Spring 7-10-2024

# Operationalizing Algorithmic Fairness: Empirical Study and Framework Proposal

Fredrik Wang
*Norwegian University of Science and Technology*, fredrikbw@outlook.com

Ilias Pappas
*Norwegian University of Science and Technology University of Agder*, ilpappas@ntnu.no

Polyxeni Vassilakopoulou
*University of Agder*, polyxenv@uia.no

# OPERATIONALIZING ALGORITHMIC FAIRNESS: EMPIRICAL STUDY AND FRAMEWORK PROPOSAL

**Fredrik Wang**

*fredrikbw@outlook.com*

*Norwegian University of Science and Technology*

**Ilias O. Pappas**

*ilpappas@ntnu.no*

*Norwegian University of Science and Technology*

*University of Agder*

**Polyxeni Vassilakopoulou**

*polyxenv@uia.no*

*University of Agder*

**Abstract**

*This paper explores how organizations approach and operationalize algorithmic fairness in practice. Through semi-structured interviews with practitioners from organizations in Norway, insights were gained around their algorithmic fairness approaches and implementations. A thematic analysis revealed key considerations around starting early, law and regulations, the business value of fairness, challenges of identifying intersectional bias and technical solutions for pursuing and continuously monitoring fairness. An Extended Sociotechnical Framework for Algorithmic Fairness is proposed to help organizations address algorithmic fairness as a multifaceted issue. The framework categorizes general and case-specific factors across technical and social domains to provide structure while emphasizing context-specificity. It suggests harmonizing technical and social components to support practitioners navigating this complex area. The study provides empirical evidence of real-world fairness operationalization. This is a critical issue as the use of artificial intelligence technologies becomes more widespread, with the potential to introduce discriminatory biases against individuals or groups. Algorithmic fairness is key for upholding equity and preventing harm to vulnerable people.*

**Keywords**: Algorithmic Fairness, Sociotechnical Systems, Responsible AI

## 1.0 Introduction

The use of different types of Artificial Intelligence (AI) technologies including machine learning is today more widespread than even. Ensuring that AI systems do not disproportionately favour or harm individuals or groups is critical. Discoveries of unfair algorithmic outcomes make headlines (Constantaras et al. 2023; Asher-Schapiro 2020; Angwin et al. 2016) and organisations that work to ensure AI fairness

can position themselves as trustworthy partners (Shollo and Vassilakopoulou, 2024). Algorithmic fairness means that algorithmic systems treat individuals and groups equitably, without discrimination or bias (Binns, 2018). The concept received increased attention from the research community, but mostly in terms of developing statistical definitions and mathematical approaches for identifying and mitigating bias (Chouldechova 2017). Statistical notions of fairness are easy to measure, however, comprehensive operationalisations of the fairness concept require domain-specific expert input and opinion (Verma and Rubin, 2018). Hence, research beyond statistical formulations is needed to gain insights about algorithmic fairness in different application domains.

The objective of this paper is to develop a better understanding of how organizations approach algorithmic fairness, from initial discussions to deployed solutions. Specifically, the paper aims to answer the following research question: How do organizations approach and implement algorithmic fairness in practice? To answer this question, we collected and analysed empirical data collected by interviewing nine participants from eight different organizations. The insights are consolidated in The Extended Sociotechnical Framework for Algorithmic Fairness and recommendations for future work.

The remainder of this paper is structured as follows. Section 2 provides an overview of relevant background literature. Section 3 describes the method followed for data collection and analysis. Section 4 presents the main findings regarding organizations' fairness approaches and implementation experiences. Section 5 provides a discussion of these findings and introduces our proposed Extended Sociotechnical Framework for Algorithmic Fairness. Finally, Section 6 concludes the paper and outlines limitations and directions for future work.

