From Explaining to Diagnosing: A Justice-Oriented Framework of Explainable AI for Bias Detection

Miriam Fahimi[*1], **Laura State**[*2, 3, 4], **Atoosa Kasirzadeh**[5]
[1] University of Paderborn,
[2] University of Pisa
[3] Scuola Normale Superiore
[4] Alexander von Humboldt Institute for Internet and Society,
[5] Carnegie Mellon University
miriam.fahimi@uni-paderborn.de, laura.state@hiig.de

Abstract

Explainable AI (XAI) methods can support the identification of biases in automated decision-making (ADM) systems. However, existing research does not sufficiently address whether these biases originate from the ADM system or mirror underlying societal inequalities. This distinction is important because it has major implications for how to act upon an explanation: while the societal bias produced by the ADM system can be algorithmically fixed, societal inequalities demand societal actions. To address this gap, we propose the RR-XAI-framework (recognition-redistribution through XAI) that builds on a distinction between socio-technical and societal bias and Nancy Fraser's justice theory of recognition and redistribution. In our framework, explanations can play two distinct roles: as a socio-technical diagnosis when they reveal biases produced by the ADM system itself, or as a societal diagnosis when they expose biases that reflect broader societal inequalities. We then outline the operationalization of the framework and discuss its applicability for cases in algorithmic hiring and credit scoring. Based on our findings, we argue that the diagnostic functions of XAI are contingent on the provision of such explanations, the resources of the audiences, as well as the current limits of XAI techniques.

Introduction

Explanations from the field of explainable artificial intelligence (XAI) have become an important tool for identifying biases in automated decision-making (ADM) systems. ADM systems, often based on opaque or complex machine learning (ML) models, make predictions that can have far-reaching implications for individuals, especially when deployed in high-stakes contexts such as credit scoring (Florez-Lopez 2010; Bono, Croxson, and Giles 2021; Onay and Öztürk 2018) or hiring (Yarger, Cobb Payton, and Neupane 2019; Raghavan et al. 2020).

Importantly, predictive ADM systems rely on past data that could reflect structural inequalities, encoding patterns of historic discrimination that can be reproduced or amplified in algorithmic outputs (Ntoutsi et al. 2020; Barocas and Selbst 2016; Barocas, Hardt, and Narayanan 2023; Alkhatib and Bernstein 2019; Fabris et al. 2022; Kasirzadeh 2022). For individuals subject to these decisions, explanations are thus not merely a matter of transparency of an ADM system but also of *justice* (Amoore 2020; Kasirzadeh and Klein 2021; Keyes and Creel 2022).

While a growing body of literature in computer science explores how explanations can support algorithmic *fairness* through the detection of biases (Ramachandranpillai, Baeza-Yates, and Heintz 2023; Zhou, Chen, and Holzinger 2020), we identify two shortcomings in XAI research: First, many XAI/fairness works define fairness through formal metrics, detaching it from legal and societal contexts. Second, the mere focus on the ADM system as the locus of analysis overlooks biases rooted outside of the model and mirrored within it – the so-called societal biases.

Responding to these shortcomings, we turn to an underexplored feminist theory of social justice, particularly Nancy Fraser's influential account of justice, comprising two dimensions: recognition and redistribution (Fraser 1997, 2000, 1995). While recognition refers to the dimension of justice that addresses the acknowledgment of diverse identities, redistribution refers to the distribution of material resources.[1] Drawing on these two principles, we foreground two diagnostic functions of XAI and propose the *RR-XAI-framework* (recognition and redistribution through XAI):

- *XAI as socio-technical diagnosis*: explanations can function as a socio-technical diagnosis when detecting forms of biases in the ADM system arising from structural inequalities (e.g., as defined by anti-discrimination law). Explanations provide an entry point for legal claims-making, institutional contestation, and technical redress. Socio-technical bias can be "fixed" within the ADM system.
- *XAI as societal diagnosis*: as a societal diagnostic, explanations make visible that automated decisions accurately mirror an unjust 'reality'. Here, explanations expose forms of biases that lie outside the ADM system and cannot be "fixed" within it. Instead, this diagnosis demands broader redistributive or political responses.

[*]Shared first authorship.

[1]Fraser labels recognition as the 'cultural dimension' of justice to combat cultural domination, nonrecognition, and disrespect and redistribution as the 'economic dimension' of justice through remedying economic exploitation and deprivation (Cesario Alvim Gomes 2018).

With the RR-XAI framework, we invite practitioners, auditors and data subjects to view explanations not only as a technical tool but also as an entry lens for understanding whether justice requires fixing an ADM system, the context within which it functions, or the world that shapes it.

Contribution Bridging perspectives from feminist justice theory, science and technology studies (STS) and XAI/fairness research, we make the following contributions:

1. We introduce the *RR-XAI-framework* as a justice-oriented conceptualization of XAI for bias detection. It is based on the following two conceptual contributions:
   (a) An understanding of bias that distinguishes between socio-technical and societal bias following STS scholar Lopez (2021) to acknowledge the different origins of bias.
   (b) An application of feminist justice theory following philosopher Fraser (1997, 2000, 1995) to theoretically ground the two justice-oriented, diagnostic functions of XAI that explanation receivers can act upon.
2. We exemplify the applicability of the framework for cases in algorithmic hiring and credit scoring.
3. We discuss the implementation and obstacles of the proposed framework.

Scope This paper focuses on the explanations used in socially consequential decision-making contexts, where ADM systems have the potential to significantly exacerbate existing societal inequalities (Benjamin 2019; Noble 2018; Scheuerman, Pape, and Hanna 2021). To illustrate these examples, this work draws on the EU legal context. As a conceptual contribution, our work has implications for a diverse range of scholars, practitioners, and the XAI/-fairness community, as we adopt a social justice-oriented lens and present our RR-XAI framework as a tool for diagnosing socio-technical and societal biases. We also believe our framework holds practical relevance: it demonstrates how explanations can prompt actions like refining ADM systems, supporting legal claims, and driving societal change. These actions have wide-ranging implications, affecting not only those directly impacted by ADM systems but also civil society more broadly. Our research is the result of an interdisciplinary collaboration drawing on our research and experiences as scholars in the field of XAI/fairness, computer and social science, and our shared grounding in feminist scholarship.

Structure We start by discussing the varieties of bias in ADM systems and by presenting Lopez's taxonomy as a useful tool for distinguishing between three forms of bias (Section: Discrimination and Biases in ADM Systems). We then review current approaches in algorithmic fairness and XAI, with a focus on how XAI is utilized for bias detection. We also briefly explore related work that approaches XAI from justice-oriented and feminist perspectives (Section: Related Work: XAI for fairness). Building on calls for justice-oriented XAI, we draw on underexplored feminist and social theories of justice to propose the RR-XAI-framework (Section: The RR-XAI-Framework). We demonstrate its applicability by two illustrative cases in hiring and credit scoring (Section: Examples). We conclude with a discussion of conceptual considerations and four key conditions for an operationalization and implementation of the framework (Section: Discussion), discuss limitations and directions for future work (Section: Limitations and Future Work) and close with an outlook (Section: Conclusion).

Discrimination and Biases in ADM Systems

ADM systems have a significant impact on various aspects of our lives. In the past, such systems have been used to predict chances on the job market (Allhutter et al. 2020), social welfare allocation (Freeman, Shah, and Vaish 2020; van Bekkum and Borgesius 2021), or a criminal's recidivism for legal proceedings (Angwin et al. 2018; Corbett-Davies et al. 2017). Given the high-stakes context of such automated decisions, there is growing public and scholarly concern over the social consequences of opaque algorithmic predictions and the harm they may cause to data subjects and society at large (Narayanan and Kapoor 2024; Noble 2018; Eubanks 2017; Chun 2021; Benjamin 2019).

As a consequence of this controversy, a new field grounded in computer science has emerged to develop methods and metrics to detect and quantify algorithmic discrimination. This field is referred to as *algorithmic fairness* (Barocas, Hardt, and Narayanan 2023; Mitchell et al. 2021; Weerts et al. 2024) (or sometimes, as fair-AI (Ruggieri et al. 2024)). In algorithmic fairness, discrimination is typically framed as *harmful bias*, further conceptualized through a pipeline model of AI development, which breaks ADM systems down into three stages – data, model, and output – each seen as a potential site where biases can arise and be mitigated (Ntoutsi et al. 2020). Along these three stages, various subcategories of bias have been identified. For example, Mehrabi et al. (2021) proposes a taxonomy of 23 distinct types of *data biases*, ranging from measurement and sampling bias to aggregation and representation bias.

