Journal of Artificial Intelligence Research 81 (2024) 1043-1078 Submitted 06/2024; published 12/2024

# Bias Mitigation Methods: Applicability, Legality, and Recommendations for Development

**Madeleine Waller** MADELEINE.WALLER@KCL.AC.UK
*Department of Informatics*
*King's College London, UK*

**Odinaldo Rodrigues** ODINALDO.RODRIGUES@KCL.AC.UK
*Department of Informatics*
*King's College London, UK*

**Michelle Seng Ah Lee** SAL87@CAM.AC.UK
*Department of Computer Science and Technology*
*University of Cambridge, UK*

**Oana Cocarascu** OANA.COCARASCU@KCL.AC.UK
*Department of Informatics*
*King's College London, UK*

## Abstract

As algorithmic decision-making systems (ADMS) are increasingly deployed across various sectors, the importance of research on fairness in Artificial Intelligence (AI) continues to grow. In this paper we highlight a number of significant practical limitations and regulatory compliance issues associated with the application of existing bias mitigation methods to ADMS. We present an example of an algorithmic system used in recruitment to illustrate these limitations. Our analysis of existing methods indicates a pressing need for a change in the approach to the development of new methods. In order to address the limitations, we provide recommendations for key factors to consider in the development of new bias mitigation methods that aim to be effective in real-world scenarios and comply with legal requirements in the European Union, United Kingdom and United States, such as non-discrimination, data protection and sector-specific regulations. Further, we suggest a checklist relating to these recommendations that should be included with the development of new bias mitigation methods.

## 1. Introduction

Artificial Intelligence (AI), particularly using machine learning (ML) techniques, is increasingly being employed in decision-making systems across various domains. Algorithmic decision-making systems (ADMS) typically rely on historical data to make predictions about individuals, which are then used to make decisions that affect them. However, this data may contain biases against certain groups or individuals, e.g., due to replicating historical societal discrimination, not being representative of the population the ADMS is deployed on, or incorrectly measuring features of individuals (Mehrabi et al., 2021), potentially leading to discriminatory decisions. In high-stakes domains such as healthcare, criminal justice, and recruitment, the potential negative impact of these systems can be significant, making it crucial to address and mitigate these biases.

ADMS are used in both private and public sectors with organisations increasingly relying on them to reduce workload and free up resources (Engin & Treleaven, 2019; Deloitte, 2021). Despite the potential advantages, there have been several instances where ADMS were involved in unfair decisions. For example, in the United States, the COMPAS system used to predict the likelihood of a criminal re-offending — influencing decisions on parole and sentencing — was found to incorrectly identify black defendants as more likely to re-offend compared to white defendants (Larson et al., 2016; Northpointe, 2019). In recruitment, Amazon's hiring tool exhibited bias against women due to the under-representation of women in the dataset of previously hired candidates (Dastin, 2022). Further, individuals may not be aware they are being treated unfairly (Wachter et al., 2021), making it challenging for users and developers to fully understand and mitigate the scope and potential harmful effects of these systems. The potential lack of transparency and accountability highlights the need for evaluation and consideration in ensuring these systems do not lead to discrimination.

As ADMS become more prevalent, the field of fairness in AI has seen an influx of literature in recent years (see Dunkelau and Leuschel (2019), Pessach and Shmueli (2023b), Hort et al. (2024) for surveys). Many bias mitigation methods have been proposed, mostly aiming to reduce bias in ADMS according to various fairness metrics. These metrics quantify different notions of fairness and measure the extent of unwanted bias in these systems. The tendency to focus on fairness metrics may risk overlooking the systemic social and legal perspectives on fairness and bias. Ensuring ADMS are fair to individuals and communities is an important cross-disciplinary issue which must consider the context and application of the systems deployed (Waller & Waller, 2020).

In this paper we describe key challenges faced by existing bias mitigation methods for ADMS. We categorise these challenges by the methods' practical applicability in real-world contexts and their misalignment with laws and regulations. Based on the limitations of existing bias mitigation methods for ADMS, we suggest recommendations to help guide the development of more effective ones. This paper is an extension of Waller et al. (2023b) and expands it in a number of ways:

- We provide more background on existing legal frameworks pertaining to non-discrimination, data protection, AI and sector-specific laws in the European Union (EU), United Kingdom (UK) and United States of America (US).
- We present a running example of ADMS used in recruitment to illustrate how regulations could impact the use of existing bias mitigation methods for ADMS.
- We suggest an expanded set of considerations to keep in mind while developing new bias mitigation methods.
- We propose a checklist which includes a series of questions to consider when creating a new bias mitigation method.

The rest of this paper is structured as follows. In Section 2 we provide the background on algorithmic fairness including types of bias mitigation methods, datasets, and metrics used, as well as regulations that pertain to the use of bias mitigation methods for ADMS. In Section 3 we discuss the limitations of existing bias mitigation methods and describe how

these motivate the recommendations made for the creation of new bias mitigation methods for ADMS. In Section 4 we suggest a checklist to be included with the development of new bias mitigation methods, relating to the recommendations made in Section 3. We discuss the legal process of proving discrimination and related work in Section 5, and conclude in Section 6.

We hope that this paper will advance the discussion towards the design of bias mitigation methods that consider the applicability in different use cases as well as the social and legal perspectives on fairness and bias.

## 2. Background

In this section we provide background on algorithmic fairness and legal frameworks relevant for bias mitigation methods for ADMS, and present an example of ADMS deployed in recruitment for use throughout the paper. Non-discrimination laws are prevalent around the world, as one of the fundamental rights encoded in the United Nations' Universal Declaration of Human Rights. For the purpose of this paper, our examples are predominantly from the EU, UK and US. However, the key principles of non-discrimination are broadly applicable to other jurisdictions.

ADMS manifest in various forms. In this paper, we focus on systems that have a binary outcome, i.e. the decision made is either positive or negative. Examples of these systems are prevalent in various fields. In criminal justice, they have been used to predict the likelihood of a criminal re-offending (Northpointe, 2019). In social services, they have been used to identify children who are likely to be at risk of neglect (Shared Intelligence and Local Government Association, 2020). In employment, they have been used to shortlist candidates for job interviews (Dastin, 2022). Bias in ADMS can arise from the dataset used for training and can be due to factors such as the historical embedding of social biases, the prevalence of individuals with particular characteristics in the data, or attribute values being measured or used incorrectly (Mehrabi et al., 2021).

Throughout this paper, we use the example of ADMS in the domain of employment to illustrate bias mitigation methods and applicable laws. Such systems have been used with adverse effects, for example the Amazon recruitment tool which was shown to be biased against women (Dastin, 2022). As a running example, consider an ML model, such as a logistic regression classifier, used to predict whether or not a candidate is a good fit for a job based on their application. The aim is for the system to assist a company in making a decision of whether or not to invite a candidate for an interview, freeing up the potentially large amount of time necessary for a human to sort through all applications (Engin & Treleaven, 2019; Deloitte, 2021). The prediction of whether a candidate is fit for the job according to the model equates to whether a candidate is selected for interview.

The model is trained on application data of previously hired candidates' such as their qualifications, number of years of experience, and education.[1] The classification for a new individual is found given their application data and is either positive, meaning they are likely to be a good fit for a job and therefore are invited for an interview, or negative, meaning they are not likely to be a good fit and are therefore not invited for an interview.

1. Here we use the example of an ML model trained on tabular data. Input data could also be textual, for example data taken from individuals' CVs or online profiles (Sandanayake et al., 2018).

If this system is used, ideally it would have a human-in-the-loop (Therese Enarsson & Naarttijärvi, 2022) where the classification is given to a human (recruiter) to aide in making the final decision. They should have other information about the system and the candidates to make an informed decision. In reality, this information might not be available and automation bias (Skitka et al., 1999) might mean the classifications from the model are taken as the decisions by the recruiter. We will assume this is the case for our running example and any unfairness present in the classifications from the model will translate to discriminatory decisions. The potential impact of making the wrong decision about who to invite for interview could be immense for an individual impacted and the company. Ensuring that these algorithmic decisions are fair is important ethically and legally.

### 2.1 Algorithmic Fairness

Many methods have been developed for the detection and mitigation of bias in binary classification, which is the type of model we focus on in this paper. ADMS based on binary classification provide positive and negative classifications for individuals, which usually translate directly to positive and negative decisions in reality.[2]

Most existing bias methods rely on some definition of fairness, quantified by a fairness metric (Garg et al., 2020). These metrics are with respect to personal protected characteristics which are the features outlined in law (UK Public General Acts, 2010; The United States Department of Justice, 2015) and include race, sex and religion (Bellamy et al., 2018). One approach to bias mitigation is *fairness through unawareness* (Grgic-Hlaca et al., 2016) which removes protected attributes from the training data. This prevents them from having any direct influence on the decision (Jorgensen et al., 2023b). However in reality, there are also many characteristics which are proxies for protected characteristics, meaning they might correlate (Wiggins, 2020). These are called sensitive attributes (Van Nuenen et al., 2020; Quy et al., 2022) and should also be carefully considered in any high-stake decision-making (Pessach & Shmueli, 2023a).

The discussions around protected attributes in this paper can be transferred to sensitive attributes if they are defined well in a particular context. In our recruitment example, the attribute 'years of experience' would likely correlate to the age of a candidate. Using this attribute to train the model could cause discrimination against candidates based on age and therefore, for the purpose of mitigating bias in ADMS, should be considered in a similar way to the protected attribute.

#### 2.1.1 Fairness Metrics

Fairness metrics quantify bias according to some defined notion of what it means to be fair with inspiration from philosophy, sociology, law (Blanchard, 1986; Verma & Rubin, 2018; Xiang & Raji, 2019). The metrics fall into two groups: *group fairness* and *individual fairness*. Group fairness metrics aim to ensure different groups of individuals receive similar positive and negative classifications across all values of a protected personal characteristic (Chakraborty, Peng, & Menzies, 2020) — a *protected attribute* in the algorithmic fairness

2. Although the decisions may appear to be immediately positive or negative, for example hiring an individual for the job, if this decision is incorrect it could have a worse negative impact on an individual in the long term (Weinberg, 2022; Jorgensen et al., 2023a).

literature. An individual is said to be in the *unprivileged group* if the value of their protected attribute defines them in the historically disadvantaged group, otherwise the individual is said to be in the *privileged group*. Individual fairness metrics seek to guarantee that individuals with similar attribute values receive the same classification (Dwork et al., 2012).