## 2.0 Background

AI systems affect many aspects of everyday life especially through algorithmic decision support (Adensamer, Gsenger, and Klausner 2021; Holten Møller, Shklovski, and Hildebrandt 2020). Such algorithmic support is used for instance in hiring (Langenkamp, Costa, and Cheung 2020), loan assessments (Sheikh, Goel, and Kumar 2020) and rankings used for recommender systems (Biega, Gummadi, and Weikum. 2018). However, studies conducted by researchers and regulators found algorithms to

reflect and even amplify historical bias, and also potentially introduce biases of their own accord (Mehrabi et al. 2022). Algorithms containing bias can unfairly discriminate against minorities or discriminate on the basis of gender, age and language. Algorithmic unfairness has been identified across a wide range of fields including welfare (Constantaras et al. 2023), healthcare (Obermeyer et al. 2019), judiciary services (Angwin et al. 2016), and education (Asher-Schapiro 2020). These serve as constant reminders that the use of AI may entail discrimination risks.

Prior research with AI experts showed that fairness is the most challenging principle for organizations to implement when it comes to responsible AI (Akbari Ghatar et al. 2023). The key practical challenges around fairness relate to the fact that it is highly context-dependent and also to the need for ongoing monitoring as AI models can change behaviour over time. Technical fixes like debiasing algorithms are important, but the "human problem" of what fairness means in a given context must also be continuously evaluated. The concept of fairness is differently used across disciplines: philosophers consider fairness in terms of morality, social scientists often consider fairness in light of social relationships, power dynamics, institutions, and markets, quantitative fields have studied questions of fairness as pure mathematical problems (Mulligan, et al. 2019). For more than 20 years researchers have been studying bias in computer systems and pointing to the risks of biased systems (Friedman and Nissenbaum 1996).

The term algorithmic fairness refers to technological solutions designed to prevent systematic advantages or disadvantages to certain groups. In other words, algorithmic fairness means that algorithmic systems treat individuals and groups equitably, without discrimination or bias (Binns, 2018). From a technical standpoint, it is possible to introduce mathematical measurments of bias that can be used to develop computational approaches to minimize discriminatory outputs in machine learning against specific groups (Chouldechova 2017). However, as fairness is not merely a technical concept it has to be approached from a sociotechnical standpoint (Dolata, Feuerriegel, and Schwabe 2022).

## 3.0 Research Method

In order to collect data, nine semi-structured interviews were performed with people from eight different organizations. Semi-structured interviews allow for the discovery

of unforeseen information as they accommodate interviewees' decisions about what is important and relevant to talk about (Schultze and Avital 2011). The interviews were guided by an interview guide which was structured into general questions first, such as background, role, and fairness impressions, and then asking about the approach followed in the specific organization. It was also sometimes beneficial to ask follow-up questions that were not in the guide as issues emerged from the participant's answers. Hence, the interviewer allowed for development of the plot (Myers and Newman 2007) during each interview based on the input of the interviewees.

The interviews were performed between February and April 2023, were transcribed and recorded, and all participants signed consent forms. Interviews were conducted over Microsoft Teams with video and audio. The organizations were selected on the basis of their experience in developing and deploying AI solutions and we aimed to cover different industries and also both public and private organizations. Table 1 provides an overview of the interviews performed.

Participants were identified in three ways. One way was by contacting those who had participated in public conferences where algorithmic fairness was a topic, or similarly had published articles or academic papers where algorithmic fairness was a topic or subtopic. The second way was using the authors´ network. The third way was using LinkedIn to search for topics like 'algorithmic fairness' and similar, to find people who worked with machine learning and AI in companies where it would be logical for fairness to be a part of their projects. When participants were recruited, they were given some instructions about what to expect the interview to be about. In this way, they would have some time to think about their views regarding the topic. Sending information about topics and questions to allow the interviewee to prepare can have a positive effect (Oates 2005). Another benefit of this is that it can alleviate some of the pressure from interviewees.

| IDs | Company | Role | Organization Size | Duration |
|---|---|---|---|---|
| R1 | State-owned enterprise | Data Scientist | >5000 | 55 min |
| R2 | State-owned enterprise | Lawyer | >5000 | 50 min |
| R3 | State-owned enterprise | Data Scientist | 500 | 50 min |
| R4 | Private Research | Research Director | 100 | 50 min |
| R5 | Insurance | Director | 1000 to 5000 | 50 min |
| R6 | Private Corporation | Data Scientist | 500 | 45 min |
| R7 | State-funded enterprise | Senior Advisor | 100 | 55 min |
| R8 | Private Corporation Company | Lawyer | 100 | 55 min |
| R9 | State-owned enterprise | Technologist | 100 | 40 min |

