Friedle, Cosima

Article

How Can Fairness Tools Impact the Understanding of Fairness and the Processes Within a Machine Learning Development Team?

Junior Management Science (JUMS)

Provided in Cooperation with:
Junior Management Science e. V.

Suggested Citation: Friedle, Cosima (2022) : How Can Fairness Tools Impact the Understanding of Fairness and the Processes Within a Machine Learning Development Team?, Junior Management Science (JUMS), ISSN 2942-1861, Junior Management Science e. V., Planegg, Vol. 7, Iss. 5, pp. 1289-1300,
https://doi.org/10.5282/jums/v7i5pp1289-1300

This Version is available at:
https://hdl.handle.net/10419/295021

Junior Management Science 7(5) (2022) 1289-1300

Junior Management Science

journal homepage: www.jums.academy

How Can Fairness Tools Impact the Understanding of Fairness and the Processes Within a Machine Learning Development Team?

Cosima Friedle

London School of Economics and Political Science

Abstract

Over the last years, a wide spread of Machine Learning in increasingly more, especially sensitive areas like criminal justice or healthcare has been observed. Popular cases of algorithmic bias illustrate the potential of Machine Learning to reproduce and reinforce biases present in the analogous world and thus lead to discrimination. The realisation of this potential has led to the creation of the research stream on fair, accountable and transparent Machine Learning. One aspect of this research field is the development of fairness tools, algorithmic toolkits that aim to assist developers of Machine Learning in identifying and eliminating bias in their models and thus ensuring fairness. The literature review on fairness tools has revealed a research gap in the impact of these on the understanding of fairness and the processes within a development team. Thus, the aim of this research was to investigate the impact that fairness tools can have on the notion of fairness and the processes in a development team. Therefore, a case study with a development team of a large, globally operating corporation has been conducted. Applying Kallinikos' theory of technology as a regulative regime and Oudshoorn and Pinch's idea of the co-construction of users and technologies on the empirical findings revealed two important conclusions. Firstly, it shows that fairness tools act as regulative regimes by shaping the understanding of fairness and the processes within a development team. Secondly, this character of fairness tools as regulative regimes needs to be understood as part of the co-construction process between the technology and the developer.

Keywords: Machine learning; Fairness; Fairness tools; Regulative regime of technology; Co-construction of user and technology.

1. Introduction

„There's software used across the country to predict future criminals. And it's biased against blacks." (Mattu, Kirchner, & Surya, 2016). This quote illustrates the potential inherent in Machine Learning (ML) to reproduce and reinforce biases and thus create discrimination (Lepri, Oliver, Letouzé, Pentland, & Vinck, 2018). Popular examples like the COMPAS, a criminal recidivism algorithm that discriminated against black people (Saxena et al., 2019) or Amazon's recruiting algorithm that was biased against women show the severe consequences that unfair AI can have (Holstein & Vaughan, 2019).

Against the backdrop of these popular examples of algorithmic bias, there have been growing concerns in both academia and practice over the deployment of ML applications. Especially since the use of ML applications has spread to increasingly more areas over the last years (Adadi & Berrada, 2018), including to contexts in which decision-making has critical consequences (Saxena et al., 2019), such as healthcare and recruitment processes, this topic has gained momentum. It is argued that given this wide spread and the areas in which these models are employed (Srivastava, Heidari, & Krause, 2019), they can have an enormous, negative impact on the life of many people (Holstein & Vaughan, 2019).

This realisation of the negative potential of ML has created the FAT ML research community which is concerned with Fairness, Accountability, and Transparency in Machine Learning (Gade, Geyik, Kenthapadi, Mithal, & Taly, 2019; Pasquale, 2015). Research within this stream is mainly focused on advancing fair, explainable and accountable algorithmic decision-making models (Adadi & Berrada, 2018). One major theme within the literature on fair ML is the development of fairness tools, which are algorithmic tools that support developers in detecting and eliminating unfairness in ML (Holstein & Vaughan, 2019).

DOI: https://doi.org/10.5282/jums/v7i5pp1289-1300

To date, the research on fairness tools has been mainly driven by two theoretical lenses. The deterministic, technical-rational perspective focuses on the technical functionalities of the tools and assumes that they can be easily implemented and will directly lead to more fairness in ML (Berk, Heidari, Jabbari, Kearns, & Roth, 2021; Calmon, Wei, Vinzamuri, Ramamurthy, & Varshney, 2017). Authors analysing fairness tools from a contingent, socio-technical perspective, criticise them for being too simplified and poorly suited to ensure fairness and call for an increased focus on the broader social context of algorithmic systems (Binns, 2018; Holstein & Vaughan, 2019).

One aspect that has been studied very little is the impact that fairness tools can have on the developers that are interacting with them and, an aspect that has been neglected so far, on the processes within a development team. Some papers suggest that fairness tools impact the knowledge of developers on fairness and bias mitigation, however, this has not been investigated empirically to date (Bellamy et al., 2019; Yan, Gu, Lin, & Rzeszotarski, 2020).

In order to fill this research gap, a case study has been conducted with a development team to evaluate the impact that fairness tools have on the understanding of fairness of the developers and the processes within the development team. The aim of the paper is to answer the research question "How can fairness tools impact the understanding of fairness and the processes within a machine learning development team?".

The essay will be structured as follows. In the literature review, current key debates and approaches to the topic will be critically evaluated and juxtaposed. Following, the theoretical framework used for sense-making of the empirical data and the methodology are outlined. In the next part, the empirical findings from the data collection are presented. The discussion embeds the findings in the context of the current literature and presents the contributions of this study. The conclusion synthesises the central aspects of this study and identifies limitations of this work as well as possibilities for future research.

2. Literature Review

2.1. Machine Learning and Algorithmic Bias

In order to embed the issue of fairness in ML in its context, a short overview of ML and algorithmic bias will be provided in the following.