Alongside these approaches to detect and mitigate bias in ADM systems, several critical scholars have emphasized that such *technically-oriented taxonomies* risk overlooking deeper structural and historical injustices (Miceli, Posada, and Yang 2022; Hildebrandt 2019; Friedman and Nissenbaum 1996; Reyero Lobo et al. 2024).

In light of these concerns, we draw on the bias taxonomy proposed by STS scholar and mathematician Paola Lopez (2021). Lopez distinguishes between technical, socio-technical, and societal bias to enable a clear conceptual separation between technical and societal sources of bias and appropriate forms of intervention. In this contribution, we adopt this taxonomy precisely because it offers a pragmatic and justice-sensitive foundation for understanding these different sources of biases in ADM systems and aligning them with structural inequalities. We value her framework not only for its alignment with critical perspectives on bias, but also for its accessibility and its potential to operationalize such perspectives in practice.

In detail, Lopez three-fold taxonomy of bias is structured as follows:

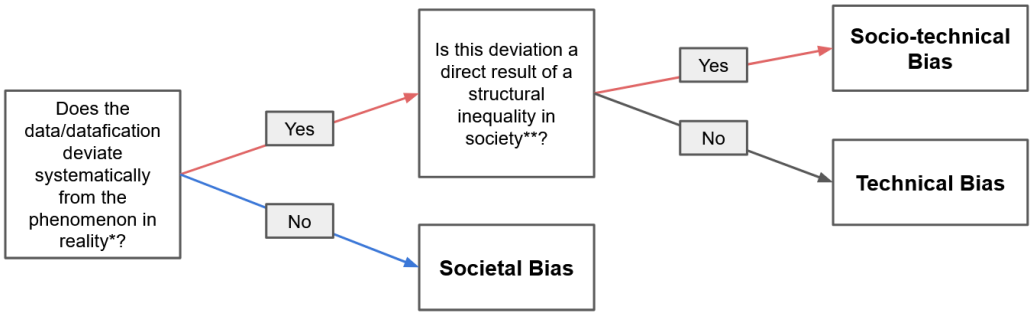

Figure 1: Bias scheme as described in (Lopez 2021). (*) Reality is a highly contested term (Bowker and Star 2000; Jacobs and Wallach 2021). (**) Structural inequalities are defined based on protected features as stated in anti-discrimination law. Figure re-created by the authors.

1. *Technical bias* refers to mismatches between what is intended to be measured and what is actually measured, due to technical or conceptual shortcomings. This type of bias is not rooted in structural inequality. This is also the reason why technical bias is not part of our justice-oriented RR-XAI framework.
2. *Socio-technical bias* arises when structural inequalities, as defined by existing legal frameworks, affect the data such that it no longer represents the phenomenon intended to be measured ('reality'). Socio-technical bias are an entry point for legal intervention against the AI provider based on existing anti-discrimination law.
3. *Societal bias* occurs when data accurately mirrors a (structurally unjust) 'reality'. The data depicts correctly that society structurally discriminates against certain social groups or marginalized communities.

To exemplify the difference between these biases, Lopez introduces the case of the Austrian public employment service (short: AMS) that designed an algorithm to predict job seekers' likelihood of integrating into the labor market. Based on this predicted likelihood, job seekers were assigned one out of three possible classes, depending on whether they had high (at least 66%), medium (between 25% and 66%) or low chances (less than 25%) to return to the job market. Only those with medium chances received AMS support, as high-chance individuals were expected to find jobs independently, and low-chance individuals were considered unlikely to succeed. The model explicitly weighted factors like female gender, care responsibilities, and non-European citizenship negatively.

1. Technical bias is the distortion that occurs when a model relies on pre-COVID labor market data to predict post-COVID employment chances, creating a mismatch between outdated patterns and a structurally changed labor market.
2. Socio-technical bias is the continued use of a binary gender classification despite the legal recognition of non-binary gender identities in Austria, reflecting how structural gender inequalities shape the AMS data and lead to the exclusion of non-binary identities.
3. Societal bias is the reflection of structural inequalities within a system, such as the statistically lower labor market chances of women, older people, or non-EU citizens. While this kind of bias can reinforce inequalities when used to allocate fewer resources to disadvantaged groups (as it was the case), Lopez argues that it can also be employed as an "emancipatory tool of diagnosis" to make these societal inequalities visible and politically actionable.

Fig. 1 illustrates how different types of bias can be identified using a diagnostic scheme based on a guiding questionnaire. The first question is: "Does the data/datafication deviate systematically from the phenomenon in reality?" If the answer is "No", and the data accurately reflects existing structural inequalities – such as the lower employment chances of women, older people, or non-EU citizens – this constitutes a societal bias. If the answer is "Yes", a follow-up question is: "Is this deviation a direct result of a structural inequality in society?" If the answer is "No", as in the case of using outdated pre-COVID data, this indicates a technical bias. If "Yes", the deviation stems from embedded structural inequalities – such as the binary coding of gender in systems despite legal recognition of non-binary identities – this is classified as a socio-technical bias.

This leaves the question of what exactly constitutes a phenomenon that can be measured – or, more fundamentally, what 'reality' is. As we critically acknowledge, reality is highly contested, since measurements are always only proxies for "unobservable theoretical constructs" (Jacobs and Wallach 2021). Following Lopez' understanding, we use the term 'reality' deliberately to highlight that datafication can *systematically* diverge from the phenomena it seeks to represent – a divergence that is not accidental but often patterned in structural inequalities that shape how certain phenomena are understood and acted upon.

As we will further discuss in the Section "Discussion", the distinction between socio-technical and societal bias is also not always clear-cut. However, drawing on this distinction remains analytically and practically valuable for identifying the different sources of bias and linking them to the corresponding justice-oriented action.

Related Work: XAI for Fairness

Explainable artificial intelligence (XAI) is a research field focused on the development of algorithmic explanations for AI (Gunning et al. 2019; Guidotti et al. 2018; Molnar 2019).[2] Different scholars in the field of XAI propose and use different definitions of what an explanation is (Mittelstadt, Russell, and Wachter 2019; Lipton 2018; Rohlfing et al. 2020). In our contribution, we build on the following understanding: "Given an audience, an XAI is one that produces details or reasons to make its functioning clear or easy to understand" (Barredo Arrieta et al. 2020).[3] Explanations are necessary not only for opaque ML models but also for interpretable ones; since some interpretable models – such as deep decision trees – can be quite complex (Molnar 2019).

XAI methods are usually divided into two categories, depending on their functionality: model-agnostic and model-specific (Guidotti et al. 2018; Molnar 2019). Instances of the first work on any, instances of the second only on particular models. The method SHAP (Lundberg and Lee 2017) – the method that we build on in the Section "Examples" – is model-agnostic. Additionally, an XAI explanation is distinguished by its validity: it is locally valid, i.e., for a specific data instance, or globally, i.e., for the full model (Guidotti et al. 2018; Molnar 2019). SHAP provides local and global explanations, and we use the local version in our work.

Most XAI methods – also SHAP – produce a static output, i.e., fixed information, with a few exceptions implementing interactivity, e.g., (Sokol and Flach 2020b), allowing for "what-if" questions by the user and interactions via voice and text. Other work exists that critically engages with some dimensions of explanations, e.g., the context of an explanation (Sloane et al. 2023), their evaluation (Nauta et al. 2023) and documentation (Sokol and Flach 2020a).

Extending (State 2021), an explanation can have different *purposes* or *functional goals* which can be broadly summarized as understanding, legal compliance, bias detection, and algorithmic recourse. Central to our work is the purpose of bias detection, which we outline in the next section.

[2]Two remarks: first, in this work, we focus on explanations for AI and exclude *transparent-by-design* approaches, sometimes also included under the umbrella of XAI. Second, we use the term AI, even if most of the methods focus on an explanation of ML models, a specific sub-field of AI, also called *sub-symbolic*. While in recent years the focus has been on developing explanations for such models, the first instances of XAI date back several decades and address so-called expert systems (Confalonieri et al. 2021). These systems are instances of *symbolic* methods, a different branch of AI.

[3]A social science perspective on XAI offers a different viewpoint, where explanations can serve as a lens to infer the relations between humans and machines (Borch and Hee Min 2022), or between explainers and the explainees (Rohlfing et al. 2020). This perspective also sheds light on XAI explanations being co-constructed and performative, including not only formalized knowledge but also intuition and different forms of expertise (DeVito, Gergle, and Birnholtz 2017), and how we still rely on narratives to make sense of automated actions (Andrejevic 2020). These works are important to consider, but not central to our argument.