**Table 1. Overview of Interviews**

Thematic analysis was adopted to analyse the qualitative data collected (Oates 2005). Thematic analysis is a method for identifying, analysing, organizing, describing, and reporting themes found within a data set. All interview transcripts were thoroughly analysed to generate initial codes summarizing key concepts and patterns found. The codes were consolidated further to form overarching themes representing important patterns within the data in relation to the research topic and objectives. This thematic analysis process allowed for a rich, detailed, and nuanced interpretation of the perspectives and experiences described by participants.

# 4.0 Findings

## 4.1 Starting Early with Fairness Considerations

One aspect of working with algorithmic fairness is that you can't simply begin considering it when your model is already deployed and affecting people. Starting early with fairness is essential for success. One of the participants stated:

> "*When working with fairness you need to start early, not just because of legal considerations, but also because it affects the design and product development of the solution.*"

Starting early also has the benefit of reducing the need for costly and time-consuming revisions further down the line of product development. Retroactively integrating fairness considerations in an AI system that is operational can be complex and costly, and it's generally more cost-effective to prevent unfairness in the first place rather than to face the aftermath of algorithmic bias.

## 4.2 The Role of Law and Regulations for Achieving Fairness

It is normally agreed upon that fairness is something one wants to achieve, the question is how it should be achieved. The interviews revealed that there are legal considerations related to achieving fairness that need to be taken into account. One of the interviewees explained how the laws have significant impact in achieving fairness:

> "*You can make a model and test its performance and fairness, but legislators can decide that certain groups should be prioritized over other groups, and then the model would have to be "unfair" first so that it complies with the law before it can be fair to other groups*."

Rapidly evolving legal and regulatory landscapes surrounding fairness can cause uncertainty for companies. Understanding and complying with laws and regulations related to fairness can be challenging. Furthermore, achieving this understanding can be expensive if the company doesn't already have these resources.

### 4.3 The Business Value of Fairness

An aspect that can serve as a motivation in several contexts, is fairness as a selling point, with fairness adding business value. Fair algorithms enhance brand reputation and foster trust among customers and stakeholders. In an era where customers increasingly value ethical business practices, companies demonstrating a commitment to fairness can differentiate themselves in the market. Investing in algorithmic fairness is not just a matter of ethics and compliance, but also a sound business strategy that drives long-term value and competitiveness as explained by one of the participants:

> "*We believe that implementing fairness, along with transparency and responsibility, will drive business value, and those who are best at it will have a competitive advantage. ... fairness will become a selling proposition*."

Fair algorithms can also lead to better and more inclusive decision-making. They can uncover and correct biases that may have traditionally limited business opportunities, such as in hiring, lending, or marketing. This leads to a more diverse and inclusive customer base and workforce, which are known to improve creativity, innovation, and profitability. Lastly, fairness can reduce the risk of costly litigation and penalties associated with unfair or discriminatory practices.

### 4.4 Challenges of Identifying Intersectional Bias

Bias may not always be so easy to spot, proxies can make it difficult to identify bias. Similarly, discrimination that only happens at intersectionality makes it difficult to understand when unfair treatment is happening. Intersectionality refers to the way different aspects of a person's identity (such as gender, race, sexual orientation, etc.) combine and overlap to expose them to various forms of discrimination or unfair treatment. Bias can occur when multiple aspects of a person's identity intersect, such as discrimination against women with immigrant background but not necessarily men with immigrant background or women on their own. This type of intersectional discrimination may be difficult to identify because the bias is not evident when only examining one aspect of identity, such as gender or background alone. One would need to analyze how different personal attributes combine before the unfair treatment resulting from their intersection is detectable. So intersectional discrimination makes

it more challenging to pinpoint precisely when and how unfair treatment is taking place within a system compared to bias along single identity dimensions. Several participants pointed out that there is a lack of systematic methods for discovering bias and unfairness, and discrimination happening at the intersectionality of attributes is an example of bias that won't be discovered easily.