ML can be described as one of the most important subsets of AI (Rossi, 2018) that enables computers to learn something without being explicitly programmed to fulfil this specific task. It is the development of a mathematical model by the use of training data (Zhang, 2020) with the aim to make predictions and classifications (Binns, 2018). Algorithmic-decision making processes supported by ML are frequently described as potentially leading to fairer decisions because they are able to prevent human bias (Lepri et al., 2018). However, this has not proven correct.

With the increasing spread of ML applications in sensitive areas, the concern over the potential of these systems to reproduce and reinforce existing biases (Chouldechova & Roth, 2020) has received a lot of attention (Holstein & Vaughan, 2019). During the learning process, ML models are likely to adopt discriminatory correlations which means that these patterns are transferred from the analogous to the digital world (Binns, 2018). Given the nature of ML models, these biases are scaled and reinforced when translated to the digital world (Garcia, 2016). Bias in ML can have a number of different sources, such as the variables, the size of the training set or the decision to deploy an algorithmic system in a certain context itself (Lepri et al., 2018).

2.2. Definition of Fairness

A significant amount of effort within the literature on fair ML has been dedicated to the elaboration of a unified definition of fairness (Holstein & Vaughan, 2019). To date, there exists a number of different fairness definitions, however, no one is commonly accepted (Binns, 2018). It can be argued that the reason for the different definitions of fairness lies in the distinct theoretical lenses that are used (Lepri et al., 2018). These approaches can be divided into two main theoretical lenses.

2.2.1. Reductionist Approaches

Authors representing the reductionist perspective focus their work on the definition of fairness on mathematical considerations. Fairness in ML is understood as a static concept (Liu, Dean, Rolf, Simchowitz, & Hardt, 2018); and various definitions of fairness are considered and compared with regards to their limitations and shortcomings (Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017). However, these are limited to a theoretical, mathematical level.

All authors within the reductionist lens agree on the fact that it is impossible to simultaneously satisfy all different fairness definitions or requirements (Kleinberg, Mullainathan, & Raghavan, 2017; Srivastava et al., 2019). On this basis, it is attempted to reduce and simplify various definitions in order to achieve a combined concept that satisfies as many requirements as possible. This is done by calculating the trade-offs between the individual definitions (Corbett-Davies et al., 2017; Kleinberg et al., 2017) with the aim to minimise the disadvantages (Chouldechova & Roth, 2020). This leads to a high level of abstraction and simplification that is performed by the authors.

Within this theoretical perspective, the importance of the specific context in which fairness in ML should be applied for the definition as well as other environmental and dynamic factors are neglected. It is assumed that one definition of fairness is applicable to all ML contexts (Agarwal, Beygelzimer, Dudík, Langford, & Wallach, 2018).

2.2.2. Contingent Approaches

Within this view, the authors share the assumption that since ML decisions affect people's lives, the definition of fairness should match people's perception of fairness (M. K. Lee

& Baykal, 2017; Saxena et al., 2019; Srivastava et al., 2019). This is underpinned by the idea that for fairness in ML to have a positive impact, it needs to be informed by the common sense of justice (Srivastava et al., 2019). By conducting empirical research, some authors discover that most people perceive demographic parity as the most suitable fairness notion (Srivastava et al., 2019), other authors find calibrated fairness to be the one that represents layperson's perception in the best way (Saxena et al., 2019).

The underlying assumption of this perspective can be described as inspired by the contingency theory. It is argued that fairness is highly contingent (Srivastava et al., 2019) and by investigating people's perceptions of fairness, a better understanding of the appropriateness of different definitions depending on the context can be generated (Saxena et al., 2019). Authors within this approach emphasise the importance of moving beyond the attempts to mathematically define fairness. Instead, they argue that it is imperative to acknowledge the contingency of fairness and establish a process of profoundly investigating the context in order to evaluate the best suitable definition of fairness for a specific problem (Lepri et al., 2018). They frequently criticise that the reductionist approaches to fairness don't reflect the highly context-specific reality of fairness in ML (M. S. A. Lee, Floridi, & Singh, 2021) and that instead, fairness needs to be negotiated every time according to the specific use case (Binns, 2018; Holstein & Vaughan, 2019).

The previous paragraphs have shown that the debate around the definition of fairness is controversial. While authors following a reductionist approach present fairness as easy to define and reduce it to a mathematical notion, authors within the contingent school of thought clearly argue for taking the context of the respective ML system into account when defining fairness.

2.3. Fairness Tools

A major theme within the literature on fair ML is the development of fairness tools. Against the backdrop of algorithmic bias and the necessity to ensure fairness in ML, researchers have developed these tools aimed at helping developers to detect and eliminate cases of bias in their models (M. S. A. Lee & Singh, 2021). These tools are typically classified according to the stage in the development process in which they can be deployed. They are typically divided into pre-processing tools, which eliminate bias from the dataset (Binns, 2018), in-processing tools that build fairness requirements directly into the algorithm (Calmon et al., 2017), and post-processing tools that evaluate and adjust the performance of a model after it has been developed (Berk et al., 2021).

The last years have seen a surge in the development of integrated toolkits in order to facilitate the implementation in development teams (Holstein & Vaughan, 2019). Companies can purchase these toolkits and integrate them into their processes. A commercial example is the AI 360 Toolkit that has been developed by IBM ('The AI 360 Toolkit', IBM Developer, 2021). Having reviewed and critically analysed the broad and interdisciplinary literature on fairness tools, two main theoretical directions could be identified.

2.3.1. Deterministic, Technical-Rational View

Authors within the deterministic, technical-rational view focus on the technical functionalities of the tools and assume that they can be easily implemented and will automatically lead to more fairness in ML (Berk et al., 2021; Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015; Hajian, Bonchi, & Castillo, 2016). In their works, they study the technical functionalities of fairness tools (Calders & Verwer, 2010), the shortcomings of individual tools (Hajian et al., 2016) and how tools can optimise the trade-off between accuracy and fairness (Kamiran, Calders, & Pechenizkiy, 2010; Kearns, Neel, Roth, & Wu, 2019).

