A Review of Bias and Fairness in Artificial Intelligence

Rubén González-Sendino[1], Emilio Serrano[1], Javier Bajo[1], Paulo Novais[2] *

[1] Ontology Engineering Group, Departamento de Inteligencia Artificial, ETSI Informáticos, Universidad Politécnica de Madrid, 28660 Madrid (Spain)
[2] ALGORITMI Research Centre/LASI, University of Minho, Braga (Portugal)

Received 16 September 2022 | Accepted 29 September 2023 | Early Access 10 November 2023

Abstract

Automating decision systems has led to hidden biases in the use of artificial intelligence (AI). Consequently, explaining these decisions and identifying responsibilities has become a challenge. As a result, a new field of research on algorithmic fairness has emerged. In this area, detecting biases and mitigating them is essential to ensure fair and discrimination-free decisions. This paper contributes with: (1) a categorization of biases and how these are associated with different phases of an AI model's development (including the data-generation phase); (2) a revision of fairness metrics to audit the data and AI models trained with them (considering agnostic models when focusing on fairness); and, (3) a novel taxonomy of the procedures to mitigate biases in the different phases of an AI model's development (pre-processing, training, and post-processing) with the addition of transversal actions that help to produce fairer models.

Keywords

Bias, Fairness, Responsible Artificial Intelligence.

DOI: 10.9781/ijimai.2023.11.001

I. Introduction

The evolution of artificial intelligence (AI) has allowed humans to be heavily supported in the decision-making process of some application domains [1]. The high degree of independence that AI is capable of exhibiting can be problematic [2], [3], especially when humans are not in the loop [4]–[6]. Automatization of decisions can come at the cost of amplifying bias and creating feedback loops [7], [8]. One of the main reasons AI can produce unfair results is due to the data with which it has been trained [9].

Although the concept of *bias* is broad, this paper adheres to the following definition: "the systematic tendency in a model to favor one demographic group/individual over another, which can be mitigated but may well lead to unfairness" [9], [10]. Therefore, the next definition needed to understand the problem this paper studies is *Fairness*, which is defined as: " the absence of prejudice or favoritism towards an individual or a group based on its inherent or acquired characteristics" [9].

In the AI scope, incorrect predictions do not necessarily indicate that the model is unfair if its development was correct [11]. An unfair model is one whose decisions are biased toward a particular group of people. Moreover, biases cannot always be avoided. Thus, techniques must be used to mitigate their consequences, which aim to increase equality in the results. Data and models can be audited with *fairness metrics*, which are used to measure fairness between two groups or similar individuals. Furthermore, the categorization of methods for bias and unfairness mitigation depends on the phase of the AI model's development in which they are used. These phases are typically pre-training, training, and post-training.

* Corresponding author.
E-mail addresses: ruben.gonzalez.sendino@alumnos.upm.es (R. González-Sendino), emilio.serrano@upm.es (E. Serrano), jbajo@fi.upm.es (J. Bajo), pjon@di.uminho.pt (P. Novais).

This paper contributes with a systematic review of bias and fairness in artificial intelligence. The purpose of a systematic review is to provide a comprehensive summary of the literature available which is relevant to several research questions. The three questions addressed in this paper are: (1) What bias affects fairness?; (2) What are the metrics to measure fairness?; and, (3) How are biases mitigated? Beyond this systematic review and the taxonomy mentioned, the final goal of this paper is to help developers and researchers identify new biases and create fairer AI models.

This systematic review differs from others by addressing the three essential questions together. Previous systematic reviews have focused primarily on measurement and mitigation, complete systematic reviews on these fields are [12]–[14]. However, this review expands upon these works by including tools for auditing algorithms (Section VI) and guidelines for fair governance (Section VII). The bias affecting AI learning has been discussed in various papers, enumerating the different types or relating with recommender or scoring algorithms [9], [15], [16]. In this review, the bias has been organized into four general categories that apply to any type of AI system.

The content of the paper is organized as follows. Section II gives the background and motivation of the paper. Section III details the search criteria for the systematic review. Section IV shows the works retrieved for review. Then sections V, VI and VII respectively offer answers to the three research questions considered. In the context of the third research question ("How are biases mitigated?"), a taxonomy of the procedures to mitigate biases in the different phases of an AI model's development is presented. To finalize, Section VIII discusses the results and Section IX concludes and gives an overview of future work.

II. Background

The use of ML-based decision-making algorithms in organizations is increasing rapidly [17]. These algorithms may generate results that reflect, reproduce, and amplify structural inequalities. The results

Please cite this article in press as:
R. González-Sendino, E. Serrano, J. Bajo, P. Novais. A Review of Bias and Fairness in Artificial Intelligence, International Journal of Interactive Multimedia and Artificial Intelligence, (2023), http://dx.doi.org/10.9781/ijimai.2023.11.001

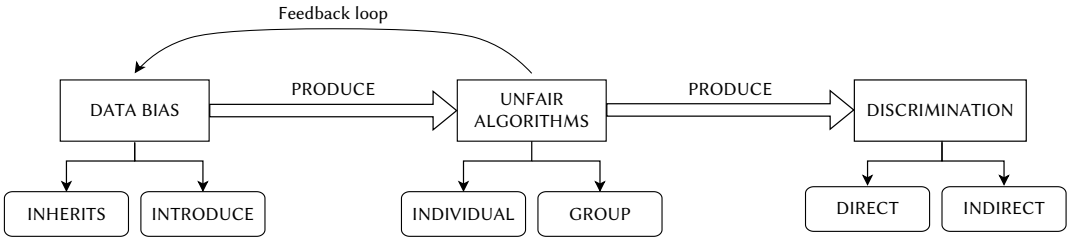

Fig. 1. Transformation of the bias in discriminatory results.

can be the product of unjust goals rooted in racist, sexist, heteronormative, nationalist, or colonialist priorities [18].