### 4.5 Technical Solutions for Ensuring and Continuously Monitoring Fairness

Participants stated that with today's toolkits, the technical aspect is a very small part of implementing algorithmic fairness. For instance:

> "*The technical implementation is a small part, the tools, and frameworks support you to check that your algorithm is implemented correctly and saves you from a lot of troubleshooting. It's a small part, but it's reassuring to have it in place*."

For classification and regression problems one can use techniques such as feature importance to see what attributes the model utilizes the most in its prediction. Through these techniques, practitioners can identify potential biases in their model. One of the participants explained a project where they revealed bias by looking at the feature importance of the model:

> "*By using feature importance methods we were able to see the model being discriminatory towards gender, and pointed out that this unfairness should be looked into even though the project is in an early phase*."

Once fairness metrics are determined it's important to continuously monitor the system against this. One of the participants stated the following:

> *"... the AI must be checked against this limit continuously. This is, for example, because the composition of the group of people the AI is used on can change, or the algorithm can become biased over time if it learns from and systematizes biases gradually.*"

## 5.0 Discussion

The research on algorithmic fairness has mostly been concerned with statistically defining fairness and then proposing methods and techniques to mitigate undesirable biases, in relation to these definitions (Agarwal et al. 2018). Whilst practitioners to some degree were also concerned with implementing statistical metrics, the overall takeaway from the interviews is that the most difficult part of algorithmic fairness is

to decide what constitutes fairness in each specific context. This requires domain knowledge and also, understanding of regulatory provisions.

Evaluating the fairness of an AI model requires a definition of fairness. Thus, understanding the context and assessing the impact of the system is pivotal algorithmic fairness. Relying on intuition for discovering unfairness is a risky strategy, but is often the chosen strategy, due to the lack of support to address the issue. A study by Holstein and colleagues also found that most industry practitioners rely on their intuitions, even though these were often found to be wrong (Holstein et al. 2019).

Data quality and sufficiency is a key challenge because data may contain historical bias which an AI model trained on these data will reflect (Roselli, et al. 2019). Similarly, data may be affected by the conscious or unconscious bias in the people who collect the data. Having enough data is also a challenge as unprivileged groups are often underrepresented. There are also cases where data isn't available, such as when all outcomes aren't observable. An example is getting a rejection for a loan, where one still doesn't know if the loan would have been paid back if it had been approved (Verma et al. 2020). Models that are biased against certain groups could continue to reject candidates from that group and we would never be able to have data on the outcomes if these decisions were not taken.

### 5.1 The Extended Sociotechnical Framework for Algorithmic Fairness

Based on the results from the thematic analysis and the literature, a framework for understanding how practitioners can advance toward algorithmic fairness has been created (Figure 1). The framework is expanding on the work of Sarker and colleagues (Sarker et al. 2019) aiming to a harmonization between technical and social components. The technical components involve things like developing mathematical definitions of fairness, implementing algorithmic mechanisms to mitigate bias, and assessing models for unfair outcomes. Meanwhile, the social components pertain to high-level issues like organizational policies, legal/regulatory landscapes, sociocultural biases, and stakeholder values. Rather than seeing these as separate concerns, harmonization aims to bridge the gap between the technical and social domains. The goal is to develop an integrated approach where the technical solutions account for relevant social factors, and social/policy decisions are informed by technical considerations. This harmonization of the technical and the social is key to operationalizing fairness in real-world applications.

The suggested framework is split into four main categories, consisting of General Technical Factors, Case-specific Technical Factors, General Social Factors, and Case-specific Social Factors. This structure distinguishes factors that have broad relevance across all organizations addressing algorithmic fairness versus those more tailored to individual situations. The classification makes the framework less overwhelming and easier to apply while also showing the context-dependent nature of algorithmic fairness initiatives, emphasizing that there is no one-size-fits-all solution to algorithmic fairness (Morse et al. 2021). By splitting factors in this way, users can identify baseline technical and social elements to address generally as well as those necessitating adaptation.