One identified limitation is that most tools are not able to account for indirect discrimination that is introduced via proxies that are correlated with sensitive attributes. Therefore, they are not suitable to eliminate systemic discrimination (Berk et al., 2021). The shortcomings that are studied within this perspective are limited to technical disadvantages, such as a loss of accuracy within a tool as compared to another tool (Kamiran et al., 2010). Shortcomings in a wider, socio-technical sense, such as implementation challenges, are not addressed within this approach.

Another aspect studied within this approach is the fairness-accuracy trade-off. It is widely acknowledged that attempts to increase fairness in ML will always be a balance between the performance of the model and fairness (Menon & Williamson, 2018); there will be no technical solution that is able to maximise both fairness and accuracy (Berk et al., 2021). However, the deterministic approach does not address the fact that in particularly sensitive contexts, such as healthcare or criminal justice, this trade-off will need to be considered under another light than in less sensitive contexts. Instead, it is assumed that an optimised balance between fairness and accuracy is applicable for every model and every context (Chen, Johansson, & Sontag, 2018).

The assumption underpinning this approach is that fairness tools are fixed objects that are implemented in a development team and then automatically eliminate the bias from ML models; regardless of the context, the perception and understanding of fairness that is present in the development team, and the general environment of the system. By studying fairness tools from a deterministic lens, they are abstracted from their implementation in real-world development teams. The fairness definition underlying these approaches is a reductionist view. It is assumed that fairness can be formalised independent of the context and then assured through the use of a tool that can be implemented regardless of the context. This means that one tool represents a certain definition of fairness, but no guidance is provided as to which tool should be applied in which context.

2.3.2. Contingent, Socio-Technical View

Within the contingent, socio-technical view, authors study the implementability of fairness tools in practice and

the interaction between tool and developer adopting a socio-technical viewpoint. These works are partly located in the realm of human-computer-interaction research.

A major critic within this approach is the focus of the deterministic view on the technical development of fairness toolkits and an evaluation based on the functional shortcomings. Authors claim that the deterministic lens is narrowly focused on mathematical models and aims at modifying complex models without taking into account the human aspect of the issue and how their solutions can be successfully translated into practice (Abdul, Vermeulen, Wang, Lim, & Kankanhalli, 2018). Issues like the usability of these toolkits in commercial contexts (M. S. A. Lee & Singh, 2021), the challenges faced by practitioners when ensuring fairness in ML (Veale, Van Kleek, & Binns, 2018) or a developer team's general challenges and needs evolving around fairness (Holstein & Vaughan, 2019) are neglected in the field.

In the last years, some papers made an effort in bridging these gaps and answering open questions. M. S. A. Lee and Singh (2021) studied the gaps between the needs of practitioners and the functionalities of fairness toolkits. These include the user-friendliness and the contextualisation of these tools. The work from Richardson, Garcia-Gathright, Way, Thom, and Cramer (2021) focusses on implementation barriers of fair ML faced by practitioner teams and the effectiveness of fairness toolkits. One of the barriers was described as the lack of a contextualisation of a fairness definition within the tools which hindered a successful implementation (Richardson et al., 2021). A study with ML practitioners is conducted in the work of Holstein and Vaughan (2019) in order to investigate their challenges and needs in the work with fairness tools.

Adopting a contingent view on fairness in ML, authors highlight that a socio-technical issue like fairness cannot be tackled only with technological tools. Instead, the socio-technical environment of ML systems and organisational processes need to be studied in order to ensure fairness (Holstein & Vaughan, 2019). The conclusion that is drawn by all the authors investigating fairness tools in ML from a socio-technical perspective is that there is a "disconnect" between the real-world situation of ML teams and the literature (Holstein & Vaughan, 2019; M. S. A. Lee & Singh, 2021; Veale et al., 2018).

One example is the call for a more holistic approach. Interviewees emphasise the importance of having a better communication between different teams that develop a ML model, such as data scientists and developers, and call for the development of tools that facilitate this coordination (Holstein & Vaughan, 2019). Another challenge expressed by practitioners is the amount of time that is required to monitor the fairness of the ML models and to work with the tools. The lack of sufficient time and human resources constitutes a major challenge in ensuring fairness (Holstein & Vaughan, 2019). Apart from the lack of time, the lack of knowledge in fairness compared to tools that require high-level expertise knowledge is identified as a challenge (M. S. A. Lee & Singh, 2021). Researchers also identify the scepticism toward a full automation of fairness processes due to quantifying challenges of fairness as a major disconnect between practice and academia. While the literature proposes a full automation with the support of fairness tools, interviewees voice concerns over this elimination of a human element (Holstein & Vaughan, 2019).

A smaller research strand within the papers that investigate fairness tools through a contingent, socio-technical lens is focused on the way that the use of fairness tools impacts the fairness understanding of developers that interact with them. While there are many papers that study the perception of fairness of lay people or users of the systems, very few studies focus on developers (Woodruff, Fox, Rousso-Schindler, & Warshaw, 2018).

Yan et al. (2020) examine the way tools enable developers to understand the sources of bias in datasets and draw meaningful conclusions. The specific toolkit used in the paper enables the developers to assess the fairness of the ML model and also leads them to reflect about systemic biases inherent in the system. The way the practitioners use the tool to explore different explanations and improve their understanding of fairness is analysed by using the sensemaking theory. The authors conclude that through interaction, fairness tools have the potential to support developers in their "sensemaking process" of assessing fair ML (Yan et al., 2020).

The work from Bellamy et al. (2019) explore how the specific toolkit AI Fairness 360 impacts the understanding of fairness of developers. The authors conclude that the toolkit shapes the knowledge of the users in terms of fairness, detection of unfairness and mitigation of bias (Bellamy et al., 2019).