These algorithms are becoming increasingly complex and *deep neural networks* (DNN) play an indispensable role in most AI-assisted tasks. These systems are "black boxes" due to the lack of transparency and explainability they exhibit. Therefore, DNNs can hide potential biases and present unexpected vulnerabilities [19], [20].

Due to previous problems, *Explainable Artificial Intelligence* (XAI) is becoming more necessary every day. Among others, XAI covers understandability, comprehensibility, interpretability, explainability, and transparency. These dimensions of XAI are essential to support and understand when discrimination occurs (explainability [21]–[24], interpretability [6], [25], [26] and transparency [1], [11], [20], [27]). XAI topics encourage *responsible AI* that considers fairness, privacy, accountability, ethics, transparency, security, and safety) [28].

Governments are focusing on explaining the decisions of autonomous systems to users. In Europe, new regulations give European citizens the right to have basic knowledge of the inner workings of automated decision-making models and to question their results [12], [29]. In Spain, there exists a similar law (BOE-A-2021-7840). This rule encourages the auditing of decisions made by an automated system. Since 2021, companies have had to inform their employers about the parameters, rules, and instructions that influence the algorithm's decisions.

There have been different incidents in which algorithms have produced unfair results. For example, the United Kingdom used an algorithm to infer the Advanced Level exam results for those students who could not take their tests[1]. In doing so, the algorithm considered the background of the students, their partners, or the school they attended. This resulted in disadvantaged ethnic minorities and people from poorer or disadvantaged backgrounds.

In the past, humans were discriminated against due to stereotypes, prejudice, or unintentional bias [30]. However, algorithms do not discriminate because they do not have the mental capacity to do so. In many cases, the problem is that human biases are transferred to the model by training data. The bias in a dataset can easily lead to an erroneous or discriminatory conclusion [31]. However, biases in AI algorithms can result not only because of issues with the training data but also from how algorithms learn over time and are used in practice [32].

Fig. 1 shows how biases are transformed into discrimination. Biases can be inherited (and later perpetuated) or introduced (and then exacerbated). The inherited bias perpetuates the existing inequality in the data structure [33]. Furthermore, bias can be introduced by assumptions in model implementation that exacerbates discrimination [34], [35]. The discrimination produced can be direct (disparate treatment) or indirect (disparate impact) [36]. Direct discrimination occurs when individuals receive less favorable treatment based on protected attributes such as sex, religion, or nationality [8]. Indirect discrimination occurs when people receive treatment based on inadequate factors. These factors are generally related to protected attributes [8]. Disparate impact is defined as a neutral rule that applies to everyone. However, the effect is more harmful to some people than to others [7].

Note that bias is subjective and related to the task. The healthcare results are not discriminatory if the diagnosis is based on sex-specific symptoms. However, the results of a hiring process could be discriminatory if it is sex-biased [7]. Note also that the problem of unfairness does not have to be addressed by necessarily reducing the use of AI. Sometimes, AI has been perceived as fairer than a human expert in the context of health and justice decisions [37]. AI decisions are made based on knowledge, unlike human decisions, which can be based on feelings.

To address bias, unfairness, and discrimination; AI must be audited following a procedure that involves: (1) the identification of potential biases that can affect fairness; (2) the selection of metrics to measure how fair AI is being; (3) and, the mitigation of the impact produced by these biases [32], [38], [39]. This paper conducts a systematic review of the state of the art with respect to these three key aspects.

III. Search Criteria for the Systematic Review

The main objective of this systematic review is to understand and analyze the fairness and bias in AI algorithms. To obtain a more detailed and comprehensive view of the field, the review examines the following three research questions (RQs):

- RQ1. What bias affects fairness?
- RQ2. What are the metrics to measure fairness?
- RQ3. How can biases be mitigated?

To strengthen the validity of the review, inclusion criteria (IC) for the studies included in this systematic review have been applied. These criteria and their justification are presented below.

- IC1. Studies must be peer-reviewed articles published in conferences, journals, press, etc.
- IC2. Studies must be conducted primarily in English.
- IC3. Studies must have been published since 2010.
- IC4. The abstract, introduction, and conclusion provide enough information.
- IC5. Studies that address the concept of fairness in artificial intelligence algorithms.
- IC6. Studies that provide enough information about bias as a product of unfairness.

This systematic review seeks research on artificial intelligence that studies the unfair results produced by bias. The search queries are considered as a base including the terms: fairness, artificial intelligence, machine learning, and bias. Furthermore, the terms FAT (fairness, accountability, and transparency) and FATE (fairness, accountability, transparency, and explainability) are considered relevant to find related research. XAI and responsible AI are excluded because these topics are considered transversal to the research questions studied and therefore can include noise in the systematic review.