**Group fairness metrics** The most commonly used group fairness metrics are *demographic parity* and *equal opportunity* (Hort et al., 2024). Demographic parity measures the difference in proportion of positive classifications between the privileged and unprivileged groups, not being concerned with the true label of the individual. In our recruitment example, for the protected attribute 'sex',[3] demographic parity quantifies the difference in the percentage of male candidates and the percentage of female candidates who get selected for interview. *Disparate impact* (Feldman et al., 2015) quantifies the same notion of fairness except that instead of finding the difference between the proportion of positive classifications across groups (taking one away from the other), it divides the proportions such that we get the proportion of female candidates who get selected for interview divided by the proportion of male candidates who get selected for interview. This gives a percentage representing the level of disparate impact in the system.

Equal opportunity (Hardt et al., 2016) measures the difference in the true positive rates between the privileged and unprivileged groups, where the true positive rate is the number of correct positive classifications out of the total positive classifications. In our recruitment example, this metric calculates the difference in the proportion of male candidates correctly selected for interview out of all male candidates selected and the proportion of female candidates correctly selected out of all female candidates selected.

*Equalised odds* (Hardt et al., 2016) strengthens equal opportunity by additionally measuring the false positive rates (number of incorrect positive classifications out of the total positive classifications). The false positive rate for our recruitment example would be the proportion of candidates selected for interview even when they are not fit for the job.

*Conditional demographic parity* (Wachter et al., 2021)[4] quantifies the difference in the proportion of positive classifications for different protected groups, conditional on a legitimate characteristic. What is 'legitimate' depends on the context and domain and would need to be pre-defined. Determining legitimate characteristics can be challenging, especially where proxies of protected characteristics are intertwined with proxies of the decisions (Pessach & Shmueli, 2023a). In our recruitment example, conditional demographic parity could mean measuring the proportion of male and female candidates selected for interview given they have the same level of education such as a masters degree in a relevant subject.

**Individual fairness metrics** Individual fairness metrics are less frequently used than group fairness metrics (Hort et al., 2024). They quantify the notion that similar individuals should be treated similarly (Chakraborty et al., 2020). For example, *fairness through awareness* quantifies the number of pairs of similar individuals receiving the same classification, where the definition of a similar individual is required as input and depends on the

3. We use the attribute 'sex' throughout the paper as the binary 'male' or 'female' to illustrate the disparities across two groups. This is a simplification often used throughout algorithmic fairness literature, and is not the same as the attribute 'gender' which considers all gender identities. The disadvantages of viewing fairness across only two groups are discussed in Section 3.1.
4. The illegal discrimination metric (Kamiran et al., 2013; Zliobaite et al., 2011) also corresponds to the fairness notion defined by conditional demographic disparity.

context (Dwork et al., 2012). Additionally, the *consistency* metric defines the similarity of individuals by considering its nearest neighbours according to the Manhattan distance and providing a score based on the number of neighbours with different classifications (Zemel et al., 2013). Different numbers of similar individuals (or nearest neighbours) can be considered and should be specified. However, in reality the meaning of this number has not been explored thoroughly (Waller et al., 2024).

*Counterfactual fairness* (Kusner et al., 2017) defines a different notion of individual fairness which specifies that an individual is treated fairly if they receive the same classification given they have any value of a protected attribute. In our recruitment example, if a candidate was not selected for interview when the value of the attribute 'sex' was either 'male' or 'female', then this would be considered counterfactually fair. Ensuring counterfactual fairness for all individuals is similar to the notion of demographic parity (Rosenblatt & Witter, 2023). Counterfactual fairness is different to fairness through unawareness as it allows us to potentially detect bias for individuals and mitigate it. Other methods focus on why an individual has been classified differently to similar individuals (Chakraborty et al., 2020; Waller et al., 2024).

#### 2.1.2 Bias Mitigation Methods

Bias mitigation methods aim to reduce bias, consequently improving the outcome measured by fairness metrics. Methods are generally split into three categories: *pre-processing* methods which modify the training data to reduce bias before training the model (e.g., Calders et al., 2009; Kamiran & Calders, 2011; Zliobaite et al., 2011; Feldman et al., 2015; Iosifidis & Ntoutsi, 2018), *in-processing* methods which modify the model to mitigate bias during training, e.g., by optimising for additional constraints that promote fairness (e.g., Calders & Verwer, 2010; Kamiran et al., 2010; Krasanakis et al., 2018; Grari et al., 2019; Iosifidis & Ntoutsi, 2019; Oneto et al., 2019; Hu et al., 2020), and *post-processing* methods which mitigate bias in the model's output, i.e., neither the model nor the data are changed, but the outputs of the model may be adjusted to improve fairness in the decisions made (e.g., Kamiran et al., 2010, 2012; Fish et al., 2016; Lohia et al., 2019).

Not all ADMS are based on binary classification. Fairness methods for other forms of ADMS include ones designed to improve fairness in regression (Calders et al., 2013; Berk et al., 2017; Chzhen et al., 2020), multi-class classification (Alghamdi et al., 2022; Putzel & Lee, 2022), clustering (Chierichetti et al., 2017; Rösner & Schmidt, 2018; Backurs et al., 2019; Bera et al., 2019; Abbasi et al., 2021; Ziko et al., 2021), outlier detection (P & Abraham, 2020), clustering from demonstrations (Galhotra et al., 2021), in online data streams (Zhang & Ntoutsi, 2019; Iosifidis & Ntoutsi, 2020), and where there is little data available (Slack et al., 2020). Many of the legal recommendations made in Section 3.2 are applicable in the creation of fairness methods for other ADMS.

### 2.2 Relevant Laws

The use of existing bias mitigation methods for ADMS will be subject to existing laws in the jurisdiction the ADMS are deployed in (and perhaps also designed and developed in). We focus on legal frameworks in the EU, UK and US due to the alignment of the values of the rule

of law, human rights, and civil liberties, amongst others.[5] These jurisdictions have firmly established equality and data protection laws, providing a solid foundation for exploring their potential impact on the legality of technical bias mitigation methods. It is important to note the distinction between laws and regulations, where laws are created by a legislative body to establish general frameworks and principles, while regulations offer specific guidelines and rules to implement and enforce these legal principles (Morgan & Yeung, 2007). Henceforth, we will specifically refer to the regulations in each jurisdiction categorised into two main groups: cross-cutting regulation and sector-specific regulation (Lawrence-Archer & Naik, 2023).

#### 2.2.1 Cross-cutting Regulation

Cross-cutting regulation encompasses guidelines that apply universally across any sector and domain. There are many cross-cutting regulations that might impact development and deployment of AI systems (Xenidis, 2023). There is a debate on which cross-cutting approaches are best for the future of AI governance, with a distinction between strengthening existing regulation (e.g., non-discrimination, data protection) or developing new technology-specific regulations that govern the particular harms that might arise from its use. For example, the European Commission outlined its plans that "EU legislation remains in principle fully applicable irrespective of the involvement of AI". However, the newly adopted EU AI Act (Edwards, 2021; European Commission, 2021) takes a very different approach of defining the technologies that are being used and the risks that each one possesses in different use cases.

In the remaining of this section we give background on regulations that have the potential to impact the use of technical bias mitigation methods used to improve the fairness of ADMS, noting that the regulations discussed here may not be the only ones applicable.

**Anti-discrimination** In the Universal Declaration of Human Rights (United Nations. General Assembly, 1949), one of the basic human rights is "equality and freedom from discrimination". This right has taken form in many regulations across different jurisdictions. For example, Article 2 of the EU's Charter of Fundamental Rights (European Union Agency for Fundamental Rights, 2007) states that

> "Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited."

The UK Equality Act 2010 (UK Public General Acts, 2010), is a regulation in the UK that exemplifies this Charter and defines types of discrimination. *Direct discrimination* is when an individual is treated less favourably because they have, are perceived to have, or are connected to someone with a protected characteristic. This would require the decision-making process to have knowledge of the protected characteristics. *Indirect discrimination* is when a policy negatively impacts an individual with a protected characteristic. This does not require knowledge of the protected characteristics and the discrimination could be due

5. Many of the considerations discussed in this paper will also be applicable to ADMS deployed in other democratic countries.

to the decisions correlating with protected characteristics (UK Public General Acts, 2010). Most non-discrimination regulations in the EU are similar in defining direct and indirect discrimination (European Commission, 2022), in relation to protected characteristics.

Other definitions in the UK Equality Act include *positive action* and *positive discrimination*. The first involves taking action to treat a "group that shares a protected characteristic more favourably than others, where this is a proportionate way to enable or encourage members of that group to: overcome or minimise a disadvantage, have their different needs met, participate in a particular activity" (UK Government, 2023). This is a legal process which attempts to address historical discrimination or imbalanced representation in a decision-making process. Positive action is voluntary and not required. Positive discrimination however occurs when a group receives preferential treatment based on a protected characteristic, without meeting the criteria specified for positive action. As a consequence, this could also entail direct discrimination against the privileged group.

In the United States, the Civil Rights Act 1964 (U.S. Department of Labor, 1964) sets out the broad principles of enforcing equality, making it illegal to discriminate based on the protected characteristics of "race, colour, religion, sex, and national origin". The assessment of discrimination can be accomplished using two approaches: *disparate treatment* and *disparate impact*. Disparate treatment (Civil Rights Division, U.S. Department of Justice, 2016a) involves examining whether a particular individual or group has been treated less favourably than others based on protected characteristics, driven by discriminatory intent. In contrast, disparate impact (Willborn, 1984; Civil Rights Division, U.S. Department of Justice, 2016b) assesses whether neutral practices (with no discriminatory intent) have been implemented that result in significant adverse effects on protected groups. These notions are similar to direct and indirect discrimination, however they focus more on intent and statistical measures across whole populations.

Unlike in the EU, positive discrimination in the US (known as *affirmative action*) is legal in some circumstances (Archibong & Sharps, 2013). It requires knowledge of individuals' protected characteristics to be able to promote opportunities amongst the unprivileged groups. This could be done using quotas (e.g., aiming to select 50% of female candidates for interview) or relaxation of selection criteria for individuals in the unprivileged group (Sabbagh, 2011). Affirmative action is widely debated, exemplified by the US Supreme Court deciding in 2023 to make it illegal to collect race data as part of the admissions process for University meaning it is no longer possible to positively discriminate based on an applicant's race (Alan et al., 2024).

**Data Protection** Data protection regulations govern data, the automated processing of data and the sharing of data. These regulations cut across different sectors and domains and are not specific to the technology being used.

One example is the General Data Protection Regulation (GDPR, 2016) which sets out comprehensive data protection law in the EU that outlines the legal collection, processing, and storage of personal data in EU member states and in the UK, as well as organisations operating outside of the EU working with data from EU organisations or individuals (Gil González & De Hert, 2019). The GDPR requires organisations to obtain explicit consent for data processing, minimise data collection, ensure data security, and grant individuals rights to access and control their personal information (GDPR, 2016). The storage

and use of personal data, in particular the use of personal characteristics (or 'special categories of personal data') in ADMS, are highly safeguarded under GDPR.