2.4. Research Question

The literature review has revealed a research gap on the impact that fairness tools have on the understanding of fairness and the processes within a development team. Specifically, to the best of my knowledge, this aspect has not been investigated empirically to date. Given the potential consequences it might have for the implementation of fairness tools, it is imperative to understand the impact they can have when used in a development team. Therefore, this dissertation aims to answer the research question "How can fairness tools impact the understanding of fairness and the processes within a machine learning development team?". In order to answer this research question, a case study has been conducted. The theoretical framework used to analyse the findings of this case study as well as the details of this case study and the methodology used will be elaborated in the following parts.

3. Theoretical Framework

During the sense-making process of the empirical findings, two theories have been deployed in order to distil the full meaning of the data.

The first theory is Kallinikos' idea of the "regulative regime of technology". He states that technology can be

considered a "distinctive regulative regime" that forms the operations within a company in important ways and therefore influences social practice. Technology is described as an objectified system consisting of processes and forces that have an impact on tasks carried out in the environment it is implemented into. One aspect of that impact is the forming of perceptions, professional rules and routines. It is important to note that regulation in this sense is not defined as a rigid, constraining force, but rather used in a broader sense (Kallinikos, 2009, 2010).

The second theory that has been used is the theory of the co-construction of users and technologies by Oudshoorn and Pinch. The theory can be understood as an augmentation of the social construction of technological systems theory (SCOT). In the co-construction theory the main idea is that with their creative capacity, users shape technology in important ways throughout the whole development and implementation process. It examines the ways users impact technology by using, altering, resisting and reconfiguring it, and the ways technology impacts users by transforming and defining them (Oudshoorn & Pinch, 2003).

4. Research Design and Methodology

4.1. Case Study Method and Case Selection

In order to answer the research question and to create a fit between the research question and the data collection, I have decided to conduct a qualitative data collection. The objective of the data collection was to understand the impact of fairness tools on the understanding of fairness and on the processes within the development team. The purpose was to generate hypotheses about the interaction between developers and fairness tools in ML development teams as well as on their definition of fairness. Therefore, a qualitative data collection was the most suitable method for answering my research question.[1]

The specific research method that has been chosen is a single case study. Since the research question aimed at explaining the impact of fairness tools on development teams, a case study was the most suitable research method (Benbasat, Goldstein, & Mead, 1987). The role that fairness tools play and the impact they have needed to be understood in depth in order to answer the research question. Conducting a case study allowed me to obtain a "holistic and real-world perspective", which was needed to address the previously identified research gap. The case study as a data collection method matched the research question because of its focus on explaining the how and its focus on contemporary events (Yin, 1994). Furthermore, since the phenomenon that was to be studied is highly complex, it was not possible to remove it from its context and study it separately (Gerring, 2004).

The unit of analysis of the case study was the ML development team of a large, globally operating company. Due to legal reasons, it has been agreed with the interviewees that the name of the company will not be revealed in this work. The company is a multinational technology company that is headquartered in the U.S. and operates in approximately 170 countries. The development team that was used for the case study is embedded in the branch of the company that develops, sells and implements AI tools, solutions and applications. The team consults its clients with the aim to ensure a successful adoption of AI and an ethical, responsible use of the technology.

The tool that is being used by the development team is a fairness toolkit that integrates features for explainability and fairness. It aims at giving the developer a comprehensive view of the model and to evaluate the biases inherent in it. Therefore, it examines the data and retraces how the model reached a decision, and on basis of that evaluates whether the model is biased or not. If the tool finds that the model is biased, it alerts the developer, who can then decide how to proceed with this information. The tool also indicates ways to alter the model in order to eliminate the bias; these include protecting sensitive features in the training data so that these features will not be taken into account for the decision-making.

The specific development team has been selected as a unit of analysis for various reasons. Firstly, the operational range spans the whole development cycle of ML. This means that the team develops the models, selects the training data and trains the algorithm before selling it to their customers. Secondly, the team uses a fairness toolkit on a daily basis since many years. This enabled me to investigate and understand the impact that this toolkit has on the team. Since an open-source version of the toolkit is available online, I was able to study the tool itself in greater detail which prepared me for the interviews and enabled me to understand the details of the functionality. Lastly, the company has agreed to grant access to developers for interviews. Considering the high sensitivity of the topic and legal requirements from corporates, this proved to be difficult to find. Many development teams have been contacted for this case study; however, many declined a cooperation for different reasons. Some claimed not to experience any fairness issues in their AI applications, others are aware of the topic but do not employ any specific fairness tools. Others, as mentioned before, use fairness tools but were not able to disclose any information due to company guidelines.

4.2. Data Collection Strategy

The case study consisted of six semi-structured, in-depth interviews that have been conducted with members from the development team. The selection of interviewees was made based on availability, tenure and position within the team. It was important for this research to achieve diversity among the interviewees in terms of tenure and position since these factors probably influence the experience of working with the fairness tool. Given the limited timeframe of this study and the current size of the team, it was not feasible to interview

[1] Taken from 2021, "How can development teams effectively use fairness tools in order to ensure fairness in Machine Learning?", Research Proposal for MY401 Course, LSE

more developers. The aim of the interviews was to understand how different team members view the role and the impact of the fairness toolkit on their work and on their understanding of fairness.[2]

The interviews were conducted over Zoom and each lasted approximately 35-45 minutes. The interview guideline included questions on the interviewees' understanding of fairness, their experience of working with the tool and how the tool impacts their understanding of fairness and the processes within the team. The participants were informed about the research beforehand with an information sheet and gave their consent to the interview being recorded. After the interviews, they were fully transcribed by using the interview recordings.

Apart from the interviews, an analysis of documents detailing the fairness toolkit and its functionality was conducted.

With the aim to cross-validate certain findings, an additional interview has been conducted with an employee from a company that develops and sells fairness toolkits. This interview enabled me to receive a second perspective to some points mentioned in the interviews with the developers and contextualise them. This interview was also conducted via Zoom and was fully transcribed afterwards.

One of the advantages of interviews in this case is that they are suited for exploring perspectives, which was one of the aims for the data collection. Conducting interviews allowed the collection of relevant and focused data. One major weakness is that the interviewees might not be able to answer all the questions or that given the artificial setting, the answers don't always reflect the reality. This challenge was sought to be addressed with the diversity among the interviewees in terms of position and experience within the team.