Bias Detection through XAI

Utilizing XAI to *detect bias* is one of the motivating factors for researchers to design XAI methods (Pedreschi et al. 2019; Guidotti et al. 2018). Since the primary goal of an explanation is to clarify how a decision was reached in an ADM system, it should also provide information about the features involved in the decision and their influence on the result of the decision. Bias becomes visible (or is detected) through an explanation when it shows that the ADM system relies on an identity category (like gender or age)[4] or a proxy for it. Bias is detected when the use of such categories or proxies cannot be justified legally or ethically. By proxy, we refer to a variable that correlates to an identity category, and that therefore acts as a substitute for it. The notion of proxy (and discrimination via a proxy) is an ongoing discussion, see for example (Tschantz 2022).

Despite the huge potential in the synergy between the field of algorithmic fairness and XAI, e.g. what we hence call *XAI/fairness*, only few works at the intersection exist yet. An early work that broadly discusses this synergy is Zhou, Chen, and Holzinger (2020): the authors discuss existing research under four different point of views: i) explanations as a guarantee of model fairness; ii) the influence of explanations on the perception of fairness; iii) the perception of the fairness of features and its relation to XAI; as well as iv) fairness and counterfactual explanations. Furthermore, the authors do not define fairness in their work, rather, they acknowledge that there are different ways to understand it.

A more extensive study is (Ramachandranpillai, Baeza-Yates, and Heintz 2023). The authors survey 56 research works in XAI/fairness and propose a taxonomy for their classification. The majority of works are grouped under the umbrella term *XAI for bias mitigation*: these are approaches that can be used to analyze and improve the fairness (the satisfaction of a fairness metric) of a given ML model. The approaches are further classified by whether they quantify bias in the *result* of the model (in the decision itself) or in the *procedure* of reaching that decision. Based on the specific XAI method used in the approach, a more fine-grained distinction can be made. The second umbrella term is *XAI for bias evaluation*: here, the authors group approaches that use XAI to analyze the fairness (here, the satisfaction of a fairness metric) of either a specific ML model or of a surrogate model. With a surrogate model, the authors refer to explanation methods that explain an ML system by approximating it from outside and thereby build the surrogate. Such a model and the explanation of this model can also be quantified by its fairness. Further, they make a distinction between whether the fairness relates to the output of the (ML or surrogate) model, to how a user understands the fairness of this model or whether it is a combination of the two.

In this work, we will utilize the local version of the explanation method SHAP (Lundberg and Lee 2017), a well-known and frequently used XAI method. According to the taxonomy of Ramachandranpillai, Baeza-Yates, and Heintz

[4]Note, these categories used in computer science are often simplified representations of more complex social identities (Keyes, Hitzig, and Blell 2021).

(2023), this approach classifies as *XAI for bias mitigation / result-oriented / post-hoc / feature-specific*.[5] The choice of an XAI method, however, is not fixed when applying our framework: we will discuss different choices and implications for the framework in the Section "Discussion".

Both surveyed papers define fairness as formalized equality goals (metrics) as put forward by the algorithmic fairness literature (Verma and Rubin 2018; Ntoutsi et al. 2020). As a result, they focus on the "abstraction of fairness" (Selbst et al. 2019) without interrogating how fairness relates to broader legal or societal structures.

Building on existing work connecting algorithmic fairness with legal frameworks (Weerts et al. 2023; Wachter, Mittelstadt, and Russell 2021; Berk et al. 2021), our study extends this perspective to the XAI/fairness literature in computer science that we surveyed. With its predominant focus on the data–model–output pipeline, existing work in XAI/fairness has so far engaged less with forms of bias that manifest outside the ADM system itself – such as societal bias, as introduced by (Lopez 2021).

In our work, we address this by adopting an understanding of bias that integrates both socio-technical and societal bias as put forward by (Lopez 2021); and by mobilizing justice theory to assess whether XAI can have diagnostic functions that can positively contribute to contest, challenge and change diverse sources of bias. Originating from these research gaps, we now outline the related work at the intersection of feminist justice theory and XAI.

Critical Feminist Scholarship on XAI

Recent scholarship across critical algorithm studies, feminist theory, STS, and computer science has called for justice-oriented approaches to XAI (Rohlfing et al. 2020; Borch and Hee Min 2022; Klumbyte, Piehl, and Draude 2023b; Keyes and Creel 2022; Amoore 2020; State and Fahimi 2023). These approaches question the dominant framing of XAI as making internal system mechanisms more accessible and more transparent. As scholars have shown, transparency is a contested and situated ideal (Burrell 2016; Ananny and Crawford 2018; Fahimi and Kinder-Kurlanda 2025), that also often fails to produce actionable or contextually meaningful explanations for those affected by algorithmic systems (Miller 2019).

In that regard, Felzmann et al. (2019) and Rohlfing et al. (2020) argue that explanations must be evaluated based on whom they serve and whether they enable action. This reframing of XAI underscores power asymmetries: AI providers define what is explainable and how, while data subjects are often left without the means to question or challenge decisions that affect them (Browne 2023; Kasirzadeh 2022; Wellner and Rothman 2020).

To address this, scholars argue for redistributing epistemic authority and designing explanations that support critical engagement, contestation, and participation (Klumbyte, Piehl, and Draude 2023a,b). Huang et al. (2022) propose a feminist and integral XAI (IXAI) approach to incorporate diverse forms of knowledge, thereby strengthening objectivity and accountability. Their framework encourages practitioners to practice humility and openness, marginalized groups to cultivate self-trust and advocacy, and AI providers to promote epistemic democracy. Similarly, Hancox-Li and Kumar (2021) critique dominant XAI methods for reinforcing Eurocentric and patriarchal assumptions, such as universalism and modularity, while neglecting critical epistemic values like context sensitivity and feature interdependence.

Our framework builds on and complements these efforts by explicitly linking bias detection to justice-oriented action. By applying our framework to SHAP (though other techniques are equally possible), we show that explanations can reveal more than internal mechanics. This framing positions explainability not just as a technical goal, but as a means for justice, particularly when explanations are used to challenge power imbalances between AI providers and those subjected to their systems. In cases of socio-technical bias, for example, an explanation can support legal claims-making and strengthen the position of data subjects towards AI providers. Importantly, by grounding our framework in Fraser's theory of justice, as proposed in the following section, we also offer a robust theoretical foundation that adapts well to the algorithmic context. Fraser's distinction between recognition and redistribution enables us to conceptualize and operationalize the broader claim of "justice". However, for XAI to function as a diagnostic for injustice – whether socio-technical or societal – explanation recipients must have the necessary resources and contextual knowledge to interpret and act upon what is revealed. We return to this challenge in detail in the discussion section.

The RR-XAI-Framework

To conceptually ground our endeavor of justice-oriented diagnostic functions of XAI, we now turn to the work of Nancy Fraser (Fraser 1997, 1995, 2000). Her theory of social justice offers a powerful lens for understanding the implications of different types of bias, particularly when it comes to deciding what kinds of diagnostic functions and interventions are appropriate, and for whom.

Fraser has proposed two pathways to achieving social justice: *recognition* and *redistribution*. On the one hand, the call for *recognition* seeks to acknowledge and appreciate differences related to identity categories (such as gender, race, or age). Consider the scenario of experiencing mansplaining at work, where you feel patronized (Turesky and Warner 2020; Stone 2022). To tackle this as an injustice and recognize it as sexist behavior, you must first understand the concept of gender, then identify your own gender identity, and finally acknowledge that gender can influence others' behavior towards you. For instance, recognizing that 'as a woman, this behavior is sexist against me' helps in addressing the issue. Thus, as an affected individual who speaks up against a sexist colleague, knowing about and using an identity category becomes necessary to point out the experienced injustice. This act of recognition is not only personal; it belongs to

[5] Local SHAP is an explanation method that explains the reasons behind a single decision (taxonomy *XAI for bias mitigation / results-oriented*). It relies on a post-hoc approach (taxonomy *post-hoc)*. The explanation surfaces the most relevant features used by an ADM system (taxonomy *feature-specific*).