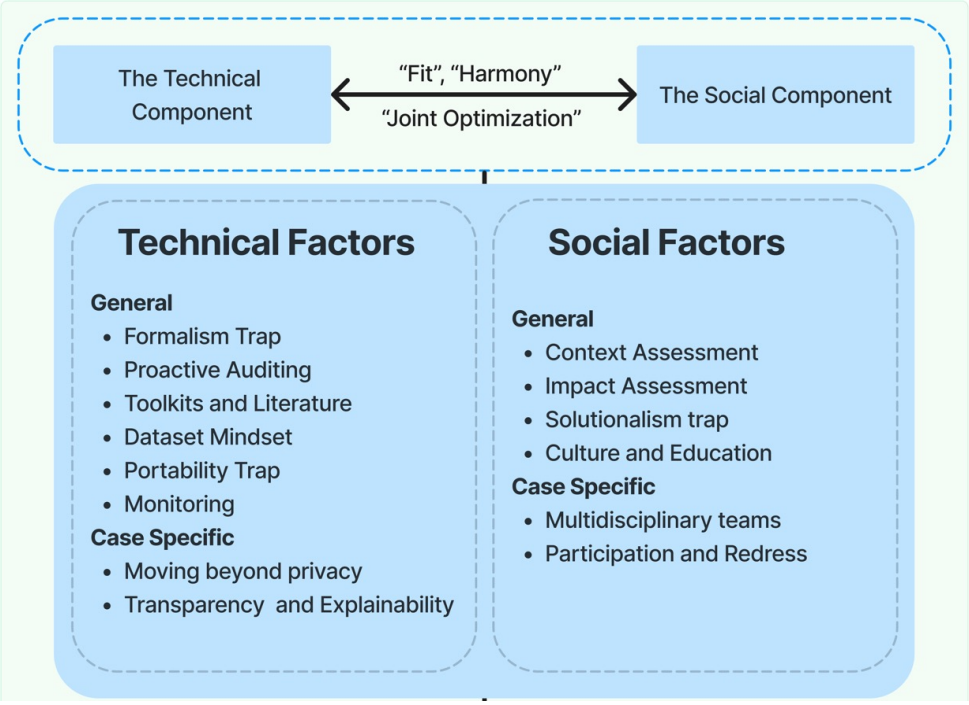


**Figure 1. Extended Sociotechnical Framework for Algorithmic Fairness.**

Factors such as performing *proactive auditing* in order to avoid bias from the start and having mechanisms in place to handle emerging bias as data and model parameters change are crucial. Using appropriate toolkits can help in properly implementing the technical part of the solution and ensuring that the outcomes are equitable. Recognizing that reusing algorithmic solutions that were designed for a specific context could lead to inaccuracies or cause harm can help prevent algorithmic systems from further marginalization and exclusion, and thus foster both inclusivity and diversity. Having a *dataset mindset* is an example of a crucial factor. Improving the quality of datasets is key for both better accuracy and fairness. Having a dataset that

better represents the real world can increase diversity. Staying up to date with technical solutions, such as the described toolkits is one way that companies can take a more structured and active approach to fairness.

In the social factors, improving *culture and education* about fairness is key. Similarly, performing an *impact assessment* can indicate who may be affected by algorithmic outcomes, and show that algorithmic systems can have significant effects on the life of individuals. An example of a case-specific social factor is: *Mechanisms for pooling knowledge across teams* so that one can develop the right solutions depending on the system and context. This factor is case-specific because it is only relevant for organizations that have multiple AI teams. It would not apply to a small company where sharing knowledge across teams is not an issue. The full list of factors included in the framework along with their descriptions is provided in Table 2 below.