[1] https://bit.ly/3dbxdbu

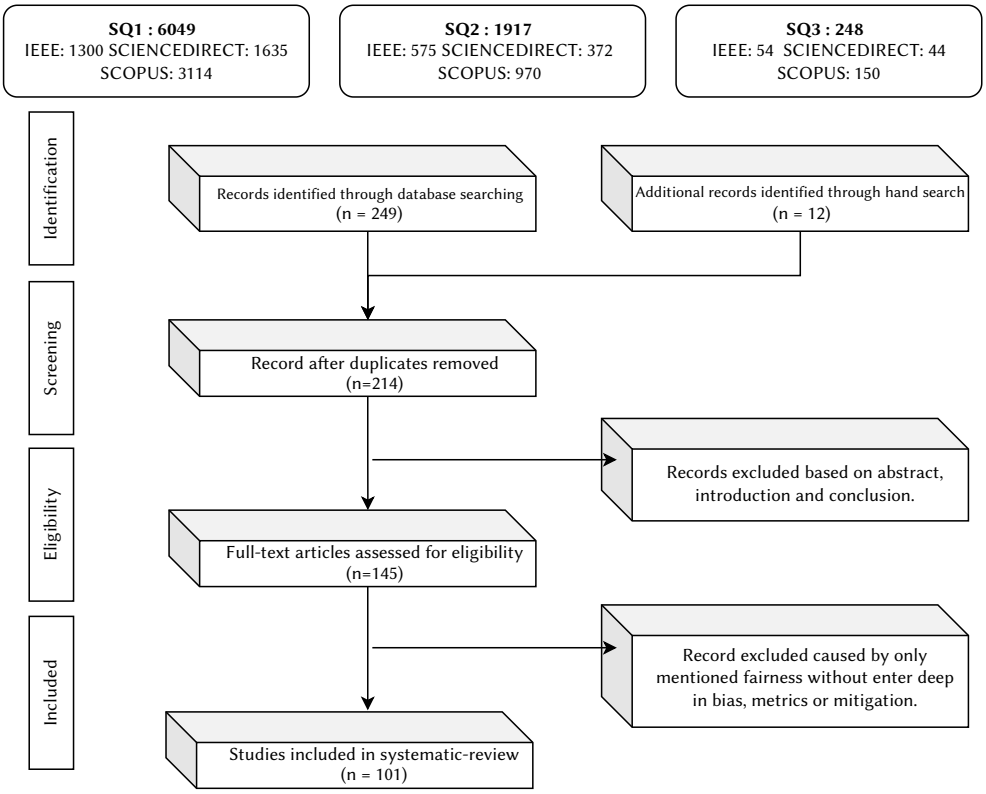

Fig. 2. PRISMA flow diagram depicting the flow of information through the different phases of the systematic review.

The systematic review focuses on the latest biases, metrics, and mitigation techniques. For this reason, the search query applies a filter by date; only articles published since 2010 will be considered. The scientific libraries used for the collection of articles will be Science Direct, Scopus, and IEEE.

The search query (SQ) will be applied only to: the title, abstract or author-specified keywords.

- SQ1 is ((‘artificial intelligence’ or ‘machine learning’) and (‘fair’ or ‘fat’ or ‘fate’ or ‘bias’ or ‘fairness’ or ‘unfair’)).
- SQ2 is ((‘artificial intelligence’ or ‘machine learning’) and (‘fair’ or ‘fat’ or ‘fate’ or ‘fairness’ or ‘unfair’)).
- SQ3 is (‘artificial intelligence’ or ‘machine learning’) and (‘fair’ or ‘fat’ or ‘fate’ or ‘fairness’ or ‘unfair’) and (‘bias’).

The three queries are used to filter from a complete set to a subset that collects the desired papers $SQ1 \supset SQ2 \supset SQ3$. The articles filtered by *SQ3* contain both topics: fairness and bias. This requirement reduces the noise caused by the broadness of the concept of bias.

IV. Retrieved Works for the Systematic Review

The three search queries used to find related research produce distinct results in terms of quantity. Fig. 2 shows the number of results for each search.

SQ1 returned thousands of studies, most of them related to an unbalanced dataset without linking it with fairness. A large number of the SQ2 results talked about fairness in algorithms, as a top-level definition without diving deep into the topic. Finally, SQ3 is the search query used to filter and read information related to fairness and bias for machine learning algorithms.

Note that the solution to bias could be similar to the solution to having unbalanced data. However, in the latter scenario, the purpose is to improve the accuracy of the model while in the former the objective is to reduce unfairness. Several studies tried to find a trade-off between accuracy and fairness.

Fig. 2 shows a flow chart with the filters applied to reduce the number of papers included in this systematic review. After applying the search queries and the inclusion criteria, the number of research works considered is 101.

Fig. 3 shows that the number of publications related to this research is growing significantly and has gained relevance in the last five years. For this reason, having skipped papers before 2010 does not seem to be a problem.

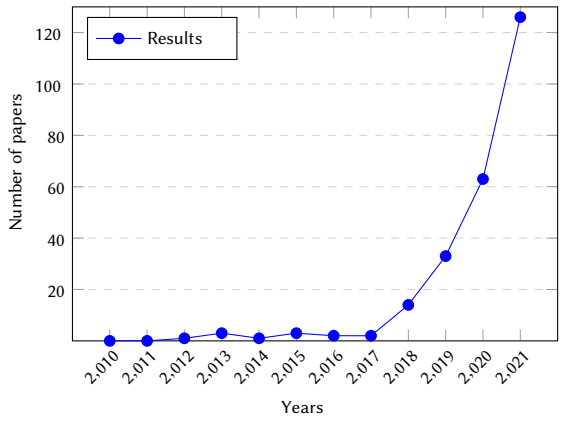

Fig. 3. Graph showing the evolution of publications related to bias and unfairness in artificial intelligence algorithms.