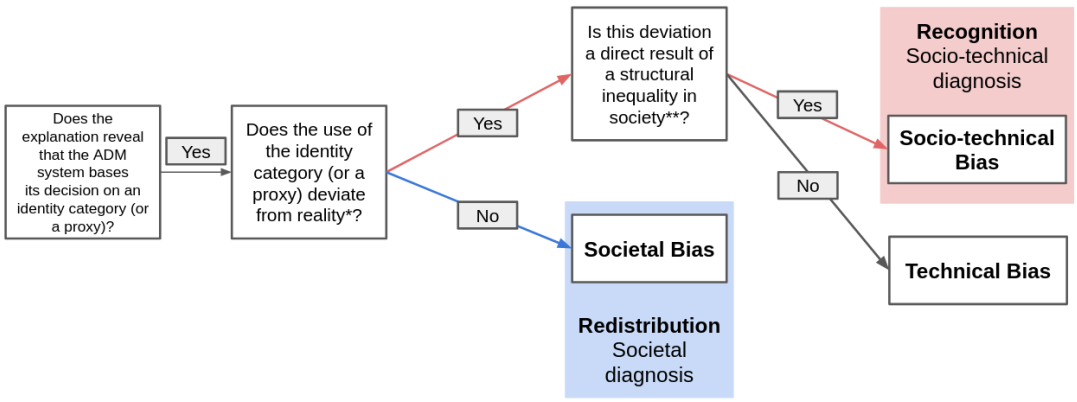

Figure 2: Adaptation of bias scheme by (Lopez 2021) for the RR-XAI-framework. (*) Reality is a highly contested term. (**) Structural inequalities are defined based on protected features as stated in anti-discrimination law.

what Fraser terms *symbolic or cultural justice*, which entails challenging dominant value patterns and making discrimination based on identities visible and actionable. Crucially, such recognition can also enable legal claims-making by anchoring lived experiences of injustice in institutional categories protected under anti-discrimination law.

On the other hand, *redistribution* aims to reshape systemic conditions (Fraser 1995).[6] Consider now, that you want to address the gender pay gap in your workplace (Abdel-Raouf and Buhler 2020; Staff 2018; Acred 2016). First, you need to know the wages of all employees. As this information is not publicly provided, you ask the CEO for transparent information to initiate an open discussion based on the exact numbers. Let's assume that an insightful CEO agrees, and you can now begin to talk about wages among your colleagues. However, changing the gender pay gap in the company cannot be easily achieved. Such a change is contingent on a transformation of gendered and capitalist class relations that assign (different) wages to men and women in the first place. In an ideal world, this would first involve recognizing the existence of the gender pay gap, followed by the redistribution of economic resources. Over time, this process would lead to the gender pay gap and the overall effects of gender becoming less important for the redistribution of wages. In the long run, redistributive justice aims towards a complete detachment from gender (and class) to bring about a systemic transformation in societal conditions.

In the following, we present our framework, grounded in this two-fold conceptualization of justice, to re-purpose XAI as diagnostic tool. We associate the detection of socio-technical bias with Fraser's concept of recognition, and the detection of societal bias with redistribution, with the aim of prompting action on these biases.

[6]Note that Fraser's definitions of categories are closely intertwined with the difference between recognition and redistribution. Here, the term identity category (Hegelian terminology) relates to a societal construct that shapes individual (or collective) identities, such as gender, race or age. Opposed, "social category" (Weberian terminology) and "economic category" (Marxian terminology) relate to systemic conditions and class relations (Fraser 2000).

Towards Recognition: XAI as Socio-technical Diagnosis

When the decision of an ADM system substantially relies on the use of an identity category in a way that misrepresents reality (socio-technical bias), the issue at stake is one of recognition. This constitutes a form of symbolic or cultural injustice: the ADM system reproduces patterns of stereotyping based on identity. For justice to be achieved, both the data subject and the AI provider must engage in a process of recognition – acknowledging that the ADM system has illegitimately used an identity category, leading to discriminatory outcomes. XAI techniques can support this process by making the influence of variables transparent, thus enabling critical reflection on how identity categories are used. Critically, recognition can also serve as a basis for legal claims. When XAI uncovers that the use of identity categories violate anti-discrimination laws, it can support claims towards institutional accountability through existing legal frameworks.

Towards Redistribution: XAI as Societal Diagnosis

When the decision of an ADM system accurately reflects reality, but that reality is itself unjust (societal bias), XAI should function as a form of societal diagnosis. In such cases, the diagnostic function aligns with Fraser's principle of redistribution by drawing attention to the socio-economic structures that shape and normalize inequality in reality. This includes how identity categories such as gender, race, or class are inscribed into data through structural patterns of access, exposure, and exclusion in reality. Justice-oriented action, in this context, does not primarily refer to changes within the ADM system, but rather to broader societal transformation. Here, structural change at the political level would be needed to ensure that ADM systems no longer reproduce societal biases. Current XAI techniques can help by making the influence of input variables visible. However, linking these variables to broader structural injustices requires interpretive work and domain-specific knowledge, contingent on the audiences of an explanation.

The RR-XAI framework integrates two key concepts: *recognition*, which involves explanations that uncover and address socio-technical bias and empower individuals with the information necessary to hold AI providers accountable, and *redistribution*, which focuses on detecting and altering societal bias. Together, these elements transform explanations into tools that drive improvements in algorithmic decision-making and promote social justice.

Fig. 2 displays the RR-XAI framework, and builds on Lopez's typology of bias as depicted in Fig. 1, by proposing two additional diagnostic steps:

1. *Does the explanation reveal that the ADM system bases its decision on an identity category (or a proxy)?* This assesses whether the system encodes, relies on, or infers identity-based information (e.g., gender, race, age) as part of its decision-making.
2. *If yes, does the use of the identity category (or a proxy) deviate from reality?* This step investigates whether the encoded identity misrepresents, essentializes, or simplifies identity categories. A deviation signals socio-technical bias, if based on structural inequalities; alignment with potentially unjust 'realities' may indicate societal bias.

Examples: Hiring and Credit Scoring

In this section, we apply the RR-XAI framework to two illustrative cases: first, we demonstrate the use of XAI as a potential socio-technical diagnosis in the context of algorithmic hiring; second, we explore its use as a societal diagnosis in credit scoring. While the presented cases are fictional, they are grounded in real-world cases of high-stakes algorithmic decision-making: For example, our algorithmic hiring example draws from the well-known Amazon case (Dastin 2018) and builds upon existing research initiatives focused on explanations for hiring algorithms (Beretta et al. 2024).[7] In the case of credit scoring, our example is inspired by the "Credit Score Calculator", an explanatory tool to understand credit scoring that has recently been provided by a leading German credit agency.[8] While such explanations are not yet standard practice, we expect their development and deployment to accelerate significantly. With the adoption of the EU AI Act in spring 2024, providers of high-risk AI systems are legally required to offer meaningful explanations.[9] Both examples discussed fall in the high-risk category. Furthermore, the importance of explanations in credit scoring is reinforced by the European Court of Justice, which ruled in 2025 that fully automated credit assessments fall under Article 22 of the GDPR and must be accompanied by meaningful information about the logic involved.[10]

[7] See also https://findhr.eu/about-findhr/

[8] https://www.schufa.de/scorechecktools/scoresimulator/

[9] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[10] https://safe-frankfurt.de/news-latest/safe-finance-blog/details/explaining-credit-scores-the-european-court-of-justice-rules-on-automated-credit-assessments.html

The two examples draw on SHAP (Lundberg and Lee 2017), a widely known and used explanation method.[11] The principal output is feature names, their relevance scores, and associated decisions. Since SHAP is a post-hoc approach, the providers of the ADM systems – a tech company and a bank – can generate explanations once the model that is the basis to the ADM system is trained. The input to SHAP is the trained model and the data for which explanations are required. The method is available as a `Python` package. While the output of SHAP is by default displayed in plots, it can be adjusted to what the providers deem most useful. Here, we assume that they reformulate the output as natural text and reduce it to the three *most important* features. These design choices are based on studies that emphasize the importance of text over numbers and mathematical notation (Miller 2019; Mittelstadt, Russell, and Wachter 2019), and additionally aim at minimizing cognitive load for users.

XAI as Socio-technical Diagnosis: Towards Recognition

A tech company uses an ADM system when hiring new employees. The system assesses the résumés (CVs) of the applicants and ranks them based on their potential fit for the open position. Applicants with a high ranking will be invited to an interview. The company provides the following explanation to an individual who was not invited to an interview:

> *We regret to inform you that your application was not selected for an interview due to a low match between your CV and the requirements of the position. The three most relevant pieces of information that informed our decision (starting with the highest relevance) are:*
>
> 1. your experience in the tech sector is below three years;
> 2. you went to an all-female college;
> 3. you do not have coding experience in JAVA.