| **Factor** | **Description** |
|---|---|
| Formalism Trap | Mathematical definitions eliminate the nuances of fairness. |
| Proactive Auditing | Aspire to implement fairness from the beginning, instead of mitigating unfairness later. |
| Toolkits and Literature | Using state-of-the-art toolkits for technical evaluation and implementation and staying updated on research. |
| Dataset Mindset | Ensuring that the data are complete and of good quality. |
| Portability trap | Recognizing that reusing algorithmic solutions, originally designed for a specific social context, could lead to misinterpretations, inaccuracies, or potentially cause harm when implemented in a different context. |
| Monitoring | Maintaining that the outcomes are fair and prevent bias and unintended consequences after initial development and deployment. |
| Moving beyond privacy | Understand that an AI system could respect privacy (by properly handling personal data) or be sheltered from sensitive attributes, but still be unfair (if it produces biased outcomes). |
| Transparency and Explainability | It is important to understand how we get the specific algorithmic outputs. Explainable AI techniques can help achieve this. |
| Context Assessment | Assessing the context of a system and how this affects how fairness is approached and defined, and who should be involved. |
| Impact Assessment | Assessing the impacts of an algorithm and potential negative outcomes necessitates understanding its social context and the varied notions of fairness within that system. |
| Solutionalism trap | Overlooking the possibility that the optimal solution may not always involve technology can lead to missteps. |
| Culture and Education | Develop a culture for fairness. Necessary for developing domain-specific guides, algorithms, metrics, ethical frameworks, and case studies. |
| Multidisciplinary teams. | Contribute to a comprehensive understanding of biases, ethics, and social implications in algorithmic systems. Foster critical thinking, challenge assumptions, and promote creative problem-solving, leading to robust and equitable solutions. |
| Participation and Redress | Affected individuals and communities should have the opportunity to participate in decision-making about algorithmic systems, and there should be mechanisms for redress if the algorithm causes harm. |

**Table 2. Factors in the Extended Sociotechnical Framework for Algorithmic Fairness**

## 6.0 Conclusion

Overall, this study aims to provide insights on real-world algorithmic fairness practices. The sociotechnical perspective taken acknowledges fairness as a multifaceted issue and contributes to algorithmic fairness research by providing an understanding of actual practices and experiences. Based on the insights from interviews with practitioners we suggest the comprehensive *Extended Sociotechnical Framework for Algorithmic Fairness*. The framework can help practitioners and organizations understand how they can approach algorithmic fairness and gain a better understanding of their own situation and context. For the research community, it provides a first step towards operationalizing algorithmic fairness. A key limitation of this study is the relatively low number of organizations that participated and the fact that they are all located in Norway. Factors in the proposed framework were also found in the literature from other countries, but further investigations in different contexts are certainly needed. Future research could also focus on one specific industry (e.g. healthcare) to create targeted frameworks accounting for particular contextual factors.

## References


Adensamer, Angelika, Rita Gsenger, and Lukas Daniel Klausner. (2021) " 'Computer Says No': Algorithmic Decision Support and Organisational Responsibility." Journal of Responsible Technology 7: 100014.

Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. (2018) "A Reductions Approach to Fair Classification." In International Conference on Machine Learning, 60–69.

Akbari Ghatar, Pouria, Ilias Pappas, and Polyxeni Vassilakopoulou. (2023) "Practices for Responsible AI: Findings from Interviews with Experts." In AMCIS 2023 Proceedings.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. (2016) "Machine Bias." .Propublica. Accessed May 26, 2022. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Asher-Schapiro, Avi. (2020) "Global Exam Grading Algorithm Under Fire for Suspected Bias." Reuters EverythingNews. Accessed May 10, 2022. https://www.reuters.com/article/us-global-tech-education-analysis-trfn-idUSKCN24M29L.

Biega, Asia J., Krishna P. Gummadi, and Gerhard Weikum. (2018) "Equity of Attention: Amortizing Individual Fairness in Rankings. " 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, 405-414.

Binns, R. (2018) “Fairness in machine learning: Lessons from political philosophy”. ACM Conference on Fairness, Accountability, and Transparency (FAccT 2018),

Chouldechova, Alexandra. (2017) “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” Big Data 5 (2): 153–63.

Constantaras, Eva, Gabriel Geiger, Justin-Casimir Braun, Dhruv Mehrotra, and Htet Aung. (2023) “Inside the Suspicion Machine.” Wired. Accessed March 26, 2023. https://www.wired.com/story/welfare-state-algorithms/.

Dolata, Mateusz, Stefan Feuerriegel, and Gerhard Schwabe. (2022) “A Sociotechnical View of Algorithmic Fairness.” Information Systems Journal 32 (4): 754–818.