4.3. Data Analysis Strategy

According to the qualitative data collection, the data has been analysed through a thematic analysis as proposed by Attride-Stirling (2001). In the first step, a coding framework was developed based on the literature review and the research question, and also on salient issues from within the data. Using this coding framework, the text has then been divided into separate sections (Attride-Stirling, 2001). These sections have then been read and analysed to identify common themes across the different interviews (Fereday & Muir-Cochrane, 2006). These themes have then been rearranged into a network of themes and clustered into basic themes, organising themes and global themes. In a next step, these thematic networks have then been further explored to detect the "underlying patterns". Consequently, the main themes within the networks have been summarised and interpreted with regard to the research question (Attride-Stirling, 2001). The thematic analysis has helped answering the research question by identifying the most important and salient aspects of the collected data and enabled me to interpret them. [3]

5. Findings

After having evaluated the interviews through a thematic analysis, three main themes emerged: Definition of Fairness, Limitations and Implementation Challenges of Fairness Tools, and the Impact of Fairness Tools. In the following, the findings corresponding to the themes will be presented.

5.1. Definition of Fairness

All the respondents agree on the difficulty of defining fairness in ML and struggle to pin the different notions they are aware of down to one unequivocable definition. They frequently equate it to the concepts of equal opportunity and equal outcome or associate it with transparency, explainability and accountability. All of the interviewees mention the importance of defining fairness differently for separate stakeholder groups.

One respondent also acknowledges that the working definition of fairness that is used in the industry will most likely differ from the one agreed on in academia.

> *"From how we work in the industry, is we have a what we call an enterprise design thinking approach for data and AI, where we look at the personas and user groups for that specific AI solution."*

A ML solution will then be judged in terms of its fairness by evaluating whether it is fair to all the different users and personas that use the system. This evaluation is conducted by ethics experts in order to then assess whether a solution is fair or not. The result of this evaluation process can then be considered as the definition of fairness.

The interviewees also emphasise that fairness needs to be defined every time depending on the context and the specific use case. In some cases, a certain form of bias might be acceptable or even wanted, which requires developers to define bias and fairness individually for every model. Therefore, as mentioned by the interviewees, the choice of the appropriate fairness definition for a specific context is not made by the fairness tool, but by the developers that then configure the tool accordingly.

It is also highlighted that the importance does not lie in choosing a particular concept of fairness, but in properly defining, understanding and operationalising one notion intentionally and transparently. For every use case, it is imperative to reveal the model of fairness that underpins the development of the model.

[2] Taken from 2021, "How can development teams effectively use fairness tools in order to ensure fairness in Machine Learning?", Research Proposal for MY401 Course, LSE

[3] Taken from 2021, "How can development teams effectively use fairness tools in order to ensure fairness in Machine Learning?", Research Proposal for MY401 Course, LSE

5.2. Limitations and Implementation Challenges of Fairness Tools

One major theme present in the interviews are the limitations and implementation challenges evolving around the use of fairness tools. These can be clustered into the three sub-themes Holistic Approach, Accuracy Trade-Offs, and Implementation Challenges.

5.2.1. Holistic Approach

Interviewees emphasise the importance of having a holistic approach when talking about fairness in ML. Instead of only focussing on the part of the development team that is concerned with the building of the model, it is fundamental to take into account the whole process of development, which includes the ML team as well as other teams that share responsibility for the development. Given the dependencies that frequently exist between various different processes, the necessity arises to monitor the whole development process to detect "opportunities for debiasing" and also weak points that can render the fairness efforts useless.

One interviewee highlighted in particular the importance of ensuring a tight connection between the business and the development area. Since the business problems that are to be solved with a ML solution and the requirements for the ML model itself are created within the business team and then transferred to the development team, it is imperative to involve the business department in the fairness efforts. This strong connection between business and development is described as success-defining, however, it is often not enabled by the use of fairness tools, as reported by interviewees.

5.2.2. Accuracy Trade-Offs

In terms of accuracy and the overall performance of a model that might be reduced when fairness tools are applied on it, all of the interviewees reject the idea of always prioritising one of the two parameters and highlight the need to consider the respective context. The process of developing a model always involves a "tweaking" of the individual parameters to achieve the desired performance and accuracy rate. The factors that guide, or should guide, this process are industry standards, the company and the corporate culture as well as the nature of the ML solution that shall be deployed, such as the objective and the data that is available.

It became evident during the interviews that the difficulty hereby mainly lies in the lack of a clear definition of the terms fairness and accuracy.

> *"What is accurate then? Is it your historical data that is accurate? Or any assumption on how the future should look like? The golden standard that you have created artificially? Is that accurate?"*

This quote clearly illustrates this challenge in labelling a model as fair or accurate. The respondent concludes that a fairness tool is not able to deal with this problematic. Instead, best practices, design methods and an ethics board need to be implemented in order to evaluate the specific context and decide which factors should be prioritised over others. This means that the decision of the trade-off is not left to the tool, but negotiated beforehand and then the tool is configured accordingly.

One of the respondents highlights the importance of moving beyond the dichotomy of accuracy and fairness. When bias is detected in a model, and it is decided to remove this bias, this has to be done regardless of the accuracy, since an unfair model is not accurate either.

> *"There is this habit of some people in the field to become very focussed on improving this one number, sometimes accuracy, and then they become tunnel-visioned on that. I don't think that is a very good way to develop models, fairness-aside".*

This quote illustrates that the formulation of a fairness-accuracy trade-off can be obsolete since accuracy is not the best and only indicator for a good model and if fairness shall be added to the set of requirements of a model, it should be treated equal to the other requirements, such as accuracy.