The explanation reveals that one of the decisive factors of the ADM system was the *gender* of the applicant, as coded through using the proxy feature *all-female college*. In this case, the proxy is easy to identify, as it can be directly inferred from the feature's value ("all-female college)" requiring no external knowledge. Therefore, the explanation revealed that a proxy of the identity category *gender* was used by the ADM system (as related to RR-XAI-framework's first question "Does the explanation reveal that the ADM system bases its decision on an identity category (or a proxy)?").

The use of this proxy could be the result of an ADM system trained on a dataset of CVs that reflects historical gender imbalances in the tech labor market: because the dataset includes more CVs from male applicants, the system learns to associate being male with being a better candidate for tech roles. According to Lopez (2021), this constitutes a deviation from a version of reality defined by anti-discrimination law, which prohibits gender-based discrimination in employment ("Does the use of the identity category (or a proxy)

[11] 216M downloads of the *shap* python package, retrieved on 2024-05-29 from https://www.pepy.tech/

deviate from reality?" in the framework). For example, in the EU, the Recast Gender Equality Directive (Parliament and the Council 2006) explicitly forbids discriminatory hiring practices based on gender. The deviation lies in the fact that the system reproduces historically rooted inequalities rather than aligning with the normative and legal expectation that gender should not influence employment opportunities. In this sense, the system does not simply reflect reality but it reinforces a biased version of it. Thus, following the RR-XAI-framework further ("Is this deviation the direct result of a structural inequality in society?"), a *socio-technical* bias can be identified: the discriminatory output arises from the interaction between unequal historical data and the biased data representation, and a deviation from legal standards.

This bias can be "fixed": one possible action to counteract this socio-technical bias is to enhance the dataset that contains the CVs such that it is balanced with respect to gender, for example, by adding synthetic CVs. The ADM system will learn not to make a distinction between men and women when searching for their fit for the job. Knowing that there is a socio-technical bias, the explanation forms also a basis to legally challenge the ADM system: because it discriminates data subjects based on their gender which is according to the sectorial law in employment, not allowed. As we explain below, the information provided by the explanation potentially supports a claim of *prima facie* discrimination. Thus, this explanation can serve as socio-technical diagnosis.

If the ADM system is already in deployment, it first has to be withdrawn to "fix" it. This was the case for the Amazon hiring algorithm, which is inspirational to our example. In 2018, Amazon introduced a new recruiting software for tech jobs. However, shortly after the introduction, the ADM system had to be withdrawn because it assigned negative ratings to CVs of female applicants: the software penalized CVs that contained words such as "women's" or "women's chess club captain". The reason for this was training data that contained more CVs from men than from women (Dastin 2018).

XAI as Societal Diagnosis: Towards Redistribution

A bank deploys an ADM system to evaluate the creditworthiness customers applying for a loan. This evaluation is based on the prediction of a credit score. The outcome can profoundly affect the customers's financial opportunities: a high score typically leads to loan approval, whereas a low score results in rejection. The bank provides the following explanation to an individual whose application was denied:

> *We are sorry to inform you that your credit score is too low to approve your loan application. The three factors that most influenced this decision (in order of importance) are:*
>
> 1. you opened your oldest checking account ten years ago;
> 2. you own four credit cards;
> 3. you do not have a real estate loan.

The explanation reveals that the feature *real estate loan* was one of the decisive factors for the denial of the loan. While this feature appears neutral at first glance, it can also stand as a proxy for gender: historic, gender-based discrimination has led to significant disparities in property ownership between men and women (Chu, Hsu, and Wang 2023; Goldsmith-Pinkham and Shue 2022). While the explanation reveals that the feature real estate loan influenced the decision, it is only by incorporating external *background knowledge* and thus by understanding that this feature serves as a proxy for gender, that one can fully recognize the use of a proxy for an identity category by the ADM system. Without this crucial knowledge, the explanation alone does not transparently disclose the influence of protected characteristics. Thus, possessing such background knowledge is essential to positively answer the RR-XAI-framework's first question: "Does the explanation reveal that the ADM system bases its decision on an identity category (or a proxy)?"

In contrast to the previous case in hiring, the use of a proxy for an identity category (real estate loan) does not deviate from 'reality' nor does it violate existing regulations. The proxy is considered a legitimate indicator of financial reliability and is also often used as a feature in credit scoring:[12] owning property is closely tied to ownership relations that historically and socially signify wealth and economic stability. However, this reality is far from objective or fixed; it is a construct shaped by historical and social conditions.

Therefore, the answer to the second question of the RR-XAI-framework "Does the use of the identity category (or a proxy) deviate from reality?" is "No". Following the RR-XAI-framework, a *societal* bias can be identified. The explanation serves here as a societal diagnosis by accurately mirroring structural injustices in society (unjust 'reality'). To address such a societal bias, broader societal changes are necessary, as we will discuss in the following section.

Discussion

In this section, we discuss four key points: i) the diagnostic functions of XAI; ii) the resources necessary to effectively utilize these diagnostic functions; iii) the distinction between socio-technical and societal bias; iv) the possibilities and limitations of current XAI techniques.

From Explaining to Diagnosing In the first example of algorithmic hiring, we have shown that the XAI can serve as a *socio-technical diagnosis* by revealing that the identity category gender has been inappropriately encoded – through a proxy feature that appears as a relevant factor in decision-making. In this case, an explanation can enable affected data subjects or advocacy groups to contest the automated decision within existing legal frameworks. For the EU context, this would be the Recast Gender Equality Directive (Parliament and the Council 2006) that prohibits the use of gender as a decisive factor in employment. The explanation can support a formal complaint by documenting that otherwise equally qualified female applicants were systematically disadvantaged by the ADM system. Drawing on Fraser's concept of *recognition*, this diagnosis made the misrepresentation based on a structural inequality as defined by the law

[12] Based on the "Credit Score Calculator" https://www.schufa.de/scorechecktools/scoresimulator/

visible and opened a path towards institutional accountability. The information provided by an explanation may be information to establish a *prima facie* discrimination claim: in cases of algorithmic discrimination, the alleged party (the provider) possesses more information. In such a context, a data subject (or another civil society actor) could establish a legal claim via a "presumption of discrimination". It is then the task of the alleged defendant to prove that the ADM aligns with anti-discrimination law (European Union Agency for Fundamental Rights 2018).

In contrast, detecting the use of proxy variables for gender in the credit scoring domain is less straightforward. First of all, the explanations demonstrated that the variable "real estate loan" was used by the ADM system. We showed that this variable correlates with gender. This can lead to a (gendered) feedback loop: those who already possess capital and assets have easier access to credit, while those without such resources face systemic barriers, reinforcing and perpetuating existing economic inequalities. However, the use of the variable does not pose a violation of the law. Because the same variable is also strongly correlated with financial reliability, it is an accurate reflection of a reality that also encodes social inequalities. Following the RR-XAI-framework, here, the explanation can serve as a *societal diagnosis*. Thus, subsequent actions needs to focus on *redistribution* to change gendered ownership structures. From this perspective, XAI's diagnostic capacity can become a critical lens through which the capitalist and gendered structures underpinning financial decision-making are made visible.

While our RR-XAI-Framework offers a pathway from *explaining to diagnosing* and justice-oriented action, the applicability of the framework is contingent on the existence of such explanations – something that is currently, unfortunately, still far from guaranteed. Following a feminist tradition of envisioning *alternative futures*, we argue that it is nevertheless essential to think through how explanations could contribute to justice, even if current AI practices do not yet support this. Imagining such futures is not a naïve exercise, but also a political and epistemic intervention: it allows us to identify the necessary conditions that must be created to make these futures possible. In the following, we further discuss some of these conditions.

Resources A second crucial condition is whether explanation recipients are equipped with interpretive knowledge and resources to respond to and act upon an explanation. Following Fricker (2007), this also raises concerns of "epistemic injustice": such knowledge is not available to all users of XAI in the same way (Barredo Arrieta et al. 2020), and people differ in their ability to make use of explanations (Felzmann et al. 2019). In the case of algorithmic hiring, this included knowledge about anti-discrimination law, while in the case of credit scoring, a very specific kind of knowledge was needed to identify that the possession of a real estate loan could be a proxy for gender. Technically, *logic reasoning* can be used to account for such knowledge (Calegari, Ciatto, and Omicini 2020; State 2021). However, there exist only few (yet increasing) number of methods that encode background knowledge exists (Calegari, Ciatto, and Omicini 2020; Beckh et al. 2021). This is, on the technical side, due to the computational complexity of logic-based XAI methods that allow encoding knowledge. On the other hand, caution has to be taken when one wants to encode legal knowledge. While a simple check of whether an explanation contains the protected attributes is feasible, coding legal knowledge always bears the danger of oversimplifying that legal reasoning is a more dynamic process (Weerts et al. 2023).