Friedman, Batya, and Helen Nissenbaum. (1996) "Bias in computer systems." ACM Transactions on Information Systems, 14 (3): 330-347.

Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. (2019) “Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?” In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–16.

Holten Møller, Naja, Irina Shklovski, and Thomas T Hildebrandt. (2020) “Shifting Concepts of Value: Designing Algorithmic Decision-Support Systems for Public Services.” In Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society, 1–12.

Langenkamp, Max, Allan Costa, and Chris Cheung. (2020) “Hiring Fairly in the Age of Algorithms.” arXiv Preprint arXiv:2004.07132.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. (2022) “A Survey on Bias and Fairness in Machine Learning.” https://arxiv.org/abs/1908.09635.

Morse, Lily, Mike Horia M Teodorescu, Yazeed Awwad, and Gerald C Kane. (2021) “Do the Ends Justify the Means? Variation in the Distributive and Procedural Fairness of Machine Learning Algorithms.” Journal of Business Ethics, 1–13.

Mulligan, Deirdre., Joshua Kroll, Nitin Kohli, and Richmond Wong. (2019) "This thing called fairness: Disciplinary confusion realizing a value in technology." Proceedings of the ACM on Human-Computer Interaction. CSCW, 1-36.

Myers, Michael D., and Michael Newman. (2007) "The qualitative interview in IS research: Examining the craft." Information and organization. 17(1): 2-26.

Oates, Briony J. (2005) Researching Information Systems and Computing. Sage.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. (2019) “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” Science 366 (6464): 447–53.

Roselli, Drew, Jeanna Matthews, and Nisha Talagala. (2019) “Managing Bias in AI.” In Proceedings of the 2019 World Wide Web Conference, 539–44.

Sarker, Suprateek, Sutirtha Chatterjee, Xiao Xiao, and Amany Elbanna. (2019) “The Sociotechnical Axis of Cohesion for the IS Discipline: Its Historical Legacy and Its Continued Relevance.” MIS Quarterly 43 (3): 695–720.

Schultze, Ulrike, and Michel Avital. (2011) "Designing interviews to generate rich data for information systems research." Information and Organization 21(1): 1-16.

Sheikh, Mohammad Ahmad, Amit Kumar Goel, and Tapas Kumar. (2020) “An Approach for Prediction of Loan Approval Using Machine Learning Algorithm.” In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 490–94. IEEE.

Shollo, Arisa, and Polyxeni Vassilakopoulou. (2024) "Beyond Risk Mitigation: Practitioner Insights on Responsible AI as Value Creation." In ECIS 2024 Proceedings.

Verma, Sahil, and Julia Rubin. (2018) “Fairness Definitions Explained.” In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). IEEE Computer Society,

Verma, Sahil, Varich Boonsanong, Minh Hoang, Keegan E Hines, John P Dickerson, and Chirag Shah. (2020) “Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review.” arXiv Preprint arXiv:2010.10596.

## APPENDIX

## Contextual data about the study participants and their organizations

This annex presents the participants in the study, including their title and experience, as well as information about the organization they work for and a description of the relevant systems or projects they have partaken in. Descriptions are made as accurate as possible without exposing sensitive information about the participant, nor the company they work for. The descriptions are provided to better understand the results.

### *R1 Data Scientist*

R1 works in a company with more than 5000 employees as a Data Scientist and has done so in the last 5 years. The company is a public agency and has initiated a project with predictive AI that can be used for decision support. The system's prediction would not be the final output but instead, be given to a case manager who would use this information along with other information in order to make a decision. R1 is thus concerned with algorithmic fairness in regard to a decision-support system that would affect people's life. R1 also works on developing other AI systems, but this is the one that is the most relevant. R1 also follows the literature on algorithmic fairness, such as by researching different toolkits that are available.

### *R2 Senior Advisor.*

R2 works in the same company as R1 and is a lawyer. They work with the same projects as R1 does but have a different role, as R2's main role is to give legal advice to different teams using machine learning. This includes making sure that the machine learning systems follow the law, and requirements such as fairness, explainability, transparency, and privacy. Assuring that the translation between law and code is correct is one task that is particularly important. R2 expertise does not lie in the technical aspects of algorithmic decision-making, instead, they use their legal expertise in order to oversee the translation between law and code that is done by developers and data scientists.