Research can be grouped into three main categories: fields in which there is a concern with bias; algorithms in which the fairness metrics are trying to be mitigated; and theoretical studies about bias and its mitigation. The fields help to understand where researchers face unfairness, thus discovering why biases affect the fields. Additionally,

TABLE I. Important Field and Algorithms Where Some Studies Are Centered

Field	Nº
Health [10], [19], [32], [40]–[46]	10
Education [9], [20], [47], [48]	4
Recruitment [17], [49], [50]	3
Travel [33]	1
Manufacturing [2]	1
Laws [1]	1
Public service [51]	1
Rent house [38]	1

Field	Nº
Deep Neural Networks [33],[34], [52]–[55]	5
Ambient intelligence [56]–[60]	5
NLP [7], [61]–[63]	4
Computer Vision [31], [64],[65]	3
GAN [66], [67]	2
Decentralizing Learning [68], [69]	2
Support Vector Machine [70], [71]	2
Decision tree [20], [72]	2
Adaptive models [32]	1
XGBoost [22]	1

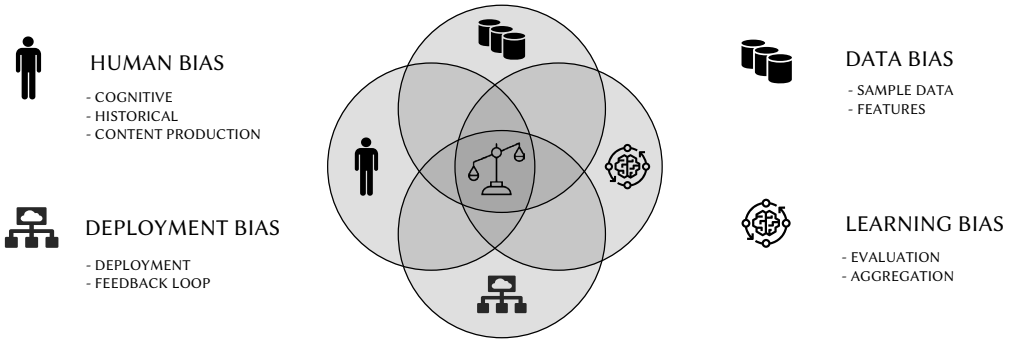

Fig. 4. An overview of bias impacting fairness.

understanding the deficiencies of the learning algorithm could help to understand the complexity of obtaining a fair output.

Table I shows the fields and algorithms most popular in this review. Health is the most relevant field in which research seeks equity. The next most crucial field is recruitment and education. This focus of the studies is understandable because of the importance of decisions in these fields. Table I also details the main algorithms used in the retrieved works in the specialized literature. Neural networks are the most widespread learning paradigm. The use of these networks is generally: decision-making, recommendation, scoring, or classification.

There are publications that address the topics of fairness and bias from a theoretical and transversal perspective in artificial intelligence. The majority of these papers talked about mitigation techniques [12]–[14], [22], [23],[73]–[76], additionally the topics of fairness metrics [22],[23], [29], [74], [76] and types of bias [15], [16], [73],[77], [78] are covered.

V. RQ1: What Bias Affects Fairness?

This section provides an overview of the biases that affect fairness. Fig. 4 shows the four phases in which the biases can be grouped. The groups to identify biases are linked to the model life cycle: production (human) bias, data bias, learning bias, and deployment bias.

The relevant consequence of biases is discrimination, reflected in the tendency to favor one individual or group over another [10]. Fig. 1 illustrates the transformation of bias into discrimination using an unfair model.

A. Human Bias

Human beings are the main factor that produces bias. For this reason, the production bias group in Fig. 4 is called *human bias*. Human-made decisions that are reflected in the data or in the model. Therefore, eliminating bias in machine learning and Artificial Intelligence without addressing the pressing concerns about bias in humans is not possible [76].

The human bias group in Fig. 4 collects how biased data are generated. Human decisions can lead to unfairness in a number of steps in AI development discussed below: data management, learning, and model deployment. These steps will be compiled in the other groups. Human biases are divided into two main subgroups: cognitive bias, and behavioral bias. Note that the information is susceptible to variation over time, producing *temporal bias* [77].

- *Cognitive bias* is a deeply ingrained part of human decision making [73], which transfers prejudices to labels [79]. Machine learning algorithms use human judgments as training data, so they propagate these biases. This *historical bias* arises even when the data is perfectly measured and sampled, for example, reinforcing a stereotype [7].
- *Behavior bias* produces distortions from reality or other applications according to user connections, activities, or interactions [9]. Furthermore, unconscious bias could be produced by *content creation*, because the way a child or an adult expresses themselves is different, in the same way as if you compare by sex or race [61]. Content creation bias is also produced when users are guided by norms or functionalities [9], and sometimes these interactions are led by an AI system [32]. Bias production in the future will be affected by unfair systems, generating new data to be used in future learning [9].

B. Data Bias

Data bias focuses on the factors that induce a biased dataset. The main tasks where data bias may occur are acquisition, querying, filtering, transforming, and cleaning [9].

Data bias is generally a distortion in *sampled data* that compromises its representatives [15], [41]. In other words, sample bias is produced when the train and test data do not represent or under-represent a population segment. Therefore, modelers play a critical role in producing data bias by including or discarding data [35], and even labeling the data [9].

The *representation* is not the only problem related to the data. *Features* could reflect discrimination by sensitive attributes: race, sex, age, socio-economic data, education, neighborhood, etc. These features are generally not legitimate for decision-making [80]. Typical sensitive variables are those collected by the General Data Protection Regulation (GDPR).

Fairness is not only attributable to protected attributes. Proxy attributes can be exploited to derive sensitive features [14]. Generally, protected characteristics and targets are highly correlated, resulting in good accuracy at the cost of diminishing fairness metrics [33]. The correlation applied as causality could lead to bias [9].