Under GDPR, there are also specific regulations around solely automated decisions or ones that have legal or similarly significant effects (i.e. decisions made with no human-in-the-loop), whilst still having to abide by the regulations for collecting and storing personal data. Article 22(3) states:

> "In case of necessity for a contract or consent, the controller must guarantee, at least, the following to the subject: the right to obtain human intervention, the right of the individual to express his or her point of view and the right to contest the decision."

However, as discussed in Section 2, a solely automated decision could be one where the human simply relies on the output from an automated system, which is a weakness of the regulations pertaining to these systems (Gil González & De Hert, 2019). Our recruitment example is still likely to fall under these regulations due to the significant effects of the decisions made. Articles 13 and 14 state that when automated decision-making (or 'profiling') is deployed, an individual has the right to "meaningful information about the logic involved" (GDPR, 2016; Hamon et al., 2022). Whether this sufficiently enforces a right to an explanation is debatable (Wachter et al., 2017). However, it is clear that current data protection regulation promotes transparency and explainability (Malgieri, 2019).

In the US, data protection law is primarily outlined by a mix of federal and state regulations. Federal regulations focus more on providing data protection for specific domains such as health (Edemekong et al., 2024), financial (Gramm, 1999), and children's data (Federal Trade Commission et al., 1998). There is no comprehensive federal law similar to the GDPR, however some states have passed their own laws, granting consumers rights over their data, for example in the state of California. The California Consumer Privacy Act (CCPA) enforces many of the same rights as GDPR such as the right to access personal data, the right to opt-out, the right to be informed whether personal data is being used (State of California Department of Justice, 2024). However, it is less strict than GDPR as it falls short of enforcing the right to be forgotten (right to have personal data removed), the right not to be subject to decisions from ADMS and the right to an explanation of decisions made by ADMS (Pardau, 2018; Rothstein & Tovino, 2019).

**AI Specific** The emergence of AI specific regulations, such as the EU AI Act (European Commission, 2021), reflects a growing recognition of the need to regulate the rapidly advancing technology. At the heart of these legislative efforts lies the challenge of defining AI itself. The EU AI Act proposes a framework that classifies use cases of AI systems into different risk categories based on their potential impact on fundamental rights and societal values. This approach could potentially lead to challenges when new applications of AI systems emerge, as they will need to be classified into appropriate categories. This classification process could spark debates regarding the placement of certain use cases within specific categories (Bosoer et al., 2023). ADMS such as our recruitment example and other high stakes applications are considered high risk by Articles 9–15 of the Act and thus are subject to particular requirements (Bosoer et al., 2023). These articles embed the key principles of traceability and explainability to produce technical documentation, have detailed

transparency provisions, ensure human supervision and maintain standards for accuracy and cybersecurity, amongst other guidelines (European Commission, 2021). The requirements do not directly enforce non-discrimination but ensure that providers of high-risk ADMS are required to demonstrate the processes involved in ensuring the decisions are non-discriminatory (Bosoer et al., 2023).

#### 2.2.2 Sector-specific Regulation

Sector-specific regulation tailors guidelines to address the unique challenges and considerations within specific industries or domains. Here, we focus on the employment, finance and public sector domains as examples but there are many other domain-specific laws that could apply to systems used within the industries.

**Employment** Non-discrimination in employment is largely covered by existing equality law. For example, the UK Equality Act 2010 (UK Public General Acts, 2010), and other similar EU Equality Acts, were specifically designed with reference to actions in the workplace such as recruitment. Since the enactment of the US Civil Rights Act of 1964 (U.S. Department of Labor, 1964), other acts have been passed to regulate more nuanced forms of discrimination such as age (U.S. Equal Employment Opportunity Commission, 1967), disability (U.S. Department of Justice Civil Rights Division, 2008) and pregnancy (U.S. Equal Employment Opportunity Commission, 1978), all of which have special considerations around affirmative action, for example relaxing the constraint that individuals have to meet certain qualifications to get the job.

**Finance** Non-discrimination laws in finance serve as safeguards to ensure fair and equitable treatment for all individuals accessing financial services and products. Globally, empirical evidence indicates the presence of financial discrimination, including in the US and the EU, especially on the basis of ethnicity and race (Stănescu & Gikay, 2020). Therefore, some countries have implemented laws and regulations specific to prohibiting discrimination and ensuring fair access to financial products and services. In the US, there are sector-specific Fair Lending Laws, such as the Equal Credit Opportunity Act and the Fair Housing Act, that prohibit discrimination in credit transactions, including transactions related to residential real estate. By contrast, in the UK and the EU, non-discrimination laws apply broadly, including in financial services, with additional requirements put forward by relevant regulators (Stănescu & Gikay, 2020). In the UK, the relevant regulations in financial services include (Powley & Stanton, 2020): consumer duty (laying out clear protection of the consumers), vulnerable customers (requirement to provide additional support to those exhibiting vulnerability), and Senior Managers and Certification Regime (SMCR) (accountability of leaders).

**Public-sector** Fairness and trust is crucial in public sector services as many of the services will be high impact (Waller & Waller, 2020). Extra provisions are often specified for ADMS deployed in the public sector. For example, under the UK's Public Sector Equality Duty (Equality and Human Rights Commission, 2023), part of the UK Equality Act, public sector services are held to a higher standard such that they must eliminate "unlawful discrimination, harassment, and victimisation" and "advance equality of opportunity between people who share a protected characteristic and those who do not" (Equality and Human

Rights Commission, 2023). Decision-making systems in the public-sector are therefore not only required to mitigate bias but actively promote equality and the reversal of systemic discrimination.

### 2.3 Related Work

In recent years, there has been a growing interest in the development of bias mitigation methods, and as a result, there have been several surveys conducted (Dunkelau & Leuschel, 2019; Pagano et al., 2023; Pessach & Shmueli, 2023b; Hort et al., 2024). These provide a detailed background to bias mitigation, including metrics and datasets used throughout the algorithmic fairness literature from a predominately technical perspective. For example, Hort et al. (2024) present a comprehensive survey of bias mitigation methods for machine learning which categorises methods from 234 papers into pre-, in-, and post-processing, and further by the approach they take. Moreover, they provide statistics about the methods in terms of datasets and metrics used. These findings of existing methods gives evidence for the claims in our paper that the recommendations we provide are based on the limitations of the majority of bias mitigation methods.

An overview of the main bias mitigation methods up to 2019 is given in Dunkelau and Leuschel (2019). These include Calders et al. (2009), Kamiran and Calders (2010), Kamishima et al. (2012), Zafar et al. (2017). Dunkelau and Leuschel (2019) discuss ethical concerns related to using protected attributes in ADMS and comment on the need for transparency and accountability, which are often overlooked in algorithmic fairness. Further, they suggest that new fairness metrics should only be introduced if they are "fundamentally different" to existing ones. This broadly aligns with our applicability recommendation $R_a3$, which recommends to be able to justify why metrics have been chosen and what real-world notions of fairness they correspond to. Corbett-Davies et al. (2024) also critique statistical-based methods such as ones that use existing group fairness methods showing that they often "harm the very groups they were designed to protect" and emphasise that we cannot formalise fairness as a metric such that it can be universally applied.

Other related works focus on the development of frameworks to be able to more easily evaluate bias mitigation methods on different models and datasets, and using different fairness metrics. These frameworks are described in Dunkelau and Leuschel (2019). Lee and Singh (2021) examine fairness toolkits, by presenting a mixed methods study which found that the toolkits represent each developer's perspective on what fairness means, which may or may not be aligned to that of the toolkit user. In this study, one participant raised the issue that some of the toolkits could result in positive discrimination in the employment sector due to the methods implemented using protected characteristics to give preferential treatment (positive discrimination) which, depending on the context, could be illegal. This demonstrates the need for toolkits to be accompanied by disclaimers and explanations on the potential applicability and legality limitations of the methods and metrics they contain.

There is existing work on creating guidance for which scenarios to deploy different fairness metrics/methods — these usually focus on which metrics to use based on the priorities of balancing different positive and negative rates. For example, Saleiro et al. (2019) propose a decision tree which helps select the fairness metric(s) that are relevant in each context. This is motivated by feedback from data scientists and policymakers that

used their toolkit: "so many different metrics and there was no guidance in how to link the different metrics and the real-world problem in hand".

## 3. Recommendations for Bias Mitigation Methods

In this section, we discuss the limitations of bias mitigation methods that are hurdles to their practical applicability in real-world scenarios, as identified throughout the literature (Dunkelau & Leuschel, 2019; Saleiro et al., 2019; Waller et al., 2023a; Corbett-Davies et al., 2024; Hort et al., 2024). We also consider the laws and regulations that may limit the usefulness of these methods in the real-world.

Whilst the development of bias mitigation methods has enriched the landscape on fairness in AI research, it remains unclear whether these methods are being used in any real-world scenarios, and further, whether they can actually be effectively and legally deployed. To ensure that new methods can be employed in practice in real-world applications, we make the following recommendations based on the limitations we collated.[6]

### 3.1 Applicability Recommendations

Bias mitigation methods have been proposed for a variety of problems: attribute effect in linear regression (e.g., Calders et al., 2013), fairness in multi-class classification (e.g., Putzel & Lee, 2022), fair clustering (e.g., Chierichetti et al., 2017), etc. Still, the majority of works focus on binary classification (Waller et al., 2023a); they make assumptions about the specific scenarios in which the proposed bias mitigation method can be applied, then design and evaluate it. There is an applicability issue as existing methods can only be applied in the context of specific dataset characteristics, models, and metrics.

#### 3.1.1 Protected Attributes

Most bias mitigation methods for binary classification optimise for some group fairness metric (Hort et al., 2024) — see Section 2. These group fairness metrics rely on the protected attribute being pre-specified and of a particular type. Specifically, most require a single pre-specified protected attribute, for example the protected attribute of 'sex' taking values 'male' or 'female'. If we assume that the protected attributes are available for individuals in the data (which may not always be the case — see Section 3.2), applying most existing bias mitigation methods requires choosing for which fairness metric to optimise (Roy et al., 2023). In our recruitment example, choosing to improve demographic parity across male and female candidates selected for interview risks reducing demographic parity with respect to race (Chen et al., 2024). Some works tackle this issue and allow multiple protected attributes (Foulds et al., 2020). For example, Zafar et al. (2017) optimise for fairness under constraints that represent fairness metrics with respect to each protected attribute, while Choi et al. (2020) create their own fairness metric. However, the use of custom metrics may present its own issues, which we discuss later in this section.