5.2.3. Implementation Challenges

During the interviews, various challenges to the implementation of fairness tools in everyday work practices have been stated. One of them is the lack of time in development that can be dedicated to fairness efforts. This is exacerbated by the fact that in many cases, when a tool indicates that a model is biased, further data needs to be collected in order to eliminate the bias from the model. This lack of time presents a hurdle for development teams in the implementation of fairness tools. This can be described as a fairness-cost trade-off, cost in that sense meaning both money and time and it advantages big companies compared to smaller companies which typically have less resources at their disposal.

Another major challenge is the lack of skilled talent. Properly understanding the tools and their functionality requires high-level math skills which are, according to the interviewees, not always represented in a development team. This is connected with the issues that companies who are about to develop a ML model often experience difficulties in finding developers and data scientists. Once they have the necessary resources, they want to start developing their model immediately without taking fairness into account. Also the fact that some toolkits are aligned with a traditional waterfall-development process makes it hard to implement them in today's iterative and agile development environments.

5.3. Impact of Fairness Tools

The theme impact of fairness tools consists of the three sub-themes Understanding of Fairness, Processes and Practices, and Fairness Automation.

5.3.1. Understanding of Fairness

The respondents state that the use of fairness tools impacts and shapes the understanding of fairness within a development team. The use of a certain tool shows the team

what is fair by indicating certain models as biased and others not.

> *"Yeah, it contributes [to the understanding of fairness], because it just visibly shows how the data is biased. And sometimes we do not even recognise as human beings that our data that we have (..) is biased."*

The interviewees explain how, by making visible the hidden bias in the data, the model contributes to a change in the understanding of fairness because of its functionality. One respondent states that sometimes, the results of the fairness tools are surprising because the development team did not expect any bias to occur in the respective model. If the model then indicates an instance of unfairness, it shows to the developers that their assessment was not correct and adds to their definition and understanding of which models are fair and which are not. Thus, the interviewees agree on the fact that the technology contributes to the awareness of fairness among the people who are working with it.

5.3.2. Processes and Practices

Apart from shaping the understanding of fairness, tools deployed in development teams to ensure the fairness of ML also limit the scope of actions of developers and the possibilities in their work, as reported by the interviewees. By giving an alert when the model seems to be biased or behaves in an unwanted way, the model gives the human the sign to pause the development process and take one step back to evaluate and, if needed, correct the bias. By giving the developers this insight, the tool enables the team to engage in a discussion and pivot the approach. This means that the developer needs to evaluate the bias instance evaluated by the model and, if it is found to be a bias that should be eliminated, take action to ensure the fairness of the model. Thus, the interviewees view the tool as limiting their possibilities of work – however not in a negative sense, but rather understood as additional support.

One respondent draws an analogy between fairness tools and compliance guidelines within a company. Given that supervising and controlling nature of fairness tools, some interviewees describe these as regulating the ML development process within a team.

> *"If you are (. . .) introducing a fairness tool, then at one point, the people or the data scientist is not able to do it, like he did before. And then of course, it's something like a regulation point."*

The fairness tool does not automatically change the development process or force actions to be taken, it rather makes the development team reflect about the model and decide whether the biases indicated by the tool shall be eliminated or not. This means that the fairness tool does not solve the bias issue automatically and independently, but requires the interaction from the developer side.

Interviewees also report that the usage of a tool impacts the development process by introducing additional checkpoints. They create a certain structure around the development and testing process and introduce new steps to it, so that the development process will be different than it was before the implementation of the tool.

5.3.3. Fairness Automation

The interviewees agree on the fact that fairness tools should not automatically eliminate bias in the future, instead of alerting a human to then make a decision. Some state that, given the sensitive and highly contingent context of ML systems, the process of eliminating bias will and should not be automated in the future. Therefore, human oversight will always be needed to make sense of the results given by a fairness tool. Others claim that the tools are still in a nascent state of development which is why they are currently not able to perform actions automatically. They agree on the importance of augmenting the regulative impact of fairness tools with human controls. It is emphasised that a fairness tool should not be used as a standalone indication of bias, but rather in addition to human judgement.

In contrast, one employee of a company which develops and sells fairness tools has been interviewed in order to cross-validate the findings and he emphasises that the tools are planned to automatically conduct the process of bias elimination in the future and thus replace the human element in the loop.

6. Discussion

In the following, it will be reflected on the empirical results, connecting them to the findings from the literature review and the research question on how fairness tools can impact the understanding of fairness and the processes within a machine learning development team. Consequently, the contributions to research and implications for practice will be presented.

6.1. Reductionist and Contingent Approaches to Fairness

The findings from the interviews show the impossibility of pinning fairness down to one single definition and that fairness needs to be treated as a contingent, context-specific concept. This finding falls in line with the idea of contingent approaches to fairness (Holstein & Vaughan, 2019; M. S. A. Lee et al., 2021; Srivastava et al., 2019). Similar to these ideas, the findings reject the idea of following a reductionist approach to fairness. Instead of attempting to analyse the trade-offs between several definitions of fairness and combine them into one simplified notion (Corbett-Davies et al., 2017; Kleinberg et al., 2017), the findings support the idea of defining fairness according to the specific use case and context. Conclusively, the empirical findings support the contingent lens on the definition of fairness and reject reductionist approaches.

6.2. Rejection of Deterministic / Technical-Rational Approaches to Fairness Tools

Analysing the empirical findings, it can be noted that they align with the socio-technical, contingent perspective, rejecting the deterministic view on fairness tools. Instead of viewing these tools as fixed solutions that can be implemented easily and directly lead to fairness, the interview results confirm the challenges encountered when implementing them and the complexity involved in the interaction between developer and tool.

The implementation challenges mentioned in the literature have been mostly confirmed by the empirical findings, namely the lack of time in development teams that can be dedicated to fairness efforts, the lack of human resources and skilled talent and the difficulties in collecting additional data Holstein and Vaughan (2019). One additional challenge that has not been mentioned in the literature to the best of my knowledge is the misfit between fairness toolkits that are sometimes designed for a waterfall development process and today's agile, iterative processes.