C. Learning Bias

Learning biases are produced in model training. This step is compounded by the difficulties of understanding and explaining the results. Typically, models learn a correct statistical pattern in favor of the majority over minorities [15], [33], leading to *aggregation* bias that amplifies the disparities between different examples in data samples [7]. At this point, the model that learns from generated experience, such as reinforcement learning, could become biased over time [32].

The performance of the model should be evaluated after each training loop. *Evaluation bias* occurs when the selected metrics are not appropriate, for example, the use of general vs. subgroup accuracy [77], and when the representation in the test data does not reflect reality.

D. Deployment Bias

Deployment bias may occur in the deployment and use of the model. The algorithm makes decisions based on patterns learned from the data. Therefore, *the deployment* of a model in a different scenario with respect to data could lead to unfair results [10]. Feedback loop is the result of the introduction of a new discriminatory decision in the data [16].

VI. RQ2: What Are the Metrics to Measure Fairness?

This section collects approaches to measure fairness. Measurement of fairness gives a quantification of the problem for further mitigation. Although these issues are most apparent in the social sciences, where fairness is interpreted in terms of the distribution of resources across protected groups, the management of bias in source data affects a variety of fields. Any domain involving sparse or sampled data is exposed to potential bias [26].

Typically, metrics used to measure fairness are divided into two groups. On the one hand, metrics measure and find the difference in equality between two selected groups. On the other hand, metrics can compare results between similar individuals whose results are disparate [81]. The final goal of both is to find discriminatory inputs [81].

- *Group Fairness or Disparate Impact.* Each group identified in the dataset receives an equal fraction of a possible outcome (applies to both positive and negative outcomes) [29], [82]. In other words, different sensitive groups should be treated equally. The two groups usually are called the Unprivileged Group (UG) and the Privileged Group (PG).
- *Individual Fairness or Disparate Treatment.* Individuals who belong to different sensitive groups with similar characteristics should be treated similarly [29], [82]. For example, applicants with the same qualifications during job applications should not be discriminated against based on their sex or race. Some positions highlight that individual fairness cannot be a definition of fairness due to: insufficiency of similar treatment, systematic bias and arbitrators, and prior moral judgments [83].

The most common metrics used to measure fairness are shown in Table II. Those metrics require ground-truth data. Other metrics are used for unsupervised problems such as Fairness Demographic Parity, Point-wise Mutual Information, Kendall Rank Correlation, t-test, and Log-likelihood Ratio [84].

A less extended classification divides algorithms into: statistics based on predicted outcomes, statistics based on predicted and actual outcomes, statistics based on predicted probabilities and actual outcomes, and similarity-based and causal reasoning [7], [36], [85]. The two most employed are: statistics based on predicted outcomes and statistics based on predicted and actual outcomes.

Statistics based on predicted outcomes can be defined as statistical parity, conditional statistical parity, and predictive equality. Furthermore, statistics based on predicted and actual outcomes can be described as calibration within groups, balance for the negative class, and balance for the positive class. Not all algorithms can satisfy all conditions simultaneously. The objective is a trade-off between the ability to classify accurately and the fairness of the resulting data [36].

A significant thing about the status of bias detection is that current strategies for detecting biases are often customized for a problem, dataset, or method [88]. This affects their generalization [88]. (1) For group fairness, there exist simple studies that use Receiver Operator Characteristics (ROC) curves for each demographic group [10], [31]. Other traditional metrics used to calculate fairness are standard deviation and skewed error ratio [65]. (2) For individual fairness, there exists Procedural Fairness [29] and Consistency Metric [22]. Procedural fairness ensures that the algorithm does not use sensitive features for prediction. Consistency Metric compares the prediction of a certain individual with the predictions of its k-nearest neighbors. The bias disparity is a concept introduced in Aequitas [89]. This bias is calculated by comparing the metric for a given group with the metric of the reference group. The metrics that could be calculated are: predicted positive, total predictive positive, predicted negative, predicted prevalence, false positive, false negative, true positive, false negative, false discovery rate, false omission rate, false positive rate, and false negative rate.

Finally, the existence of agnostic models helps to comply with responsible artificial intelligence. The most common agnostic models include explainability methods. There exist specific agnostic models that focus on helping to improve fairness in the different development phases: AIF360 [23], FairLearn [44], [90], LFIT [50], Aequitas [89], LimeOut [29], MAML [67], the What-If toolkit (WIT) [91], and Audit AI [92].

The previous toolkits help to audit machine learning models for discrimination and bias. The most popular are Aequitas (Carnegie Mellon University), AIF360 (IBM), FairLearn (Microsoft) and WIT (Google). Some of these tools, in addition to measure, also help with mitigation. This is the case of FairLearn and AIF360.

Aequitas and WIT are especially suitable as audit tools. WIT provides a graphical interface in which the behavior of an algorithm can be tested visually. The tool integrates the fairness indicator developed in TensorFlow. Aequitas runs a full report on biases. This report is expected to be used by developers, analysts, and policymakers.

VII. RQ3: How to Mitigate Bias?

As a result of the systematic review of this research question, a taxonomy has been proposed to aggregate bias mitigation procedures. Fig. 5 displays this new taxonomy which is considered from the point of view of data science. The stages where mitigation techniques can be applied include pre-training, training, and post-training.

- Mitigating bias in the pre-training phase is the most effective manner of correcting bias since it transforms the dataset. However, bias may appear after training, hindering developers from dealing with it in the first iteration of the process [73].
- Training is the most efficient stage for handling bias. These methods are often unsupervised and do not involve adulterating the underlying data set [73]. Not including sensitive features such as gender or race is not enough to mitigate discrimination, considering that other derivative features are introduced. Instead, adding fairness to the objective function is more efficient [93].
- Post-training is an ideal phase to calculate most of the previously revised metrics [73]. However, mitigating biases in this phase should be the last option [67].