In addition, protected characteristics such as 'gender', 'race' and 'age' cannot be fully captured as a binary attribute (Weinberg, 2022). Some methods (Quadrianto & Sharmanska, 2017; Zafar et al., 2017) tackle this by allowing for multi-valued and numerical

6. We denote recommendations related to applicability as $R_a$ and to regulations as $R_r$.

protected attributes in their optimisation constraints, while Feldman et al. (2015) allows for multi-valued protected attributes as long as they have a natural ordering.

When creating a new bias mitigation method, if access to protected attributes is assumed, the number of protected attributes the method aims to improve fairness for and the type of these attributes should be clearly highlighted.

With all of the above in mind, we come to our first applicability recommendation.

**$R_a$1: Protected Attributes**

When developing a new bias mitigation method, efforts should focus on ensuring its applicability in scenarios where there are multiple pre-specified protected attributes and that they could be of any type including:

- binary,
- multi-valued, or
- numerical.

#### 3.1.2 Model Types

As outlined in Section 2, different bias mitigation methods target different stages of the ML pipeline: the training data (pre-processing), during model training (in-processing), and the model's outputs (post-processing). Pre-processing and most post-processing methods[7] are model-agnostic and thus can be applied in a wider range of scenarios.

In-processing methods can be model-agnostic (Krasanakis et al., 2018; Iosifidis & Ntoutsi, 2019; Oneto et al., 2019). However, many are designed for a specific model e.g., Naïve Bayes models (Calders & Verwer, 2010; Choi et al., 2020), Bayesian networks (Mancuhan & Clifton, 2014), logistic regression (Bechavod & Ligett, 2017; Zhang et al., 2018), gradient tree boosting (Grari et al., 2019), decision trees (Kamiran et al., 2010), neural networks (Hu et al., 2020) and can improve fairness more effectively than model-agnostic ones (Caton & Haas, 2020). Developing an in-processing model-specific method is only possible if there is access to the model and, obviously, these methods can only be used on the specific models for which they were designed (Aggarwal et al., 2019).

There are many considerations when choosing the stage in the ML pipeline at which bias should be mitigated. Our next recommendation instead highlights the risk for many methods to never be used in the real-world due to not being applicable in different scenarios, or to not being specific about the scenarios in which they could be used.

7. A sub-category of post-processing methods, intra-processing methods, require some knowledge of the model used e.g., decision tree nodes (Kamiran et al., 2010), posterior probabilities (Calders & Verwer, 2010), to mitigate the bias in the output (Savani et al., 2020).

**$R_a$2: Model Types**

When developing a new bias mitigation method, efforts should be made to ensure its applicability across scenarios involving various ML models, especially when the inner workings of the models are not accessible, or to justify the implementation for specific models.

#### 3.1.3 Fairness Metrics

As previously stated, bias mitigation methods aim to optimise for some metric which quantifies a notion of fairness. The notion of fairness that should be optimised for depends on the intended use case, hence the metric deployed with a bias mitigation method must be representative of the scenario in which it is applied. For example, a developer may choose to optimise for equalised odds in a system that predicts who will pay back their loan. Equalised odds is a suitable metric here as it is important to ensure equal proportions of individuals who are correctly and incorrectly predicted to pay back a loan across the unprivileged and privileged groups. This requires choosing a method that optimises for equalised odds (e.g., Grari et al., 2019; Iosifidis & Ntoutsi, 2019; Hu et al., 2020) over ones that do not (e.g., Krasanakis et al., 2018; Choi et al., 2020). Thus, the metric the bias mitigation method aims to improve limits the applications in which it can be used. Some bias mitigation methods allow the user to choose the fairness metric to optimise for (e.g., Quadrianto & Sharmanska, 2017; Zhang et al., 2018). This is useful as it allows the method to be applied in different scenarios. Further, optimising with respect to one fairness metric can have a detrimental effect on the performance with respect to another metric (Kleinberg et al., 2016; Quadrianto & Sharmanska, 2017; Garg et al., 2020)

Some methods also create new fairness metrics (Mancuhan & Clifton, 2014; Fish et al., 2016; Choi et al., 2020). Often, their relationships with real-world notions of fairness are not explored thus it is difficult to know in what scenarios they might be applied as well as their relationship with existing metrics.

**$R_a$3: Fairness Metrics**

When developing a new bias mitigation method, attempts should be made to ensure it can be applied to optimise for a range of fairness metrics, allow for the specification of the fairness metric to optimise for or justify why a specific (set of) metrics have been chosen by describing what notions of fairness they correspond to.

#### 3.1.4 Evaluation of Methods

Whilst the works proposing bias mitigation methods provide results to evaluate their effectiveness, these results are not robust as they are vulnerable to changes in the experimental setup (Friedler et al., 2019). Given a scenario, a bias mitigation method is chosen based on the characteristics of the dataset, model used, and the notion of fairness chosen to optimise for. However, one question arises: how can we ensure that the method will indeed improve

fairness? Each method includes experiments using a chosen model trained on publicly available datasets, with results reported using different metrics. However, the results are not easily comparable across methods due to the multitude of these choices. Other factors also impact the values of computed fairness metrics, e.g., different distributions of positive and negative labels across the (un)privileged groups in the training data (Iosifidis & Ntoutsi, 2019) or different proportions of training/testing data (Hort et al., 2024).

Experiments on different datasets and models using different fairness metrics can give very different results. Non-robust evaluations may be used to incorrectly justify the use of a bias mitigation method which could lead to potentially discriminatory systems being deployed.

**$R_a$4: Evaluation of Methods**

The evaluation of new bias mitigation methods should be thorough and robust such that the method's effectiveness in various scenarios can be understood. There should be evaluations provided with different datasets, varied train-test splits and models, using different fairness (and performance) metrics when possible.

#### 3.1.5 Trade-offs

Closely linked to the evaluation of bias mitigation methods, there are potential trade-offs associated with applying methods to ADMS. One such trade-off is between fairness and performance. For example, Cardoso et al. (2019) trained models on biased datasets to show that existing methods optimising for the disparate impact metric (Kamiran & Calders, 2011; Feldman et al., 2015) improve fairness but decrease the model's performance. Other methods also showed a decrease in performance (Iosifidis & Ntoutsi, 2018; Oneto et al., 2019), while Zhang et al. (2018) included a parameter to control the fairness-performance trade-off. However, there is limited discussion as to the acceptable level of decrease in performance to achieve a given level of fairness in the decisions.

Improving fairness with respect to one metric may be at a detriment to another (Kleinberg et al., 2016; Quadrianto & Sharmanska, 2017). The notions of group and individual fairness conflict, thus targeting only one notion does not fully capture fairness. Targeting only individual fairness (Ruoss et al., 2020) ensures that similar individuals are treated the same. However, in our recruitment example, if the similar individuals are all female candidates and they are all not selected for an interview, then individual fairness would be satisfied but group fairness (demographic parity) with respect to sex would not be satisfied. Similarly, only improving fairness with respect to one protected attribute could risk decreasing fairness with respect to another (Chen et al., 2024).

Another trade-off of applying bias mitigation methods could be reducing interpretability. Interpretability of ADMS is not as easy to measure (Miller, 2022) thus it is rarely considered. The impact of bias mitigation methods on transparency and explainability is discussed further in Section 3.2.5.

All potential trade-offs need to be considered when creating a bias mitigation method. The variability in results using different datasets, models, and metrics, also impacts these

trade-offs, making it difficult to ever be certain on a method's applicability and effectiveness. However several frameworks have been developed (e.g., Bellamy et al., 2018; Cardoso et al., 2019; Friedler et al., 2019; Schelter et al., 2020) to explore these differences (see Section 5).

**$R_a$5: Trade-offs**

The potential trade-offs of the application of new bias mitigation methods should be thoroughly evaluated and discussed.

### 3.2 Legality Recommendations

There are many existing laws that regulate the use of ADMS, specifically with the aim of ensuring they do not discriminate. As outlined in Section 2, these include cross-cutting regulation and sector-specific regulation. These regulations, which often differ between different jurisdictions, contribute to governing the use of technical bias mitigation methods for ADMS. In this section, we bridge the gap between prominent bias mitigation methods for binary classification ADMS and regulation around non-discrimination and data protection, with a focus on US, EU and UK laws by providing recommendations around the key considerations for creating new bias mitigation methods.

#### 3.2.1 Legal Definitions of Fairness

As previously discussed, the notion of fairness should be chosen depending on the context of the decision-making system. This context should also include the laws pertaining to the application of system (Centre for Data Ethics and Innovation, 2020; Nachbar, 2020). Each regulatory body has its own definition of fairness to which decision-makers must adhere (European Commission, 2022).

Regulation in UK and EU defines discrimination as direct and indirect. Direct discrimination focuses on an individual and whether they have been treated differently based on protected characteristics. Simply removing protected attributes from ADMS, fairness through unawareness (Grgic-Hlaca et al., 2016), will ensure ADMS avoid directly discriminating as there is no possibility for them to base their decisions on these protected attributes.[8] Some methods are considered more reliant on protected attributes than others, with some specifically aiming to minimise reliance on these attributes (and therefore being less likely to risk direct discrimination) and others making no attempt to (Jorgensen et al., 2023b). Further, individual fairness metrics such as consistency and context-dependent definitions (see Section 2) can be useful in assessing whether a system is directly discriminating by comparing similar individuals and ensuring they are equally treated.

However, as discussed in Section 2.2.1, the removal of protected attributes could still result in indirect discrimination due to proxies. Existing metrics such as demographic parity can assist in measuring indirect discrimination, yet, in EU case law, statistical measures have rarely been highly considered, with courts favouring individual contextual factors (Wachter

8. Some works argue that if decisions are based on attributes that correlate closely with the protected attributes, this could also be classed as direct discrimination (Adams-Prassl et al., 2023), however without knowledge of the protected characteristics this would be very difficult to assess.

et al., 2021). It is difficult (if not impossible) to quantify this notion in a metric (Xenidis, 2020), however conditional demographic disparity metric (Wachter et al., 2021) was created with this aim.

Disparate treatment in US non-discrimination regulation involves assessing the intent of the decision-maker (Civil Rights Division, U.S. Department of Justice, 2016a), which does not translate easily to an algorithmic fairness definition as 'intent' cannot be attributed to ADMS. The burden would revert back to those implementing or deploying the system (Ashton, 2023). Applying bias mitigation techniques such as the methods discussed in this paper could be seen as attempting to ensure the system is fair, and as such hide any bad intent which is a potential weakness of disparate treatment and why it is not usually discussed in relation to algorithmic fairness.

There are also considerations around the disparate impact of a system (see Section 2) being less than 80% (Civil Rights Division, U.S. Department of Justice, 2016b). This definition has been widely adopted throughout the algorithmic fairness literature (Calders & Verwer, 2010; Kamiran et al., 2012; Feldman et al., 2015; Zafar et al., 2017), focusing on the impact of decisions made and whether protected attributes are used in the decision-making. However, to measure disparate impact using existing metrics, knowledge of individuals' protected attributes are required.