What is more, financial and time resources for legal action against socio-technical bias, as well as political and collective influence to counteract societal bias, are needed. Such resources are not always available to a single data subject, or – as in the case of collective influence – cannot be mobilized by a single person. Therefore, we suggest that structured pathways should be established and strengthened for data subjects to share explanations with civil society organizations. Some non-profit organizations have already taken important steps in this direction by collecting individual cases of algorithmic discrimination.[13]

Distinction between Socio-Technical and Societal Bias Another key condition lies in the ability to differentiate between societal and socio-technical bias. This distinction might be difficult to sustain in practice, as also discussed in the work of Lopez (2021). It stems from the fact that classifying a given bias as either socio-technical or societal is not merely an empirical question, but reflects a commitment to a particular version of 'reality'. Consider again the case of the hiring algorithm trained on historical data in which women were underrepresented. We have classified this as a socio-technical bias, insofar as the training data failed to adequately represent qualified female applicants. From this perspective, we claimed that the data is distorted and could be corrected through technical interventions such as rebalancing. However, one could equally argue that this reflects a societal bias: the model faithfully mirrors broader gendered structures in the tech labor market, which very much shapes access to employment and career trajectories today (Wajcman and Young 2023). At the end, when it comes to gender, all actions are somehow constituted by it (Hu and Kohler-Hausmann 2020). This has important implications for how we can define the justice-oriented functions of XAI. While we have associated socio-technical bias with recognition and societal bias with redistribution, this is not a rigid nor exclusive mapping: correcting a socio-technical bias can also lead to questioning the undergirding, systemic conditions that lead to a deviation in the data. For instance, the Amazon hiring case was followed by a public controversy about sexist big tech institutions.[14] Likewise, detecting societal bias can also demand action against the institution that deploy an ADM system when it treats algorithmic outputs as neutral or inherently authoritative. Such detection can and should also create the possibility of deciding not to use the ADM system at all.

[13] See: https://algorithmwatch.org/en/press-release-reporting-form-collect-cases/

[14] See for instance in The Guardian: https://www.theguardian.com/technology/2018/oct/11/tech-gender-problem-amazon-facebook-bias-women

XAI Techniques We have already discussed that logic-based approaches to XAI may offer the needed capability to add knowledge into the explanation. Here, we would like to discuss another consideration: different XAI methods surface different information about the ADM. For example, the SHAP method informs about the most relevant features but not about their encoding in the ADM system. Therefore, SHAP is suitable to indicate socio-technical bias based on the use of a protected category of data but not based on the encoding of a category. A suitable method to learn about the encoding are counterfactual explanations (Wachter et al. 2017): such explanations display data instances similar to the instance in focus (the data subject) but with a different decision. These alternative instances can only change, based on the options that the encoding provides. Thus, for example, if gender is encoded as a binary variable, the alternative data instances can only change along this binary axis, i.e., the alternative gender is male (if the original is female), or female (if the original is male). This leads us to a follow-up question: which explanation methods are best suited for the RR-XAI-framework, and (how) can a combination of explanation methods substitute the need for interpretational work and background knowledge? While addressing this question goes beyond the scope of the present paper, it defines a clear direction for future research: advancing XAI methods that not only surface patterns of bias but also do so in ways that are accessible and actionable for diverse audiences.

Last, *algorithmic recourse* is a field that is often mentioned when discussing actions following XAI explanations (Sullivan and Kasirzadeh 2024). Algorithmic recourse centers on the provision of both an explanation and a recommendation to the data subjects so that they can obtain a favorable decision (Karimi et al. 2020) and builds upon *counterfactual explanations* (Kasirzadeh and Smart 2021). However, the focus is on providing the additional recommendation (the "actionable knowledge") on how to reach the favorable decision.[15] What connects our work with algorithmic recourse is that it relies on explanations to induce actions. However, algorithmic recourse relies on explanations to generate concrete recommendations ("how to act") and that lead to an action of the data subject to change its *own* conditions. Trying to seek change through fostering individual knowledge can, however, also fall into the trap of a (gendered) individualization of responsibilities (Powell et al. 2022; Rohlfing et al. 2020).

Limitations and Future Work

This work faces some limitations and avenues for future work. Frist, we acknowledge that we discussed our framework for two illustrative and fictional cases that take into account only a single dimension of discrimination (i.e. gender). While this focus was due to demonstration purposes, gender needs to be understood in intersection (Crenshaw 2019), by integrating other axes of discrimination.

Second, we have not yet acknowledged the limitations of explanations that build on "approximations" (e.g., surrogates). A specific observation that different XAI explanations can provide different values of faithfulness ("how true an explanation is to the underlying algorithm") for different groups (Balagopalan et al. 2022), introducing potentially additional bias in the explanation. Explanations can also be manipulated so that they appear fairer than they should be (Aïvodji et al. 2019). These limitations need to be taken into account when providing explanations and evaluating their diagnostic value. Sound technical knowledge of the methods but also an honest communication of limitations by developers is essential so expectations towards explanations remain grounded (Sokol and Flach 2020a; Smart and Kasirzadeh 2024).

Much more practical and situated research is necessary to understand the extent to which justice-oriented XAI can be effectively implemented in real-world settings, including identifying obstacles and situations where it may not be applicable. To explore these aspects, empirical studies should focus on examining how AI providers and diverse explanation receivers could actually utilize the RR-XAI-framework.

We acknowledge that this work would benefit from deeper integration of legal expertise to enhance the actionability of the diagnostic functions of XAI methods. This includes not only expertise in anti-discrimination law but also guidance on how to compose explanations in compliance with EU law (Bringas Colmenarejo, State, and Comandé 2025).

Conclusion

In this paper, we introduced the RR-XAI framework to disentangle and operationalize two critical dimensions of bias in algorithmic decision-making: socio-technical and societal bias. Building on Lopez's conceptual distinction and Fraser's theory of justice, we demonstrated how XAI can transcend its role as a transparency tool to become a powerful diagnostic instrument for justice. By framing XAI as both a socio-technical and societal diagnostic, we have demonstrated that explanations could carry significant legal and political weight. At the same time, we underscore the limits of the diagnostic function of current XAI techniques, and how justice-oriented action is contingent on the provision of explanations, resources and existing limits to explanation techniques. Ultimately, we hope this contribution advances the development of justice-oriented XAI techniques that are both technically robust and socially relevant.

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation program under Marie Sklodowska-Curie Actions (g.a. number 860630) for the project "NoBIAS - Artificial Intelligence without Bias". Views and opinions expressed are those of the authors and do not reflect those of the EU. Neither the EU nor the granting authority can be held responsible for them. Additionally, Laura State acknowledges funding by Volkswagen Foundation.

[15] Approaches may also incorporate different constraints, concerning the actionability (changes need to be possible, i.e., the birth data cannot be altered) and the plausibility (the change must result in a realistic data instance) on a proposed recommendation (Karimi et al. 2020).

Author Contributions

Miriam Fahimi and Laura State share first authorship of the paper. Shared (Miriam/Laura): ideation and conceptualization, RR-XAI-framework, first draft, revision/editing. Miriam: discrimination and biases in ADM systems, critical feminist scholarship on XAI, feminist theory of justice (Fraser). Laura: XAI for fairness, examples, figures. Atoosa: problem framing, narrative direction, editorial input.

References

Abdel-Raouf, F.; and Buhler, P. M. 2020. *The Gender Pay Gap: Understanding the Numbers*. Routledge. ISBN 978-1-00-019550-7.

Acred, C. 2016. *Gender Equality?* Independence Educational Publishers. ISBN 978-1-86168-729-6. Google-Books-ID: 5L0ujwEACAAJ.

Aïvodji, U.; Arai, H.; Fortineau, O.; Gambs, S.; Hara, S.; and Tapp, A. 2019. Fairwashing: The risk of rationalization. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 161–170. PMLR.

Alkhatib, A.; and Bernstein, M. 2019. Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–13. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-5970-2.

Allhutter, D.; Cech, F.; Fischer, F.; Grill, G.; and Mager, A. 2020. Algorithmic Profiling of Job Seekers in Austria. How Austerity Politics Are Made Effective. *Frontiers in Big Data*, 3.

Amoore, L. 2020. *Cloud Ethics. Algorithms and the Attributes of Ourselves and Others*. Durham: Duke University Press. ISBN 978-1-4780-0778-4 978-1-4780-0831-6.

Ananny, M.; and Crawford, K. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3): 973–989. Publisher: SAGE Publications.