There is a limitation with pre-processing and post-processing because manipulating the data leads to an outcome that may not be realistic due to the perturbation of the original distribution [53].

A fair governance category is also included in the taxonomy, where mitigation is possible without applying complex algorithms.

A. Pre-Training

Pre-processing modifications could be made to the sample or features and labels to produce fair data that neutralize discriminatory effects [65]. However, this approach cannot eliminate discrimination that may come from the algorithm itself [7]. The taxonomy shown in Fig. 5 reflects the three main techniques studied in the literature: resampling, fair representation, and re-weighting.

1. Resampling

Resampling is used to change the size of the data set that affects the distribution without transforming the data. Resampling methods are divided into undersampling and oversampling [65], [67]. Undersampling techniques are based on eliminated samples from the dataset; meanwhile, oversampling means generating (or repeating) data samples to augment the original dataset.

These techniques have been transferred from data-balancing problems to the fairness domain. In fairness mitigation, different algorithms are tested for data augmentation, while techniques for undersampling are less popular [94]. Successive data augmentations may be computationally expensive if the dataset contains many features [79].

The two hegemonic approaches to oversampling are: the Synthetic Minority Oversampling Technique (SMOTE); and, the Generative Adversarial Networks (GANs) [65]. GANs have been used to produce synthetic tabular data to improve demographic parity [66], allowing fairness to be increased while maintaining precision in prediction.

In addition to altering the number of samples, another approach to improve fairness can be to reduce the number of features in the data by *feature selection*. A simple method is to remove sensitive features that could produce bias in prediction. However, this is not enough, since protected attributes could be encoded or correlated with other features [12].

2. Fair Representation

Fair representation is obtained by eliminating information that can link a person to a protected group [73]. *Learning fair representation* (LFR) is a popular algorithm for finding a latent representation that encoded data while preserving fairness [22]. Protected information can be hidden or explicit, giving more or less weight to its representation [76]. However, LFR improves fairness at the cost of complicating the explainability of the results [74].

3. Re-Weighting

Re-weighting is the method more widely used to transform the data by modifying the weight in the data set [75]. Note that not all learning algorithms accept weighted samples [73]. Re-weighting means that certain instances from a privileged group, more likely to have a favorable outcome, will get a lower weight. Similarly, instances of an unprivileged group will receive a higher weight [22].

4. Other Categories of Bias Mitigation in Preprocessing

In addition to the three main categories explored above, less popular methods used to mitigate bias in the preprocessing phase include: Privileged Group Selection Bias (PGSB) [13], [14], disparate impact remover [23],[76], and optimized preprocessing [23], [74].

B. Training

When mitigating biases in training time, algorithms are modified to improve fairness rather than just precision. The advantage of addressing biases in this phase see Fig. 5, is that data and prediction can be used to evaluate fairness. Regularization and adversarial training are the most common methods for this purpose according to the revised literature. Other emerging approaches are: decentralized learning, fair linear regression, fair-n, DeepFair, multimodal models, and fairlet clustering. These approaches are discussed below.

1. Regularization

Regularization is a well-known technique in machine learning. Regularization is used to correct underfitting or overfitting when training the model. This method can also be used to mitigate biases and unfairness [73]. In contrast, adding regularization methods to a machine learning model can complicate the explanation and interpretation of its results [22]. Regularization for mitigating biases can be: implicitly adding constraints that disentangle the association between model predictions and sensitive attributes; or explicitly adding constraints by updating the model loss function to minimize the performance difference between different protected groups [65].

Regularization methods in the loss function of deep neural networks can help reduce the difference in prediction disparity between different groups [33]. Regularization can also penalize high correlations between sensitive attributes and outcomes. The following reports [13], [14], [95] employ L2 regularization to weight examples equally in several machine learning models, such as support vector machines (SVMs) and logistic regressions (LRs).

2. Adversarial Training

In the field of AI, *Adversarial Learning* is a technique in which multiple neural networks compete with each other to improve the predictive accuracy [96]. The fairness of machine learning models can be improved by mitigating bias through the use of adversarial learning; this process is called *adversarial debiasing* [31].

Adversarial debiasing involves training two neural networks where one network learns to predict the outcome, and the other network identifies and removes any biases in the training data that could affect the prediction of the first network. The second network, also known as the "adversary", attempts to find and exploit weaknesses in the first predictions, thus forcing the first network to become more robust and resistant to bias [97].

In adversarial debiasing, the goal is to reduce evidence of any biases related to protected attributes in the predictions [73]. Evidence of protected attributes can be reduced, and prediction accuracy can also be improved in certain cases [75]. Scores between different demographic groups can be balanced, promoting: demographic parity, equality of odds, and equality of opportunity [97].