Fairness definitions in law might differ from the technical definitions used in algorithmic fairness. The development of a new bias mitigation method should include which legal definitions of fairness it does (not) satisfy (direct versus indirect discrimination, disparate impact versus disparate treatment) and in which legal jurisdictions these definitions fall under.

**$R_r$1: Legal Definitions of Fairness**

When developing a new bias mitigation method, the fairness metric(s) chosen to optimise for should be consistent with some legal notion of fairness. This should be specified or a justification as to why the metric chosen does not align with existing legal frameworks should be provided.

#### 3.2.2 Data Protection

The majority of bias mitigation methods require the identification of the protected attributes before they can be applied. However, under UK, EU and US data protection laws (GDPR, 2016; UK Information Commissioner's Office, 2021; State of California Department of Justice, 2024) the collection, processing and storage of personal characteristics should be justified and is held to high standards of transparency as discussed in Section 2. Often organisations might not collect protected attributes due to concerns or misconceptions around the legality of using them to audit their systems (Centre for Data Ethics and Innovation, 2023) or the potential cost and effort of collecting this data could lead organisations to prioritise gathering only some protected attributes, if any. Further, similar to the collection of all data, information about protected attributes could be incomplete or unreliable. More consideration should be given to cases where only partial details of pro-

tected characteristics are available for some individuals, and to determining the best way to record various values of these protected attributes.

Without somewhat complete information about individuals' protected characteristics, many of the bias mitigation methods become unusable. Some methods infer protected characteristics from other attributes (in other words pre-specifying sensitive attributes) (Chen et al., 2019). However, unless there is a direct correlation between a proxy and a protected attribute, there is a risk of already discriminating in the prediction of protected attributes. Some policy bodies such as the European Committee of Social Rights and the United Nations Special Effort on Extreme Poverty and Human Rights have advocated for an ethical responsibility to collect sensitive data to facilitate legal proceedings which would be useful for the assessment and mitigation of algorithmic discrimination using existing methods (Wachter et al., 2021). However, the relationship of existing bias mitigation methods to data protection is rarely considered (Haeri & Zweig, 2020).

**$R_r$2: Data Protection**

When developing a new bias mitigation method, the possibility of not having access to the protected attributes for all (or any) individuals should be considered. New methods should aim to be less reliant on pre-specified protected attributes for bias mitigation.

#### 3.2.3 Positive Discrimination

After the application of a bias mitigation method to an ML model, the classifications might differ from the unmitigated model based on a protected attribute value (Grari et al., 2019). This could cause individuals from the unprivileged group to be favoured over others (Wachter et al., 2020), otherwise known as *positive discrimination*. For example, imagine our recruitment system is used to select 6 candidates for interview. There are 10 male candidates and 10 female candidates who apply and the model predicts 4/10 male candidates and 2/10 female candidates as being fit for the job. Given the model predictions are solely used to decide who gets selected for interview, the disparate impact is 50%, which is not acceptable (>80% is the acceptable level in US regulation as discussed in Section 3.2.1). After the application of a bias mitigation method, 3/10 male candidates and 3/10 female candidates are now selected for interview meaning the disparate impact is 100%, in other words there are the same proportion of male and female candidates selected for interview. However, increasing the number of female candidates selected for interview, based on them being female, is positive discrimination.[9] When there are limited resources to enact decisions, meaning there is a strict maximum of how many positive decisions can be made, there is a higher risk of positive discrimination to achieve better fairness metric values. If a system could select every candidate for interview, it would be very 'fair' (although accuracy would likely be very poor). All bias mitigation methods that have knowledge of protected

9. This is a further example of the conflict between group and individual fairness. A notion of group fairness is satisfied here (disparate impact) however a male candidate could find that similar individuals to him are classified differently.

attributes risk positively discriminating. Post-processing methods, in particular, are especially risky, as they often directly alter classifications based on protected characteristics to optimise particular metrics.

In some jurisdictions, such as in the UK under the UK Equality Act (UK Public General Acts, 2010; Centre for Data Ethics and Innovation, 2020), changing any outcome based on protected attributes is unlawful except in special cases of positive action. However, positive action usually involves practices aimed at promoting fairness as opposed to using the protected attributes directly for changing decisions thus is not highly related with the deployment of ADMS.

US regulation is less strict and positive discrimination (affirmative action) can be deployed in some cases through the use of quotas or relaxing selection criteria. This is comparable to the example above of the mitigated model ensuring there is some proportion of female candidates selected for the interview which could, in turn, reduce the proportion of male candidates selected (Xiang & Raji, 2019). However, affirmative action is highly disputed and it would need to be justified thoroughly if the application of a bias mitigation method means positively discriminating.

The issues around the potential to positively discriminate are inherently due to the definition of the commonly used group fairness metrics, for which it might be necessary to treat individuals in the privileged group worse in order to reach a satisfactory equality. Another framing of fairness is to ensure, for each protected group, a minimum threshold of positive classifications (or true positives, false positives, etc.) is met to guarantee no group is 'levelled down' (Mittelstadt et al., 2023). However, as previously stated, this could be difficult given fixed number of available positive classifications.

Potential positive discrimination resulting from existing bias mitigation methods is rarely discussed. However, reducing discrimination without positively discriminating (Mancuhan & Clifton, 2014) is an important consideration for any method to be applied in a real-world scenario.

**$R_r$3: Positive Discrimination**

New bias mitigation methods should avoid positive discrimination occurring as a result of applying the method if possible. If the new method has the potential to positively discriminate (i.e. a mitigated model's classifications are changed based on any value of a protected attribute), this should be discussed and scenarios where this might be necessary should be outlined.

#### 3.2.4 Sector-specific

Sector-specific regulations highlight the need for context-specific bias mitigation approaches. For example, ADMS deployed in employment in the US are subject to the Civil Rights Act and also other acts with specific provisions for age discrimination, disabilities and pregnancy. For decision-making, the characteristics are likely not considered in the same way as each other and other characteristics. The different caveats of discrimination for different protected characteristics highlight that they cannot all be treated in the same way when introducing non-discrimination measures to ADMS.

The financial services regulations in the UK also emphasise the contextual importance of the application of bias mitigation methods. Mitigation should be considered on a case-by-case basis depending on what is the most proportionate way of addressing potential consumer harm, regardless of the type of technology or algorithm used in the process (Financial Conduct Authority, 2018).

Regulations specific to the public-sector such as the UK's Public Sector Equality Duty have extra obligations to ensure transparency and the promotion of equal opportunities. ADMS should not be employed in the public sector unless every step of the design process and decision-making is fully explainable, if deemed appropriate for use at all (Waller & Waller, 2020).

Sector-specific regulations are much more difficult to discuss in relation to bias mitigation methods, as designing methods for one sector, or to comply with one set of regulations, would take significant understanding of the regulation, and could limit its applicability. However, there should be consideration of any sector-specific regulations that might impact the application of new bias mitigation methods in different domains. Further, this motivates the need for methods that focus on explaining decisions such that the bias can be assessed given knowledge of the context in which they are applied.

**$R_r4$: Sector-specific**

Sector-specific regulations can define nuances to ensure equality in different domains, for example health, finance, and the public sector. The creation of new methods should consider how regulations in different domains may impact the use of the method.

#### 3.2.5 Transparency and Explainability

Regulations such as (UK) GDPR ensure the protection of personal data and outline protections against opaque ADMS, however these transparency and explainability regulations are vague and their meaning is contested (Wachter et al., 2017). Due to the lack of a federal approach to ADMS so far in the US, these principles are not specifically regulated in all scenarios. However, as more laws are crafted, it is becoming increasingly important to prioritise transparency and explainability, especially in high-impact systems. These aspects are essential for fostering trust in the technology and ensuring accountability of decisions made using ADMS. Recent developments, such as the EU AI Act and various white-papers (Centre for Data Ethics and Innovation, 2020; European Commission, 2021; Schöffer et al., 2021), underscore the critical role of transparency and explainability for individuals impacted by decisions in high-stakes scenarios. For an individual to contest a decision (a right which is ensured in GDPR), they must be able to query why a decision was made. In Section 5 we discuss the need for an explanation for proving to a court that a decision was discriminatory.

Decisions made by ADMS can be difficult to explain due to their opacity, or lack of understanding of how the technology works. Arguably, using any of the bias mitigation methods discussed further increases the system's opacity and adds another layer of automated processing that requires explanation. Every bias mitigation method works differently, with various design choices made in the process (many discussed throughout this paper).

There are existing explainability methods for ADMS (Holzinger et al., 2022) but the application of bias mitigation methods to ADMS could impact their usefulness. The impact that the application of a bias mitigation method may have on the transparency of the system is rarely acknowledged (Chakraborty et al., 2020; Waller et al., 2024).

In order to consider the context of a system and not rely on pre-specified protected attributes which may not be available, we suggest exploring explainable AI (XAI) methods specifically designed for bias detection and mitigation. Stakeholders can then consider the reasoning for the system's classification and decide whether it is fair based on the context and application of the system. There is a plethora of research into XAI, but its use cases in relation to fairness are limited (Begley et al., 2020; Grabowicz et al., 2022; Calegari & Sabbatini, 2023; Ciatto et al., 2024; Waller et al., 2024). There might still be issues with a stakeholder evaluating the system's fairness but the concept of fairness is inherently human-oriented, context-specific and culturally dependent (Barabas et al., 2020), meaning that it is difficult to automate, cannot be reduced to a metric, and requires some level of human input also for accountability purposes (Veale et al., 2018).

**$R_r$5: Transparency and Explainability**

New bias mitigation methods should either:

- focus on embedding transparency and explainability by ensuring the impact of applying a method to a system is highly interpretable, or
- be able to explain the decisions made such that stakeholders can evaluate whether the decision was fair or not according to some notion of fairness, allowing greater consideration of context.

Overall, there is a need for cross-disciplinary considerations on issues such as fairness. Whilst not a new recommendation (Van Nuenen et al., 2020; Aran et al., 2021; Cheng et al., 2021; Wachter et al., 2021; Weinberg, 2022; Hort et al., 2024), it has not yet been universally adopted in current research around fairness and bias in AI. There is still a gap between the algorithmic fairness and the non-discrimination legal communities. Indeed,

> "Aligning the design of autonomous systems with contextual equality, flexibility, and the judicial interpretation of the comparative aspects of non-discrimination law would greatly benefit all parties involved" (Wachter et al., 2021).

## 4. Key Considerations in the Development of Bias Mitigation Methods

We propose a checklist which we suggest should be included along with the creation of any new bias mitigation methods, relating to the recommendations for applicability and legality that we introduced in Section 3. We believe that including the checklist will make it easier for developers of ADMS to understand the limitations of different methods.