Andrejevic, M. 2020. Shareable and un-sharable knowledge. *Big Data & Society*, 7(1): 205395172093391.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2018. Machine Bias. *Nieman reports*, 37–. Publisher: Harvard University.

Balagopalan, A.; Zhang, H.; Hamidieh, K.; Hartvigsen, T.; Rudzicz, F.; and Ghassemi, M. 2022. The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. In *FAccT*, 1194–1206. ACM.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Barocas, S.; and Selbst, A. D. 2016. Big Data's Disparate Impact.

Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable Artificial Intelligence (XAI). Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, 58: 82–115.

Beckh, K.; Müller, S.; Jakobs, M.; Toborek, V.; Tan, H.; Fischer, R.; Welke, P.; Houben, S.; and von Rüden, L. 2021. Explainable Machine Learning with Prior Knowledge: An Overview. *CoRR*, abs/2105.10172.

Benjamin, R. 2019. *Race After Technology. Abolitionist Tools for the New Jim Code*. Cambridge: UKPolity. ISBN 978-1-5095-2640-6.

Beretta, A.; Ercoli, G.; Ferraro, A.; Guidotti, R.; Iommi, A.; Mastropietro, A.; Monreale, A.; Rotelli, D.; and Ruggieri, S. 2024. Requirements of eXplainable AI in Algorithmic Hiring. In *AIMMES*, volume 3744 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1): 3–44.

Bono, T.; Croxson, K.; and Giles, A. 2021. Algorithmic fairness in credit scoring. *Oxford Review of Economic Policy*, 37(3): 585–617.

Borch, C.; and Hee Min, B. 2022. Toward a Sociology of Machine Learning Explainability. Human–Machine Interaction in Deep Neural Network-Based Automated Trading. 9(2): 20539517221111361. Publisher: SAGE Publications Ltd.

Bowker, G. C.; and Star, S. L. 2000. *Sorting Things Out. Classification and its Consequences*. Inside technology. Cambridge, Mass. [u.a.]: The MIT Press, 1. mit press paperback ed.. edition. ISBN 978-0-262-02461-7.

Bringas Colmenarejo, A.; State, L.; and Comandé, G. 2025. How should an explanation be? A mapping of technical and legal desiderata of explanations for machine learning models. *International Review of Law, Computers & Technology*, 1–32.

Browne, J. 2023. AI and Structural Injustice: A Feminist Perspective. In Browne, J.; Cave, S.; Drage, E.; and McInerney, K., eds., *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*, 328–346. Oxford University Press. ISBN 978-0-19-288989-8.

Burrell, J. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. 3(1): 2053951715622512. Publisher: SAGE Publications Ltd.

Calegari, R.; Ciatto, G.; and Omicini, A. 2020. On the integration of symbolic and sub-symbolic techniques for XAI: A survey. *Intelligenza Artificiale*, 14(1): 7–32.

Cesario Alvim Gomes, J. 2018. Nancy Fraser's tridimensional approach to justice: Contributions and provocations to the practice of domestic litigators. 16(3): 1021–1024.

Chu, C. Y. C.; Hsu, P.-H.; and Wang, Y.-T. 2023. The gender gap in the ownership of promising land. *Proceedings of the National Academy of Sciences*, 120(24): e2300189120. Publisher: Proceedings of the National Academy of Sciences.

Chun, W. H. K. 2021. *Discriminating Data. Correlation, Neighborhoods, and the New Politics of Recognition*. Cambridge, MA, USA: MIT Press. ISBN 978-0-262-04622-0.

Confalonieri, R.; Coba, L.; Wagner, B.; and Besold, T. R. 2021. A historical perspective of explainable Artificial Intelligence. *WIREs Data Mining Knowl. Discov.*, 11(1).

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and 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*, KDD '17, 797–806. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-4887-4.

Crenshaw, K. 2019. *On Intersectionality. Essential Writings*. New York: New Press. ISBN 978-1-62097-270-0.

Dastin, J. 2018. Amazon scraps secret AI recruiting tool that showed bias against women | Reuters.

DeVito, M. A.; Gergle, D.; and Birnholtz, J. 2017. "Algorithms ruin everything": #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, 3163–3174. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-4655-9.

Eubanks, V. 2017. *Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press, first edition edition. ISBN 978-1-250-07431-7.

European Union Agency for Fundamental Rights. 2018. Handbook on European non-discrimination law.

Fabris, A.; Messina, S.; Silvello, G.; and Susto, G. A. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6): 2074–2152.

Fahimi, M.; and Kinder-Kurlanda, K. 2025. Friction in the Materialities of Value. Relating Transparency, Algorithms and Credit Scoring. In Burkhardt, M.; Seitz, T.; Ochs, C.; and Kropf, J., eds., *Frictions: Conflicts, controversies and design alternatives in digital valuation*, number vol. 9, issue 2 (2023) in Digital Culture & Society, 141–159. Bielefeld: transcript. ISBN 978-3-8376-6358-7.

Felzmann, H.; Villaronga, E. F.; Lutz, C.; and Tamò-Larrieux, A. 2019. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, 6(1): 2053951719860542. Publisher: SAGE Publications Ltd.

Florez-Lopez, R. 2010. Effects of missing data in credit risk scoring. A comparative analysis of methods to achieve robustness in the absence of sufficient data. *Journal of the Operational Research Society*, 61(3): 486–501. Publisher: Taylor & Francis _eprint: https://doi.org/10.1057/jors.2009.66.

Fraser, N. 1995. From Redistribution to Recognition? Dilemmas of Justice in a 'Post-Socialist' Age. *New Left Review*, (I/212): 68–93.

Fraser, N. 1997. *Justice interruptus: critical reflections on the "postsocialist" condition*. New York London: Routledge,Taylor & Francis Group. ISBN 978-0-415-91795-7 978-0-415-91794-0.

Fraser, N. 2000. Rethinking Recognition. *New Left Review*, (3): 107–120.

Freeman, R.; Shah, N.; and Vaish, R. 2020. Best of Both Worlds: Ex-Ante and Ex-Post Fairness in Resource Allocation. In *Proceedings of the 21st ACM Conference on Economics and Computation*, EC '20, 21–22. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-7975-5.

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press. ISBN 978-0-19-170684-4.

Friedman, B.; and Nissenbaum, H. 1996. Bias in Computer Systems. *ACM Transactions on Office Information Systems*, 14(3): 330–347.

Goldsmith-Pinkham, P.; and Shue, K. 2022. THE GENDER GAP IN HOUSING RETURNS.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Pedreschi, D.; and Giannotti, F. 2018. A Survey Of Methods For Explaining Black Box Models. arXiv:1802.01933.

Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; and Yang, G. 2019. XAI - Explainable artificial intelligence. *Sci. Robotics*, 4(37).

Hancox-Li, L.; and Kumar, I. E. 2021. Epistemic Values in Feature Importance Methods. Lessons from Feminist Epistemology. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 817–826. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8309-7.

Hildebrandt, M. 2019. The Issue of Bias. The Framing Powers of Machine Learning.

Hu, L.; and Kohler-Hausmann, I. 2020. What's Sex Got to Do With Fair Machine Learning? 11.

Huang, L. T.-L.; Chen, H.-Y.; Lin, Y.-T.; Huang, T.-R.; and Hung, T.-W. 2022. Ameliorating Algorithmic Bias, or Why Explainable AI Needs Feminist Philosophy. 8(3). Publisher: University of Western Ontario.

Jacobs, A. Z.; and Wallach, H. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 375–385. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8309-7.

Karimi, A.; Barthe, G.; Schölkopf, B.; and Valera, I. 2020. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *CoRR*, abs/2010.04050.

Kasirzadeh, A. 2022. Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 349–356. ArXiv:2206.00945 [cs].

Kasirzadeh, A.; and Klein, C. 2021. The ethical gravity thesis: Marrian levels and the persistence of bias in automated decision-making systems. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 618–626.

Kasirzadeh, A.; and Smart, A. 2021. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 228–236.

Keyes, O.; and Creel, K. 2022. Artificial Knowing Otherwise. *Feminist Philosophy Quarterly*, 8(3/4). Number: 3/4.

Keyes, O.; Hitzig, Z.; and Blell, M. 2021. Truth from the machine: artificial intelligence and the materialization of identity. *Interdisciplinary Science Reviews*, 46(1-2): 158–175. Publisher: SAGE Publications.

Klumbyte, G.; Piehl, H.; and Draude, C. 2023a. Explaining the ghosts: Feminist intersectional XAI and cartography as methods to account for invisible labour. ArXiv:2305.03376 [cs].