TABLE II. The Most Popular Metrics to Measure Fairness

Metric name	Target	Definition
Equal Opportunity Difference (EOD)[22], [23], [29], [74],[76], [82], [86]	Group	Measures the difference in true positive rates (TPR) between an unprivileged group and a privileged group. $\text{TPR} = \left[\frac{\text{True Positive (TP)}}{\text{TP} + \text{False Negative (FN)}}\right]$ (1) $\text{EOD} = \text{TPR}_{\text{UG}} - \text{TPR}_{\text{PG}}$ (2)
Odds Difference (OD) [22], [23], [74], [76], [79]	Group	Computes the difference of false positive rate (FPR) and true positive rate (TPR) between unprivileged and privileged groups. $\text{FPR} = \left[\frac{\text{False Positive (FP)}}{\text{FP} + \text{True Negative (TN)}}\right]$ (3) $\text{OD} = (\text{FPR}_{\text{UG}} - \text{FPR}_{\text{PG}}) + (\text{TPR}_{\text{UG}} - \text{TPR}_{\text{PG}})$ (4)
Statistical Parity Difference (SPD) [22], [23], [74], [75],[75], [76], [79], [86]	Group	Calculates the difference in the probability of favorable results (Predicted as Positive (PPP)) between the unprivileged group and the privileged group. $\text{PPP} = \left[\frac{\text{TP} + \text{FP}}{\text{Total Population (N)}}\right]$ (5) $\text{SPD} = \text{PPP}_{\text{UG}} - \text{PPP}_{\text{PG}}$ (6)
Disparate Impact (DI) [22], [23], [74],[76]	Group	Compares the proportion of individuals who receive a positive output for two groups: an unprivileged group and a privileged group. $\text{DI} = \frac{\text{PPP}_{\text{UG}}}{\text{PPP}_{\text{PG}}}$ (7)
Theil Index (TI) [22],[23], [87]	Group/ Individual	Subclass of the generalized entropy index (using alpha = 1). The entropy index is a measure of inequality in a group or individual with respect to the fairness of the algorithm outcome. $\text{TI} = \frac{1}{N}\sum_{i=1}^{N}\frac{b_i}{\mu}\ln\frac{b_i}{\mu}, b = \textit{predicted} - \textit{labeled}$ (8)

3. Emerging Approaches

Decentralizing the learning is an application of the blockchain strategy to machine learning models where models are built with a distributed and collaborative approach. In this scheme, some methods that have been examined from the point of view of fairness in the literature are: Swarm Learning (SL) [69] and Federated Learning (FL) [68].

The *Fair linear regression* and *Fair-n* approaches are useful for cases with more than one sensitive variable. *Fair linear regression* is based on the Hilbert-Schmidt independence criterion. This allows it to deal with several sensitive variables simultaneously [93]. *FaiR-N* introduces fairness and robustness regularization techniques to the loss function based on an approximation of the distance of data points to the decision boundary during training [53].

Fairlet clustering can be used in cases with unsupervised data [71], where detecting bias is a complex task. In clustering problems, fairness is defined in terms of consistency in that the balance ratio of data with different sensitive attribute values remains constant for each cluster.

DeepFair is a solution for a recommender system [55]. The recommender system relies on Collaborative Filtering (a set of the user's preferences on the items). This amount of data affects minorities negatively. The solution proposes a Deep Learning based on Collaborative Filtering that provides recommendations with an optimal balance between fairness and accuracy.

Multimodal models can understand and process information from multiple heterogeneous sources of information that can help reduce or correct bias and unfairness [49]. However, in multimodal model, detecting the origin of the bias is a very challenging task [17].

C. Post-Training

When addressing bias in this phase, the results of the model are modified by correcting decisions that could harm the fair representation of different subgroups in the final decision process [7]. As shown in Fig. 5, the most commonly used methods in post-training include: equalized odds, calibrated equalized odds, and reject option classification.

- *Equalized odds* adds a post-learning step to determinate optimal probabilities to change output labels. Equalized odds enforce fairness and precision [22], [23], [65],[73]–[75].
- *Calibrated equalized odds*, starting from the score outputs of a calibrated classifier, optimizes the probabilities with which to change the output with an equalized odds objective [23], [65], [73]–[75].
- *Reject option classification* gives favorable outcomes to protected unprivileged groups and unfavorable outcomes to privileged ones. This method uses a confidence band around the decision boundary with the highest uncertainty [22], [23], [65], [73]–[75].

D. Fairness Governance Practices

Fair governance minimizes biases while avoiding mitigation methods revised above for pre-training, training, and post-training. As shown in Fig. 4, the three main areas in which unfair results are reduced are: team, data, and models.

1. Team

This section refers to the people involved in developing an artificial intelligence algorithm. The impact of the modeler is not the only one. For example, imposed precision requirements may go against fairness indicators.

Diversity is the first dimension of the teams that needs to be improved [98]. Furthermore, this diversity has to affect all levels of the hierarchy [4]. Until then, algorithms and their associated biases will become mirrors of structural discrimination rather than bridges to opportunity, equality, and efficiency [99].

Creating diverse teams, as well as *cross-disciplinary* teams of data scientists and social scientists [5], is also essential to reducing bias and unfairness.