**Applicability Considerations**

Does your method require access and specification of protected attributes? ➢ $R_a1$
- If yes, what type of attributes can the protected attributes be? Binary/Multi-valued/Numerical?

What models can your method be applied to? ➢ $R_a2$
- Does your method require access to the inner workings of the model?

What fairness metrics have you chosen to optimise for? ➢ $R_a3$
- Have you justified why you have chosen these metrics?

Have you thoroughly evaluated your method on different datasets and models and using different fairness and performance metrics? ➢ $R_a4$
- Have differences in results from different evaluation setups been discussed/explained/justified?

Have any trade-offs of using your method been discussed? ➢ $R_a5$

**Legality Considerations**

What legal definitions of fairness does your method address? ➢ $R_r1$
- If none, have you justified why not?

Have you considered if your method works when you do not have full access to protected attributes, potentially due to misinterpretations of data protection regulations? ➢ $R_r2$

If your method changes classifications using protected characteristics, have you discussed how this could be illegal positive discrimination in some contexts? ➢ $R_r3$

If thinking about specific domains that your method might be applied in, have you discussed how your method complies with sector-specific regulations? ➢ $R_r4$

Have you discussed the impact of applying your method to the transparency and explainability of the system? ➢ $R_r5$

The recommendations presented stem from the limitations of existing bias mitigation methods for binary classification on tabular data (Dunkelau & Leuschel, 2019; Saleiro et al., 2019; Waller et al., 2023a; Corbett-Davies et al., 2024; Hort et al., 2024). As previously mentioned in Section 2, bias mitigation methods have been proposed for various settings including ADMS based on regression (Calders et al., 2013; Berk et al., 2017; Chzhen et al., 2020), multi-class classification (Alghamdi et al., 2022; Putzel & Lee, 2022), and clustering (Chierichetti et al., 2017; Rösner & Schmidt, 2018; Backurs et al., 2019; Bera et al., 2019; Abbasi et al., 2021; Ziko et al., 2021). Further, bias in different types of data such as text, images and graphs have been explored, though sometimes it is not as straightforward in these settings how bias translates to discriminatory decisions.

Some of our recommendations can be transferred to these other settings. As an example, consider a decision-making system that uses clustering to decide the best placement of polling locations for an election (Abbasi et al., 2021). The same considerations of applicability and legality can be used to ensure that the chosen locations do not disadvantage protected groups from enacting their right to vote, i.e., what notions of fairness should be optimised for (Chierichetti et al., 2017). Similarly, assessing discrimination in ADMS that rely on data such as images (Nadeem et al., 2020; Yang et al., 2024), text (Dash et al., 2019; Schaaf et al., 2021) or graphs (Fisher et al., 2020; Chuang et al., 2022) is even more challenging. An example is a facial recognition system based on image data, used by the police to flag individuals who are on their watchlist, potentially leading to an arrest or intervention (Mansfield, 2023). In such systems, protected characteristics within the data are not as easy to identify and additional labelling or metadata is needed to determine whether data instances belong to a protected group — thus complicating the definition of fairness metrics. However, it is still important to take into account the applicability considerations we propose, such as the chosen notions of fairness and the evaluation of bias mitigation methods applied. For example, the measured fairness of ADMS based on facial recognition varies depending on the datasets, models, and evaluation setups used, all of which should be discussed and clearly explained, with the selected fairness notion justified. Furthermore, legality considerations related to sector-specific regulations, transparency, and explainability are still crucial for all ADMS, independent of the type of data used for training. For legality considerations, the focus is often on the application of a system and the impact it has on individuals or groups.

## 5. Discussion

In this section we discuss the legal process of proving discrimination and further highlight the need for explainable decisions.

Under EU or UK law, when an individual believes they are being discriminated against, they must give evidence of this discrimination, otherwise known as establishing *prima facie* discrimination. Specifically, they must demonstrate that (Wachter et al., 2021):

> "(1) a particular harm has occurred or is likely to occur; (2) the harm manifests or is likely to manifest significantly within a protected group of people; and (3) the harm is disproportionate when compared with others in a similar situation."

Once prima facie discrimination has been established, the burden of proof switches to the alleged discriminator and they must show that they have acted legally. In some contexts — for example, in the public sector under the UK's Public Sector Equality Duty — failure of the alleged discriminator to demonstrate sufficient consideration of the fairness of the system can be enough to establish liability. Generally, the alleged discriminator can demonstrate they are not liable (Wachter et al., 2021):

> "(1) by refuting that a causal link between the differential results and a protected ground does not exist, or (2) by acknowledging that differential results have occurred but providing a justification that is based on the pursuit of a legitimate interest in a necessary and proportionate manner."

The use of ADMS might obscure individuals' awareness of discrimination. However, if they can demonstrate prima facie discrimination resulting from ADMS, the alleged discriminators would likely need to justify how they ensured the system's fairness. This justification could entail explaining the chosen metrics used, evaluating various types of discrimination, and addressing other questions in the checklist in Section 4.

Potentially due to the difficulty for an individual to assess the fairness of decisions made using ADMS, there is a lack of case law which makes it challenging to interpret the above conditions for shifting the burden of proof, as well as non-discrimination laws generally relating to ADMS. As these systems become more widely adopted and transparency improves, case law will likely guide development of regulation by helping to understand judicial decisions when interpreting existing regulation, such as the definitions of direct, indirect, and positive discrimination where algorithmic systems are used in decision-making.

To aide individuals impacted by ADMS to be able to prove whether they have been treated fairly, it is crucial to be able to explain the logic of the decisions made to show they were made for legitimate reasons. This includes justifying the use of ADMS, the reasons behind all design choices and ways to describe the internal workings of the systems used to all stakeholders. Due to the context specificity of fairness and the reliance on pre-specified protected attributes, using XAI for detecting bias should be explored further. Existing work in symbolic knowledge extraction and injection aligns with this goal, specifically by aiming to identify biases or bugs in training data or within the operation of black-box models. Ruggieri et al. (2023) propose the 'train–extract–fix–inject' (TEFI) loop, which corresponds to the detection and mitigation of bias within ADMS. This approach focuses on the use of symbolic knowledge to be able to explain the underlying logic of decisions to data scientists and other stakeholders, with the ultimate aim to satisfy the right to an explanation under the GDPR.

Although it is debatable whether existing technical XAI methods can be used as a justification for the use of ADMS, they are a crucial step in establishing better interactions with individuals impacted in order to ensure they have some agency in the decisions that impact them.

## 6. Conclusion

Existing bias mitigation methods for binary classification ADMS have significant limitations. Building upon Waller et al. (2023b), we proposed recommendations for creating new methods that are effective, applicable in a wide range of scenarios and compliant with legal requirements such as non-discrimination, data protection and sector-specific regulations. Using an example of ADMS deployed in recruitment, we illustrated these recommendations and highlighted the need to revise current approaches to bias mitigation. We also provided a checklist for developers of new methods to use to guide their development processes. Our recommendations aim to guide the advancement of research in this crucial area of AI to ensure that bias mitigation methods are developed in a more responsible manner.

## Acknowledgments

This work was supported by the UK Research and Innovation Centre for Doctoral Training in Safe and Trusted Artificial Intelligence[10] [grant number EP/S023356/1]. Madeleine Waller, Odinaldo Rodrigues and Oana Cocarascu are affiliates of the King's Institute for Artificial Intelligence.[11] Madeleine Waller is also partially funded by The Alan Turing Institute's Enrichment Scheme. The authors are grateful for Mackenzie Jorgensen's contribution to initial discussions of this work. The authors further thank Karen Yeung, Paul Waller and the Legal, Ethical & Accountable Digital Society (LEADS) Lab at the University of Birmingham for their valuable feedback on the legal aspects of this article.

## References

Abbasi, M., Bhaskara, A., & Venkatasubramanian, S. (2021). Fair clustering via equitable group representations. In *FAccT '21: Conference on Fairness, Accountability, and Transparency*, pp. 504–514.

Adams-Prassl, J., Binns, R., & Kelly-Lyth, A. (2023). Directly discriminatory algorithms. *The Modern Law Review*, *86*(1), 144–175.

Aggarwal, A., Lohia, P., Nagar, S., Dey, K., & Saha, D. (2019). Black box fairness testing of machine learning models. In *Proceedings of the ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/SIGSOFT FSE*, pp. 625–635.

Alan, A., Ennabe, M., Sulaiman, A., & Weinand, M. (2024). Discontinuation of affirmative action: Consequences for black educational equity, neurosurgical residency, and medical diversity, with consideration of potential adversity as a new path forward. *World Neurosurgery*, *X*, 100339.

Alghamdi, W., Hsu, H., Jeong, H., Wang, H., Michalak, P. W., Asoodeh, S., & Calmon, F. P. (2022). Beyond Adult and COMPAS: Fairness in multi-class prediction. arXiv:2206.07801.

Aran, X. F., van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and discrimination in AI: A cross-disciplinary perspective. *IEEE Technology and Society Magazine*, *40*(2), 72–80.

Archibong, U., & Sharps, P. W. (2013). A comparative analysis of affirmative action in the United Kingdom and United States. *Journal of Psychological Issues in Organizational Culture*, *3*(S1), 28–49.

Ashton, H. (2023). Definitions of intent suitable for algorithms. *Artificial Intelligence and Law*, *31*(3), 515–546.

Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., & Wagner, T. (2019). Scalable fair clustering. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 405–413.

10. https://www.safeandtrustedai.org
11. https://www.kcl.ac.uk/ai

Barabas, C., Doyle, C., Rubinovitz, J. B., & Dinakar, K. (2020). Studying up: reorienting the study of algorithmic fairness around issues of power. In *FAT* '20: Conference on Fairness, Accountability, and Transparency*, pp. 167–176.

Bechavod, Y., & Ligett, K. (2017). Learning fair classifiers: A regularization-inspired approach. arXiv:1707.00044.

Begley, T., Schwedes, T., Frye, C., & Feige, I. (2020). Explainability for fair machine learning. arXiv:2010.07389.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv:1810.01943.

Bera, S. K., Chakrabarty, D., Flores, N., & Negahbani, M. (2019). Fair algorithms for clustering. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, (NeurIPS)*, pp. 4955–4966.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M. J., Morgenstern, J., Neel, S., & Roth, A. (2017). A convex framework for fair regression. arXiv:1706.02409.

Blanchard, W. (1986). Evaluating social equity: What does fairness mean and can we measure it?. *Policy Studies Journal*, *15*(1).

Bosoer, L., Cantero Gamito, M., & Rubio-Marin, R. (2023). *Non-Discrimination and the AI Act*. Casadei (ed.), Law and Digitalization. Arazandi. Available at: `https://ssrn.com/abstract=4666071`.