Klumbyte, G.; Piehl, H.; and Draude, C. 2023b. Towards Feminist Intersectional XAI: From Explainability to Response-Ability. ArXiv:2305.03375 [cs].

Lipton, Z. C. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10): 36–43.

Lopez, P. 2021. Bias does not equal Bias. A socio-technical Typology of Bias in Data-Based Algorithmic Systems. *Internet Policy Review*, 10(4).

Lundberg, S. M.; and Lee, S. 2017. A Unified Approach to Interpreting Model Predictions. In *NIPS*, 4765–4774.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6): 115:1–115:35.

Miceli, M.; Posada, J.; and Yang, T. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP): 34:1–34:14.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267: 1–38.

Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; and Lum, K. 2021. Algorithmic Fairness. Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8(1): 141–163.

Mittelstadt, B. D.; Russell, C.; and Wachter, S. 2019. Explaining Explanations in AI. In *FAT*, 279–288. ACM.

Molnar, C. 2019. *Interpretable Machine Learning*.

Narayanan, A.; and Kapoor, S. 2024. *AI snake oil: what artificial intelligence can do, what it can't, and how to tell the difference*. Princeton Oxford: Princeton University Press. ISBN 978-0-691-24913-1 978-0-691-24964-3.

Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; and Seifert, C. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.*, 55(13s): 295:1–295:42.

Noble, S. U. 2018. *Algorithms of Oppression. How Search Engines Reinforce Racism*. New York: New York University Press. ISBN 978-1-4798-4994-9 978-1-4798-3724-3.

Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdl, W.; Vidal, M.-E.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; Kompatsiaris, I.; Kinder-Kurlanda, K.; Wagner, C.; Karimi, F.; Fernandez, M.; Alani, H.; Berendt, B.; Kruegel, T.; Heinze, C.; Broelemann, K.; Kasneci, G.; Tiropanis, T.; and Staab, S. 2020. Bias in Data-Driven Artificial Intelligence Systems—An Introductory Survey. *WIREs Data Mining and Knowledge Discovery*, 10(3): e1356. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1356.

Onay, C.; and Öztürk, E. 2018. A review of credit scoring research in the age of Big Data. *Journal of Financial Regulation and Compliance*, 26(3): 382–405. Publisher: Emerald Publishing Limited.

Parliament, T. E.; and the Council. 2006. Directive 2006/54/EC of the European Parliament and of the Council.

Pedreschi, D.; Giannotti, F.; Guidotti, R.; Monreale, A.; Ruggieri, S.; and Turini, F. 2019. Meaningful Explanations of Black Box AI Decision Systems. In *AAAI*, 9780–9784. AAAI Press.

Powell, A. B.; Ustek-Spilda, F.; Lehuedé, S.; and Shklovski, I. 2022. Addressing Ethical Gaps in 'Technology for Good'. Foregrounding Care and Capabilities. *Big Data & Society*, 9(2): 20539517221113774. Publisher: SAGE Publications Ltd.

Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 469–481. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6936-7.

Ramachandranpillai, R.; Baeza-Yates, R.; and Heintz, F. 2023. FairXAI - A Taxonomy and Framework for Fairness and Explainability Synergy in Machine Learning.

Reyero Lobo, P.; Kwarteng, J.; Russo, M.; Fahimi, M.; Scott, K.; Ferrara, A.; Sen, I.; and Fernandez, M. 2024. A Multidisciplinary Lens of Bias in Hate Speech. In *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '23, 121–125. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0409-3.

Rohlfing, K. J.; Cimiano, P.; Scharlau, I.; Matzner, T.; Buhl, H. M.; Buschmeier, H.; Esposito, E.; Grimminger, A.; Hammer, B.; Häb-Umbach, R.; Horwath, I.; Hüllermeier, E.; Kern, F.; Kopp, S.; Thommes, K.; Ngomo, A.-C. N.; Schulte, C.; Wachsmuth, H.; Wagner, P.; and Wrede, B. 2020. Explanation as a Social Practice. Toward a Conceptual Framework for the Social Design of AI Systems. 1–1. Conference Name: IEEE Transactions on Cognitive and Developmental Systems.

Ruggieri, S.; Alvarez, J. M.; Pugnana, A.; State, L.; and Turini, F. 2024. Can We Trust Fair-AI? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13): 15421–15430.

Scheuerman, M. K.; Pape, M.; and Hanna, A. 2021. Auto-Essentialization. Gender in Automated Facial Analysis as Extended Colonial Project. *Big Data & Society*, 8(2): 20539517211053712. Publisher: SAGE Publications Ltd.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 59–68. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6125-5.

Sloane, M.; Solano-Kamaiko, I.; Yuan, J.; Dasgupta, A.; and Stoyanovich, J. 2023. Introducing contextual transparency for automated decision systems. *Nat. Mac. Intell.*, 5(3): 187–195.

Smart, A.; and Kasirzadeh, A. 2024. Beyond model interpretability: Socio-structural explanations in machine learning. *AI & SOCIETY*, 1–9.

Sokol, K.; and Flach, P. A. 2020a. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *FAT**, 56–67. ACM.

Sokol, K.; and Flach, P. A. 2020b. One Explanation Does Not Fit All. *Künstliche Intell.*, 34(2): 235–250.

Staff, T. N. Y. T. E. 2018. *The Gender Pay Gap: Equal Work, Unequal Pay*. The Rosen Publishing Group, Inc. ISBN 978-1-64282-117-8. Google-Books-ID: pvGCDwAAQBAJ.

State, L. 2021. Logic Programming for XAI: A Technical Perspective. In *ICLP Workshops*, volume 2970 of *CEUR Workshop Proceedings*. CEUR-WS.org.

State, L.; and Fahimi, M. 2023. Careful Explanations: A Feminist Perspective on XAI. In *EWAF*, volume 3442 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Stone, K. L. 2022. *Panes of the Glass Ceiling: The Unspoken Beliefs Behind the Law's Failure to Help Women Achieve Professional Parity*. Cambridge: Cambridge University Press. ISBN 978-1-108-42759-3.

Sullivan, E.; and Kasirzadeh, A. 2024. Explanation Hacking: The perils of algorithmic recourse. *arXiv preprint arXiv:2406.11843*.

Tschantz, M. C. 2022. What is Proxy Discrimination? In *FAccT*, 1993–2003. ACM.

Turesky, M.; and Warner, M. E. 2020. Gender Dynamics in the Planning Workplace: The Importance of Women in Management. *Journal of the American Planning Association*, 86(2): 157–170.

van Bekkum, M.; and Borgesius, F. Z. 2021. Digital Welfare Fraud Detection and the Dutch SyRI Judgment. *European Journal of Social Security*, 23(4): 323–340. Publisher: SAGE Publications Ltd.

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, 1–7. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-5746-3.

Wachter, S.; Mittelstadt, B. D.; and Russell, C. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Comput. Law Secur. Rev.*, 41: 105567.

Wachter, S.; et al. 2017. Counterfactual Explanations without Opening the Black Box. *Harv. JL & Tech.*, 31: 841.

Wajcman, J.; and Young, E. 2023. Feminism Confronts AI: The Gender Relations of Digitalisation. In Browne, J.; Cave, S.; Drage, E.; and McInerney, K., eds., *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*, 47–64. Oxford University Press. ISBN 978-0-19-288989-8.

Weerts, H.; Xenidis, R.; Tarissan, F.; Olsen, H. P.; and Pechenizkiy, M. 2024. The Neutrality Fallacy: When Algorithmic Fairness Interventions are (Not) Positive Action. ArXiv:2404.12143 [cs].

Weerts, H. J. P.; Xenidis, R.; Tarissan, F.; Olsen, H. P.; and Pechenizkiy, M. 2023. Algorithmic Unfairness through the Lens of EU Non-Discrimination Law: Or Why the Law is not a Decision Tree. In *FAccT*, 805–816. ACM.

Wellner, G.; and Rothman, T. 2020. Feminist AI. Can We Expect Our AI Systems to Become Feminist? *Philosophy & Technology*, 33(2): 191–205.

Yarger, L.; Cobb Payton, F.; and Neupane, B. 2019. Algorithmic equity in the hiring of underrepresented IT job candidates. *Online Information Review*, 44(2): 383–395. Publisher: Emerald Publishing Limited.

Zhou, J.; Chen, F.; and Holzinger, A. 2020. Towards Explainability for AI Fairness. In *xxAI@ICML*, volume 13200 of *Lecture Notes in Computer Science*, 375–386. Springer.