### *R3 Data Scientist.*

R3 is educated as a sociologist but now works as a data scientist in a company with around 500 employees. R3 works in a company that specializes in auditing and

controlling various systems and solutions. They work in the company's artificial intelligence department, where tasks include auditing machine learning systems and algorithms, and this is where the relevance of algorithmic fairness comes from in the work that R3 does. R3 follows the literature regarding algorithmic fairness and other publications about artificial intelligence and has also authored papers about artificial intelligence and fairness. They work both on implementing machine learning in their own systems and processes and also auditing other companies' use of machine learning and algorithms. Certain projects R3 has worked on were in relation to analyzing and auditing machine learning algorithms and checking for certain biases.

### *R4 Research Director.*

R4 works as a researcher specializing in machine learning in a company with around 100 employees, R4 has 20+ years of experience. R4 works tightly with both companies and research institutions. R4 stays updated with algorithmic fairness research, and the increase in literature is part of why R4 has taken a special interest in the field. R4 often works on projects where R4 or R4's team only has partial responsibility such as only being in charge of the technical implementations, whereas another team has the superior responsibility, which may include deciding the fairness definition. Their task in these projects is usually to design the algorithm used in a solution and implement fairness accordingly.

### *R5 Department Director.*

R5 works for an insurance company with a number of employees between 1000 and 5000 and has a background in economics. They work as a department director and has 10 years of experience. In order to process insurance claims and decide insurance premiums, the company employs thousands of machine learning models. R5 has a long experience with insurance and the use of machine learning within the insurance context. Algorithmic fairness is vital for R5 along with other aspects of RAI. Fairness is a relatively new concept in regards to the use of algorithms, but at the same time seen as very important, and a key factor for the future in terms of reputation and business value.

### *R6 Data Scientist.*

R6 has worked with algorithmic fairness both as a researcher as well as working as a Data Scientist. They work for a company that makes safety software and has around 500 employees. The current company of R6 is in the process of implementing more and more machine learning in order to streamline their solutions, although it's still at an early stage. R6 has previous experience working for an IT consulting company, where among other things they would provide solutions for implementing algorithmic fairness in AI systems. R6 also follows the literature and has attended several conferences on fairness in AI. Through this work as well as staying up to date with the literature, R6 has a good overview of existing solutions and toolkits.

### *R7 Senior Advisor.*

R7 has a background in the social sciences and is now working as a senior advisor in a company with around 100 employees. They have more than 5 years of experience working with the use and effects of AI. R7 works for a company specializing in consumer rights, such as ensuring fair treatment when a system uses algorithmic decision-making. R7 thus provides a different view on algorithmic fairness, as they "represent" those affected by algorithms, as opposed to those who design and deploy them. As a consequence of this, R7 doesn't always have all of the tools for detecting algorithmic unfairness at their disposal, as they may not have all the data or outcomes available. Instead, they employ different methods for bias detection, such as algorithmic auditing and unsystematic approaches.

### *R8 Lawyer.*

R8 is a lawyer who specializes in AI. R8 has worked at their current company for 3 years and the company has around 100 employees. R8 follows the research that is done and has a particular interest in algorithmic fairness. They work with client companies that wish to ensure that their AI systems are in line with legal regulations, which include ensuring algorithmic fairness. R8 is concerned with how the use of artificial intelligence challenges legal principles, and how bias in algorithms is a challenge to the principle of justice.

***R9 Data Scientist.***

R9 works as a Data Scientist for a company with around 100 employees specializing in digitalization and privacy. R9 has 10+ years of experience working with AI for different companies. Among other focus areas, the company that R9 works for leads artificial intelligence projects where different companies can try out and evaluate their systems. These projects often revolve around privacy and RAI, and around half of the projects are also concerned with fairness. R9 has partaken in these projects where algorithmic fairness is important, and the projects operate in several different contexts such as healthcare, welfare, and surveillance, where both technical and organizational solutions have been proposed to mitigate bias and implement algorithmic fairness.