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *ICDM Workshops 2009, IEEE International Conference on Data Mining Workshops*, pp. 13–18.

Calders, T., Karim, A., Kamiran, F., Ali, W., & Zhang, X. (2013). Controlling attribute effect in linear regression. In *2013 IEEE 13th International Conference on Data Mining*, pp. 71–80.

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

Calegari, R., & Sabbatini, F. (2023). The psyke technology for trustworthy artificial intelligence. In *AIxIA 2022 – Advances in Artificial Intelligence*, pp. 3–16, Cham. Springer International Publishing.

Cardoso, R. L., Jr., W. M., Almeida, V. A. F., & Zaki, M. J. (2019). A framework for benchmarking discrimination-aware models in machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES*, pp. 437–444.

Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. arXiv:2010.04053.

Centre for Data Ethics and Innovation (2020). Review into bias in algorithmic decision-making. Available at: `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/957259/Review_into_bias_in_algorithmic_decision-making.pdf`.

Centre for Data Ethics and Innovation (2023). Enabling responsible access to demographic data to make AI systems fairer. Available at: https://www.gov.uk/government/publications/enabling-responsible-access-to-demographic-data-to-make-ai-systems-fairer/report-enabling-responsible-access-to-demographic-data-to-make-ai-systems-fairer.

Chakraborty, J., Peng, K., & Menzies, T. (2020). Making fair ML software using trustworthy explanation. In *35th IEEE/ACM International Conference on Automated Software Engineering, ASE*, pp. 1229–1233. IEEE.

Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, p. 339–348. Association for Computing Machinery.

Chen, Z., Zhang, J. M., Sarro, F., & Harman, M. (2024). Fairness improvement with multiple protected attributes: How far are we?. In *2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE)*, pp. 944–944. IEEE Computer Society.

Cheng, L., Varshney, K. R., & Liu, H. (2021). Socially responsible AI algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, *71*, 1137–1181.

Chierichetti, F., Kumar, R., Lattanzi, S., & Vassilvitskii, S. (2017). Fair clustering through fairlets. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pp. 5029–5037.

Choi, Y., Farnadi, G., Babaki, B., & den Broeck, G. V. (2020). Learning fair naive bayes classifiers by discovering and eliminating discrimination patterns. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pp. 10077–10084. AAAI Press.

Chuang, Y.-N., Lai, K.-H., Tang, R., Du, M., Chang, C.-Y., Zou, N., & Hu, X. (2022). Mitigating relational bias on knowledge graphs. arXiv:2211.14489.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., & Pontil, M. (2020). Fair regression via plug-in estimator and recalibration with statistical guarantees. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Ciatto, G., Sabbatini, F., Agiollo, A., Magnini, M., & Omicini, A. (2024). Symbolic knowledge extraction and injection with sub-symbolic predictors: A systematic literature review. *ACM Computer Survey*, *56*(6).

Civil Rights Division, U.S. Department of Justice (2016a). Section VI — proving discrimination — intentional discrimination. Available at: https://www.justice.gov/crt/fcs/T6Manual6.

Civil Rights Division, U.S. Department of Justice (2016b). Section VII - proving discrimination — disparate impact. Available at: https://www.justice.gov/crt/fcs/T6Manual7.

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2024). The measure and mismeasure of fairness. *Journal of Machine Learning Research*, *24*(1).

Dash, A., Shandilya, A., Biswas, A., Ghosh, K., Ghosh, S., & Chakraborty, A. (2019). Summarizing user-generated textual content: Motivation and methods for fairness in algorithmic summaries. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–28.

Dastin, J. (2022). Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of Data and Analytics*. Auerbach Publications.

Deloitte (2021). AI and Automated Decision Making. Available at: https://www2.deloitte.com/uk/en/pages/deloitte-analytics/articles/ai-and-automated-decision-making.html.

Dunkelau, J., & Leuschel, M. (2019). Fairness-aware machine learning: An extensive overview. Available at: https://stups.hhu-hosting.de/downloads/pdf/fairness-survey.pdf.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. S. (2012). Fairness through awareness. In *Innovations in Theoretical Computer Science*, pp. 214–226.

Edemekong, P., Annamaraju, P., & Haydel, M. (2024). Health insurance portability and accountability act. StatPearls [Internet], Available at: https://www.ncbi.nlm.nih.gov/books/NBK500019/.

Edwards, L. (2021). The EU AI Act: a summary of its significance and scope. *Ada Lovelace Institute*, *1*.

Engin, Z., & Treleaven, P. C. (2019). Algorithmic government: Automating public services and supporting civil servants in using data science technologies. *Computer Journal*, *62*(3), 448–460.

Equality and Human Rights Commission (2023). Public Sector Equality Duty: guidance for public authorities. Available at: https://www.gov.uk/government/publications/public-sector-equality-duty-guidance-for-public-authorities/public-sector-equality-duty-guidance-for-public-authorities.

European Commission (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM/2021/206 final.. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

European Commission (2022). A comparative analysis of non-discrimination law in Europe 2022. Available at: https://www.equalitylaw.eu/downloads/5812-a-comparative-analysis-of-non-discrimination-law-in-europe-in-2022.

European Union Agency for Fundamental Rights (2007). EU charter of fundamental rights. Available at: http://fra.europa.eu/en/eu-charter/article/21-non-discrimination.

Federal Trade Commission, M., Commission, F. T., et al. (1998). Children's online privacy protection rule ("coppa"'). Available at: https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD*, pp. 259–268.

Financial Conduct Authority (2018). Price discrimination in financial services. Available at: `https://www.fca.org.uk/publications/research/price-discrimination-financial-services`.

Fish, B., Kun, J., & Lelkes, Á. D. (2016). A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 144–152.

Fisher, J., Mittal, A., Palfrey, D., & Christodoulopoulos, C. (2020). Debiasing knowledge graph embeddings. In Webber, B., Cohn, T., He, Y., & Liu, Y. (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 7332–7345. Association for Computational Linguistics.

Foulds, J. R., Islam, R., Keya, K. N., & Pan, S. (2020). An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 1918–1921.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, pp. 329–338.

Galhotra, S., Saisubramanian, S., & Zilberstein, S. (2021). Learning to generate fair clusters from demonstrations. In *AIES: AAAI/ACM Conference on AI, Ethics, and Society*, pp. 491–501.

Garg, P., Villasenor, J. D., & Foggo, V. (2020). Fairness metrics: A comparative analysis. In *IEEE International Conference on Big Data*, pp. 3662–3666.

GDPR (2016). General Data Protection Regulation (GDPR) – Official Legal Text. Available at: `https://gdpr-info.eu/`.

Gil González, E., & De Hert, P. (2019). Understanding the legal provisions that allow processing and profiling of personal data—an analysis of GDPR provisions and principles. In *Era Forum*, Vol. 19, pp. 597–621. Springer.

Grabowicz, P. A., Perello, N., & Mishra, A. (2022). Marrying fairness and explainability in supervised learning. In *FAccT '22: Conference on Fairness, Accountability, and Transparency*, pp. 1905–1916. ACM.

Gramm, P. (1999). Gramm-Leach-Bliley Act. In *Vol. Public Law 106–102). Washington, DC: United States Congress.*

Grari, V., Ruf, B., Lamprier, S., & Detyniecki, M. (2019). Fair adversarial gradient tree boosting. In *2019 IEEE International Conference on Data Mining, ICDM*, pp. 1060–1065.

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. In *Proc. NIPS Symposium on Machine Learning and Law*, Vol. 1.

Haeri, M. A., & Zweig, K. A. (2020). The crucial role of sensitive attributes in fair classification. In *IEEE Symposium Series on Computational Intelligence, SSCI*, pp. 2993–3002.

Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G., & De Hert, P. (2022). Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. *IEEE Computational Intelligence Magazine*, *17*(1), 72–85.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3315–3323.

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2022). Explainable AI methods — a brief overview. In *International workshop on extending explainable AI beyond deep models and classifiers*, pp. 13–38. Springer.

Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM J. Responsib. Comput.*, *1*(2).

Hu, T., Iosifidis, V., Liao, W., Zhang, H., Yang, M. Y., Ntoutsi, E., & Rosenhahn, B. (2020). Fairnn — conjoint learning of fair representations for fair decisions. In *Discovery Science — 23rd International Conference, DS*, Vol. 12323, pp. 581–595.

Iosifidis, V., & Ntoutsi, E. (2018). Dealing with bias via data augmentation in supervised learning scenarios. *Jo Bates Paul D. Clough Robert Jäschke*, *24*.

Iosifidis, V., & Ntoutsi, E. (2019). Adafair: Cumulative fairness adaptive boosting. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM*, pp. 781–790.

Iosifidis, V., & Ntoutsi, E. (2020). FABBOO — online fairness-aware learning under class imbalance. In *Discovery Science - 23rd International Conference, DS*, Vol. 12323 of *Lecture Notes in Computer Science*, pp. 159–174.

Jorgensen, M., Richert, H., Black, E., Criado, N., & Such, J. (2023a). Not so fair: The impact of presumably fair machine learning models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, p. 297–311, New York, NY, USA. Association for Computing Machinery.

Jorgensen, M., Waller, M., Cocarascu, O., Criado, N., Rodrigues, O., Such, J., & Black, E. (2023b). Investigating the legality of bias mitigation methods in the United Kingdom. *IEEE Technology and Society Magazine*, *42*(4), 87–94.

Kamiran, F., & Calders, T. (2010). Classification with no discrimination by preferential sampling. In *Proceedings of the 19th Machine Learning*, Vol. 1.

Kamiran, F., & Calders, T. (2011). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In *ICDM, The 10th IEEE International Conference on Data Mining*, pp. 869–874.

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. In *12th IEEE International Conference on Data Mining, ICDM*, pp. 924–929.

Kamiran, F., Zliobaite, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, *35*(3), 613–644.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD*, Vol. 7524, pp. 35–50.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv:1609.05807.

Krasanakis, E., Xioufis, E. S., Papadopoulos, S., & Kompatsiaris, Y. (2018). Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW*, pp. 853–862.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm. Available at: `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`.

Lawrence-Archer, A., & Naik, R. (2023). Effective protection against ai harms. Available at: `https://www.awo.agency/files/AWO%20Analysis%20-%20Effective%20Protection%20against%20AI%20Harms.pdf`.

Lee, M. S. A., & Singh, J. (2021). The landscape and gaps in open source fairness toolkits. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pp. 699:1–699:13. ACM.

Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., & Puri, R. (2019). Bias mitigation post-processing for individual and group fairness. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 2847–2851.

Malgieri, G. (2019). Automated decision-making in the eu member states: The right to explanation and other "suitable safeguards" in the national legislations. *Computer law & security review*, *35*(5), 105327.