These implementation challenges described by the interviewees serve as a support to the claim made in the contingent approach that tools need to take into account the socio-technical environment of the ML model in order to be successfully implemented. It also shows that the claim from the deterministic perspective that tools do not face any difficulties when being implemented in a team (Berk et al., 2021; Hajian et al., 2016) does not hold true.

In general, the empirical findings emphasise the importance of considering the context of ML systems. Examples are the trade-off between accuracy and fairness which needs to be negotiated according to the respective context, as well as the holistic approach that is requested in order to consider the whole development context of a system. This falls in line with the socio-technical approaches in the literature (Holstein & Vaughan, 2019; M. S. A. Lee & Singh, 2021) which argue that fairness toolkits are too focused on the technical aspects and are neglecting the context of the respective systems.

6.3. Fairness Tools as Regulative Regimes

Concerning the impact that fairness tools have in development teams, the empirical findings reveal that fairness tools impact and shape the understanding of fairness within a team in important ways. In this sense, the technology of fairness tools contributes to both the awareness and the understanding of fairness by indicating the developers instances of unfairness. Fairness tools can also impact the processes within a team by creating a structure around the development and testing process and introducing new steps and checkpoints. What makes these findings significant is that they can be seen as an enactment of Kallinikos' theory on technology as a regulative regime (Kallinikos, 2009).

By alerting biased decisions made by an algorithm and thus indicating the developer to modify the model to eliminate this bias, the tool clearly shows the developer instances of bias and thus shapes their understanding of fairness. The output of the technology determines the next steps for the developers and therefore forms the way they define and understand fairness. By limiting the scope of actions for developers, the tools regulate the development process and enable the team to engage in a discussion and pivot their approach. Thus, the technology of fairness tools can be understood as a "regulating practice" that shapes the operations and development processes within a team and governs the social practice by raising the awareness for fairness and shaping its understanding (Kallinikos, 2010). In future investigations, it would have to be investigated in a comparative study how the implementation of different tools in one development team impacts the understanding of fairness.

The application of Kallinikos' theory on the impact of fairness tools can be regarded as an augmentation to the current, limited research on the impact of fairness tools on the understanding of fairness of developers. The empirical findings confirm the ideas present in the literature on how fairness tools enable developers to better understand unfairness and fairness and shape their knowledge in terms of bias mitigation (Bellamy et al., 2019; Yan et al., 2020). The sensemaking theory used in the existing literature cannot be confirmed since the methodological setting used in this study did not allow for a close observation of sensemaking loops that are supported by fairness tools. Thus, the analysis of the impact of fairness tools through the lens of Kallinikos' theory reveals the regulating impact these tools have on developers interacting with them.

Another aspect of the role that fairness tools play in development teams is the degree of automation and independence of these tools. Here, the empirical findings reflect the results from the literature review. While the interviewee who works in a company that develops fairness tools emphasises that these tools will function automatically in the future, without including a human in the loop, the interviewees are more concerned about a full automation. Given the sensitive and context-dependent area, they highlight the importance of human control, and this scepticism is also reflected in the literature (Holstein & Vaughan, 2019). This raises questions about the future role of fairness tools in development teams and their desired degree of independence and automation. This will need to be investigated in more detail in future research.

6.4. Co-Construction of Developers and Fairness Tools

The empirical findings have revealed that, apart from the fairness tool impacting the understanding of fairness of the developer and the processes within the team, the tools are also shaped by the choices made by the developers. This becomes evident at the choices developers make about the definition of fairness and desired fairness-accuracy trade-offs depending on the context. Instead of the tool having these choices already pre-programmed, the developers shape them by making choices and configuring the tools accordingly. This rejects the deterministic view on fairness tools that assumes that choices around fairness, such as the definition of fairness and fairness-accuracy trade-offs can be simplified and built into a fairness tool.

Applying Oudshoorn and Pinch's theory on the co-construction of users and technologies, this process of mutual shaping between developer and fairness tool can be understood as a process of co-construction. The character of fairness tools as regulative regimes can then be viewed as an embedded part of this process. By acting as a regulative regime and shaping the understanding of fairness and development processes, the fairness tools transform and influence the developers. Simultaneously, developers shape fairness tools by making choices and configuring the tools accordingly.

This finding also supports the socio-technical perspective and challenges the deterministic perspective on fairness tools because it confirms that tools cannot be simply implemented into a development team and expected to yield in more fairness. Fairness tools are no fixed objects that function independently from the development team and the socio-technical environment. Instead, they are actively shaped by the developers through choices and then configured accordingly; and the developers and development processes are also shaped and influenced by the fairness tools. Rather than focussing on the technical functionalities and shortcomings of fairness tools, this process of mutual shaping, of co-construction between user and technology, needs to be brought into focus.

6.5. Contributions to Theory and Implications for Practice

These findings contribute to the existing literature by explaining the impact of tools on the understanding of fairness with the use of the theory on regulative regimes which challenges the deterministic, technical-rational view on fairness tools which assumes that choices on fairness are built into a tool that then directly eliminates bias in ML and leads to fairness. Simultaneously, it supports the contingent, socio-technical perspective by showing the importance of the context and environment of ML systems.

Furthermore, the findings contribute to the understanding of the interaction between developer and fairness tool by describing the mutual shaping of developer and fairness tool as a co-construction process, of which the regulative character of fairness tools can be understood as one part.

The empirical findings also entail important implications for the practice of development teams. Having analysed how developers construct fairness tools and vice versa, development teams should make an increased effort to educate all members of the team on fairness in ML. Since the choices, both conscious and subconscious, taken by developers, influence the functionality of fairness tools, they should understand for instance what consequences different definitions of fairness have and how accuracy and fairness can be traded off against each other, depending on the context. Apart from that, companies that implement fairness tools in their teams should be aware of this co-construction process and actively shape it. The findings also show that companies cannot simply implement fairness tools and expect them to eliminate algorithmic bias from their models. Instead, they need to actively construct the functionality of the tool.

