(1), 45-55.

Yetim, F. (2011). A set of critical heuristics for value sensitive designers and users of persuasive systems.

Zhong, Z. (2018). A tutorial on fairness in machine learning. *Medium*.