Mancuhan, K., & Clifton, C. (2014). Combating discrimination using bayesian networks. *Artificial Intelligence and Law*, *22*(2), 211–238.

Mansfield, T. (2023). Operational testing of facial recognition technology..

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 115:1–115:35.

Miller, T. (2022). Are we measuring trust correctly in explainability, interpretability, and transparency research?. arXiv:2209.00651.

Mittelstadt, B., Wachter, S., & Russell, C. (2023). The unfairness of fair machine learning: Levelling down and strict egalitarianism by default. arXiv:2302.02404.

Morgan, B., & Yeung, K. (2007). *An introduction to law and regulation: text and materials*. Cambridge University Press.

Nachbar, T. B. (2020). Algorithmic fairness, algorithmic discrimination. *Fla. St. UL Rev.*, *48*, 509.

Nadeem, M., Bethke, A., & Reddy, S. (2020). Stereoset: Measuring stereotypical bias in pretrained language models. arXiv:2004.09456.

Northpointe (2019). Practitioner's Guide to COMPAS Core. Available at: `https://s3.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf`.

Oneto, L., Donini, M., Elders, A., & Pontil, M. (2019). Taking advantage of multitask learning for fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES*, pp. 227–237.

P, D., & Abraham, S. S. (2020). Fair outlier detection. In *Web Information Systems Engineering (WISE)*, Vol. 12343 of *Lecture Notes in Computer Science*, pp. 447–462.

Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., Araujo, M. M., Santos, L. L., Cruz, M. A., Oliveira, E. L., et al. (2023). Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, *7*(1), 15.

Pardau, S. L. (2018). The california consumer privacy act: Towards a european-style privacy regime in the united states. *Journal of Technology Law & Policy*, *23*, 68.

Pessach, D., & Shmueli, E. (2023a). *Algorithmic Fairness*, pp. 867–886. Springer International Publishing.

Pessach, D., & Shmueli, E. (2023b). A review on fairness in machine learning. *ACM Computer Survey*, *55*(3), 51:1–51:44.

Powley, H., & Stanton, K. (2020). *Financial conduct in the UK's banking sector: Regulating to protect vulnerable consumers*, pp. 206–235. Routledge.

Putzel, P., & Lee, S. (2022). Blackbox post-processing for multiclass fairness. In *Proceedings of the Workshop on Artificial Intelligence Safety 2022 (SafeAI 2022) co-located with the Thirty-Sixth AAAI Conference on Artificial Intelligence*, Vol. 3087.

Quadrianto, N., & Sharmanska, V. (2017). Recycling privileged learning and distribution matching for fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pp. 677–688.

Quy, T. L., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, *12*(3).

Rosenblatt, L., & Witter, R. T. (2023). Counterfactual fairness is basically demographic parity. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*, 14461–14469.

Rösner, C., & Schmidt, M. (2018). Privacy preserving clustering with constraints. In *45th International Colloquium on Automata, Languages, and Programming, ICALP*, Vol. 107 of *LIPIcs*, pp. 96:1–96:14.

Rothstein, M. A., & Tovino, S. A. (2019). California takes the lead on data privacy law. *Hastings Center Report*, *49*(5), 4–5.

Roy, A., Horstmann, J., & Ntoutsi, E. (2023). Multi-dimensional discrimination in law and machine learning - A comparative overview. In *FAccT '23: Conference on Fairness, Accountability, and Transparency*, pp. 89–100. ACM.

Ruggieri, S., Alvarez, J. M., Pugnana, A., State, L., & Turini, F. (2023). Can we trust Fair-AI?. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(13), 15421–15430.

Ruoss, A., Balunovic, M., Fischer, M., & Vechev, M. T. (2020). Learning certified individually fair representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Sabbagh, D. (2011). Affirmative Action: The U.S. Experience in Comparative Perspective. *Daedalus*, *140*(2), 109–120.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2019). Aequitas: A bias and fairness audit toolkit. arXiv:1811.05577.

Sandanayake, T. C., Limesha, G. A. I., Madhumali, T. S. S., Mihirani, W. P. I., & Peiris, M. S. A. (2018). Automated cv analyzing and ranking tool to select candidates for job positions. In *Proceedings of the 6th International Conference on Information Technology: IoT and Smart City*, ICIT '18, p. 13–18. Association for Computing Machinery.

Savani, Y., White, C., & Govindarajulu, N. S. (2020). Intra-processing methods for debiasing neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., & Lin, H. (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Schaaf, N., de Mitri, O., Kim, H. B., Windberger, A., & Huber, M. F. (2021). Towards measuring bias in image classification. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks*, pp. 433–445. Springer.

Schelter, S., He, Y., Khilnani, J., & Stoyanovich, J. (2020). Fairprep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. In *Proceedings of the 23rd International Conference on Extending Database Technology, EDBT*, pp. 395–398.

Schöffer, J., Machowski, Y., & Kuehl, N. (2021). A study on fairness and trust perceptions in automated decision making. In Glowacka, D., & Krishnamurthy, V. R. (Eds.), *Joint Proceedings of the ACM IUI 2021 Workshops co-located with 26th ACM Conference on Intelligent User Interfaces (ACM IUI 2021)*, Vol. 2903 of *CEUR Workshop Proceedings*.

Shared Intelligence and Local Government Association (2020). Using predictive analytics in local public services..

Skitka, L., Mosier, K., & Burdick, M. (1999). Does automation bias decision-making?. *International Journal of Human-Computer Studies*, *51*(5), 991–1006.

Slack, D., Friedler, S. A., & Givental, E. (2020). Fairness warnings and fair-maml: learning fairly with minimal data. In *FAT* '20: Conference on Fairness, Accountability, and Transparency*, pp. 200–209.

Stănescu, C.-G., & Gikay, A. A. (2020). *Discrimination, Vulnerable Consumers and Financial Inclusion: Fair Access to Financial Services and the Law*. Routledge.

State of California Department of Justice (2024). California consumer privacy act (ccpa). Available at: https://oag.ca.gov/privacy/ccpa.

The United States Department of Justice (2015). The Fair Housing Act. Available at: https://www.justice.gov/crt/fair-housing-act-1.

Therese Enarsson, L. E., & Naarttijärvi, M. (2022). Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts. *Information & Communications Technology Law*, *31*(1), 123–153.

UK Government (2023). Positive action in the workplace. Available at: https://www.gov.uk/government/publications/positive-action-in-the-workplace-guidance-for-employers/positive-action-in-the-workplace.

UK Information Commissioner's Office (2021). The UK GDPR. Available at: https://ico.org.uk/for-organisations/data-protection-and-the-eu/data-protection-and-the-eu-in-detail/the-uk-gdpr/.

UK Public General Acts (2010). Equality Act 2010. Available at: https://www.legislation.gov.uk/ukpga/2010/15/contents.

United Nations. General Assembly (1949). Universal declaration of human rights. Available at https://www.un.org/en/about-us/universal-declaration-of-human-rights.

U.S. Department of Justice Civil Rights Division (2008). Americans with disabilities act of 1990, as amended. Available at: https://www.ada.gov/law-and-regs/ada/.

U.S. Department of Labor (1964). Civil Rights Act of 1964. Available at: https://www.archives.gov/milestone-documents/civil-rights-act.

U.S. Equal Employment Opportunity Commission (1967). The age discrimination in employment act of 1967. Available at: https://www.eeoc.gov/statutes/age-discrimination-employment-act-1967.

U.S. Equal Employment Opportunity Commission (1978). The pregnancy discrimination act of 1978. Available at: https://www.eeoc.gov/statutes/pregnancy-discrimination-act-1978.

Van Nuenen, T., Ferrer, X., Such, J. M., & Coté, M. (2020). Transparency for whom? Assessing discriminatory Artificial Intelligence. *Computer*, *53*(11), 36–44.

Veale, M., Kleek, M. V., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, p. 440. ACM.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, p. 1–7, New York, NY, USA. Association for Computing Machinery.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76–99.

Wachter, S., Mittelstadt, B., & Russell, C. (2020). Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. *West Verginia Law Review*, *123*, 735.

Wachter, S., Mittelstadt, B. D., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law Security Review*, *41*, 105567.

Waller, M., Rodrigues, O., & Cocarascu, O. (2023a). Bias mitigation methods for binary classification decision-making systems: Survey and recommendations. arXiv:2305.20020.

Waller, M., Rodrigues, O., & Cocarascu, O. (2023b). Recommendations for bias mitigation methods: Applicability and legality. In *Aequitas 2023: Workshop on Fairness and Bias in AI, co-located with ECAI 2023, Kraków, Poland*.

Waller, M., Rodrigues, O., & Cocarascu, O. (2024). Identifying reasons for bias: An argumentation-based approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(19), 21664–21672.

Waller, M., & Waller, P. (2020). Why predictive algorithms are so risky for public sector bodies. Available at SSRN 3716166.

Weinberg, L. (2022). Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ML fairness approaches. *Journal of Artificial Intelligence Research*, *74*, 75–109.

Wiggins, B. (2020). Calculating Race: Racial Discrimination in Risk Assessment. Available at: `https://academic.oup.com/book/0/chapter/287737508/chapter-pdf/39813059/oso-9780197504000-chapter-5.pdf`.

Willborn, S. L. (1984). The disparate impact model of discrimination: Theory and limits. *American University Law Review*, *34*, 799.

Xenidis, R. (2020). Tuning EU equality law to algorithmic discrimination: Three pathways to resilience. *Maastricht Journal of European and Comparative Law*, *27*(6), 736–758.

Xenidis, R. (2023). Beyond bias: algorithmic machines, discrimination law and the analogy trap. *Transnational Legal Theory*, *14*(4), 378–412.

Xiang, A., & Raji, I. D. (2019). On the legal compatibility of fairness definitions. arXiv:1912.00761.

Yang, J., Jiang, J., Sun, Z., & Chen, J. (2024). A large-scale empirical study on improving the fairness of image classification models. In Christakis, M., & Pradel, M. (Eds.), *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis, ISSTA 2024, Vienna, Austria, September 16-20, 2024*, pp. 210–222. ACM.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS*, Vol. 54, pp. 962–970.

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML*, Vol. 28, pp. 325–333.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES*, pp. 335–340. ACM.

Zhang, W., & Ntoutsi, E. (2019). FAHT: an adaptive fairness-aware decision tree classifier. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*, pp. 1480–1486.

Ziko, I. M., Yuan, J., Granger, E., & Ayed, I. B. (2021). Variational fair clustering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, pp. 11202–11209.

Zliobaite, I., Kamiran, F., & Calders, T. (2011). Handling conditional discrimination. In *11th IEEE International Conference on Data Mining, ICDM*, pp. 992–1001.