For companies that develop and sell fairness toolkits, the findings signify that they should educate their clients in more detail about the impact of fairness tools and how they are co-constructed and influenced by the choices made by developers.

7. Conclusion

The aim of this research was to investigate the impact that fairness tools can have on developers and on the processes within a development team.

The literature review has revealed the tensions between the deterministic, technical-rational and the socio-technical, contingent approach to fairness tools. It has been analysed that the existing, but very limited literature on the impact of fairness tools argues that fairness tools impact the knowledge of developers on fairness tools. The empirical findings from the case study conducted with the ML development team has confirmed this and revealed the extent of this impact. The application of Kallinikos' theory on the regulative regime of technology brought to light the character of fairness tools as regulative regimes in development teams.

The empirical findings on the impact of fairness tools on developers, against the backdrop of the current literature, led to a novel understanding of the relation between developer and tool. The use of Oudshoorn and Pinch's theory enabled an understanding of the interaction between the developer and the technology, the fairness tool, as a process of co-construction in which both elements shape and construct the other.

With these two main conclusions, the research question can be answered as follows. By shaping the understanding of fairness and the processes within a development team, fairness tools act as a regulative force. This process of regulating has to be understood in the wider context of a co-construction process between the technology, in this case the fairness tool, and the user, who is the ML developer in this case.

These findings contribute to the limited literature on the impact of fairness tools by confirming and extending the impact of tools on the knowledge of developers and offering a theoretical framework to analyse the impact of this technology. It also adds to the contingent, socio-technical perspective on fairness tools by highlighting the importance of the context for the functionality of the tools. Lastly, it challenges the deterministic perspective present in the current literature by showing how the impact of fairness tools is not determined by their technical characteristics, but by the way they are constructed through conscious choices of the developers.

7.1. Limitations & Further Studies

Although this research contributes to both research and practice in significant ways, it also exhibits important limitations.

Firstly, it is imperative to note that the limited scope of this work constrains it to the most relevant and significant findings that answer the research question. In future work,

the impact of fairness tools on developers and the concrete processes of co-construction will have to be studied further and in more detail.

Another limitation is the possibly constrained eagerness of the respondents to talk openly in the interviews. Given the high sensitivity of the topic, it is possible that certain thoughts have not been expressed freely. This could be overcome in future works by the conduction of an anonymised study, like it is also used by Binns (2018).

Regarding the general reliability of the data, it can also be noted that through an observation of the development team over a longer period of time, richer data could have been collected concerning the interaction with the tool. Due to the COVID-19 Pandemic, the majority of employees is currently still working from home, which is why this has not been possible and should be conducted in future works on the topic.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–18).

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning* (pp. 60–69).

Attride-Stirling, J. (2001). Thematic networks: An analytic tool for qualitative research. *Qualitative Research*, *1*(3), 385–405.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... Mojsiloviç, A. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4–1.

Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The case research strategy in studies of information systems. *MIS Quarterly*, 369–386.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, *50*(1), 3–44.

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. *Conference on Fairness, Accountability and Transparency*, 149–159.

Calders, T., & Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277–292.

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 3995–4004).

Chen, I., Johansson, F. D., & Sontag, D. (2018). Why Is My Classifier Discriminatory? *Advances in Neural Information Processing Systems*(31), 3543–3554.

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

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797–806).

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259–268).

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, *5*(1), 80–92.

Gade, K., Geyik, S. C., Kenthapadi, K., Mithal, V., & Taly, A. (2019). Explainable AI in Industry. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3203–3204.

Garcia, M. (2016). Racist in the Machine: The Disturbing Implications of Algorithmic Bias. *World Policy Journal*, *33*(4), 111–117.

Gerring, J. (2004). What is a case study and what is it good for? *American Political Science Review*, *98*(2), 341–354.

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2125–2126).

Holstein, K., & Vaughan, J. W. (2019). Opportunities for Machine Learning Research to Support Fairness in Industry Practice. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16.

IBM Developer. (2021). *The AI 360 Toolkit: AI models explained.* https://developer.ibm.com/articles/the-ai-360-toolkit-ai-models-explained/.

Kallinikos, J. (2009). The regulative regime of technology. In *ICT and innovation in the public sector* (pp. 66–87). Springer.

Kallinikos, J. (2010). *Governing through technology: Information artefacts and social practice*. Springer.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining* (pp. 869–874). IEEE.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2019). An Empirical Study of Rich Subgroup Fairness for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 100–109). ACM.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. *Proceedings of Innovations in Theoretical Computer Science (ITCS)*.

Lee, M. K., & Baykal, S. (2017). Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1035–1048). ACM.

Lee, M. S. A., Floridi, L., & Singh, J. (2021). Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI and Ethics*, 1–16.

Lee, M. S. A., & Singh, J. (2021). The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–13).

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy & Technology*, *31*(4), 611–627.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. In *International Conference on Machine Learning* (pp. 3150–3158). PMLR.

Mattu, J. L., Kirchner, L., & Surya, J. A. (2016). Machine Bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Menon, A. K., & Williamson, R. C. (2018). The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency* (pp. 107–118). PMLR.

Oudshoorn, N., & Pinch, T. (2003). How Users Matter: The Co-Construction of Users and Technologies. *The MIT Press*.

Pasquale, F. (2015). *The black box society*. Harvard University Press.

Richardson, B., Garcia-Gathright, J., Way, S. F., Thom, J., & Cramer, H. (2021). Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–13).

Rossi, F. (2018). Building trust in artificial intelligence. *Journal of International Affairs*, *72*(1), 127–134.

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., & Liu, Y. (2019). How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 99–106).

Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2459–2468).

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14.

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–14).

Yan, J. N., Gu, Z., Lin, H., & Rzeszotarski, J. M. (2020). Silva: Interactively Assessing Machine Learning Fairness Using Causality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13).

Yin, R. K. (1994). Case study research: Design and methods, applied social research. *Methods Series*, *5*.

Zhang, X.-D. (2020). *A Matrix Algebra Approach to Artificial Intelligence*. Springer Singapore.