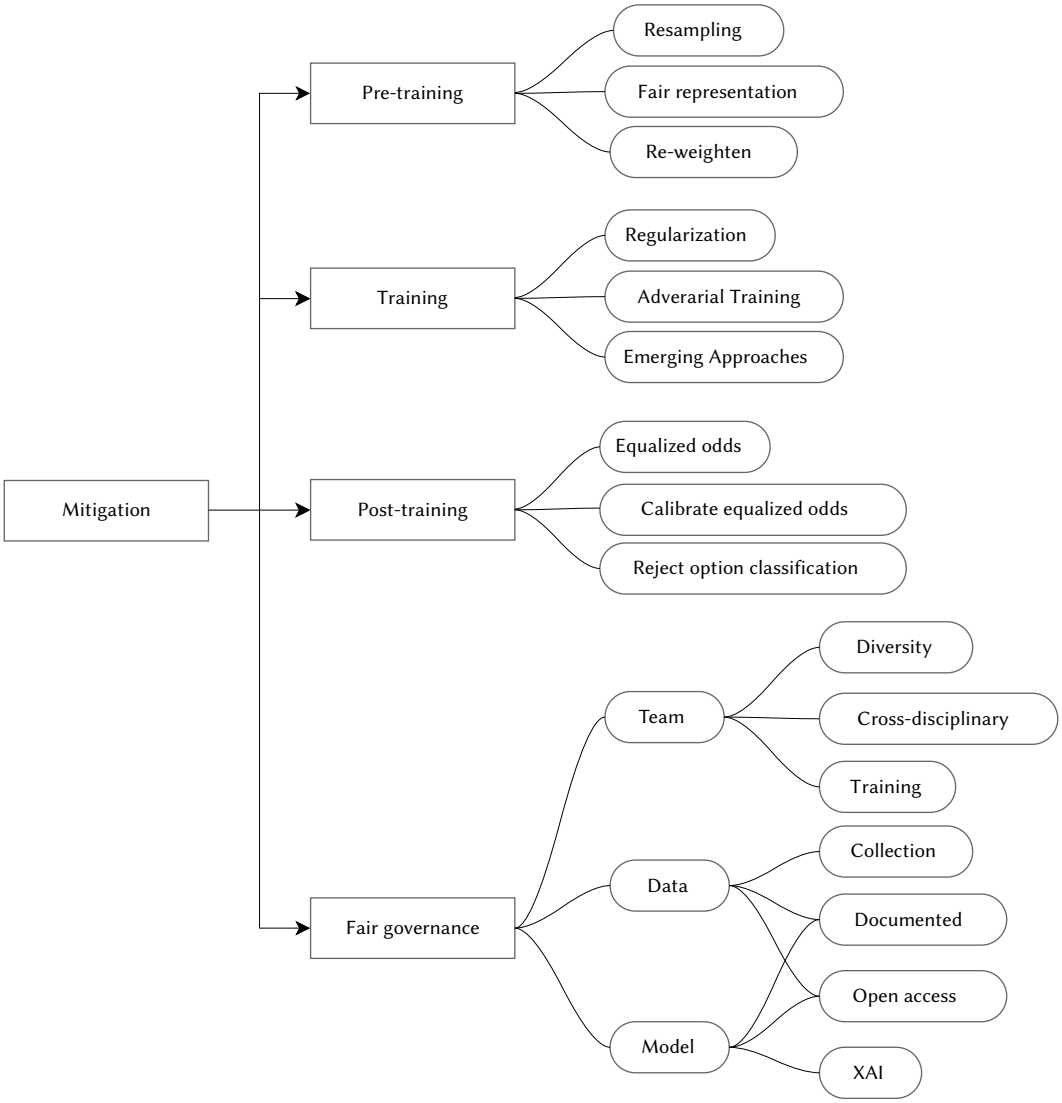

Fig. 5. Taxonomy of bias mitigation.

To further improve AI development, continuous training in fairness and ethics is recommended for team members, and stakeholders [4]. In addition, cultivating emotional intelligence to process better, identify, and confront discordance [6]. Tools, such as packages to calculate and audit bias, can be an important aid in addressing and standardizing this problem [9].

2. Data

The data is where the knowledge resides, and the algorithms will learn that information. Important keys at this point could be selecting features, transformations, or data labelling.

Data collection can be improved by prioritizing data on sex, race, ethnicity, etc. As a result, complete information related to sensitive features is available in the datasets. Therefore, knowledge of the sensitive issues present in real life is improved. This allows representative problems to be reduced [5].

Data-based decisions must be *documented* with original priorities and the necessary annotations [6]. This allows answering which data have been used consistently with inclusion patterns [100].

Opening access to train the data used to build the model leads to better transparency and trustworthiness. In addition, this allows third parties to detect possible biases [9].

3. Model

This topic gives suggestions for the training and evaluation phases. Furthermore, during the deployment it is essential to understand and explain the results. The model perpetuates or creates bias independently of the origin of this bias. Fig. 4 shows that under the fair governance category of bias mitigation methods, the "model" and "data" sections share the "documented" and "open access" recommendations explained above.

Models open to the community would allow their testing to detect new biases and demonstrate their efficacy [9]. Also, documentation of assumptions on parameters or metrics should be transparent and available [6].

Furthermore, applying *Explainable Artificial Intelligence* (XAI) allows understanding the model results and the importance of each feature in the predictions [4]–[6].

VIII. Discussion

This section explains the results obtained in this systematic review and the answers to the research questions revised. Section III has detailed the search criteria for this systematic review. These criteria pursue robust and unbiased results. Among others, only peer-reviewed

research works have been considered, which is the first of the six inclusion criteria discussed. Although peer review is not a guarantee of a good level of confidence in the published results, it is seen as the foremost process for research validation [101].

A. RQ1: What Bias Affects Fairness?

Regarding RQ1, results were quite broad because the term *biases* encompasses different concepts within AI, including poorly balanced datasets which can lead to performance issues. However, this paper specifically focuses on biases that are related to injustice. Many studies on this question analyze biases for specific use cases, such as classification or recommendation. Other publications listed biases that could appear in the training cycle (from data acquisition to algorithm deployment).

As a result of this question, Fig. 4 groups biases in the following steps of an algorithm development process: Human Bias (Data Generation), Data Bias, Learning Bias, and Deployment Bias. In all the groups, the focus is placed on the human factor, which is responsible for the decisions made.

The critical step detected in this point is Human Bias. As discussed earlier, this is where historical data for training is generated in a way that can introduce biases. The bias in the following steps could be reduced or eliminated by reducing the cognitive bias presented in the dataset.

B. RQ2: What Are the Metrics to Measure Fairness?

The topic of metrics has only appeared in a few of the retrieved publications, being the most widespread metrics detailed in Table II. Another clear result is that the most used metrics are the ones that focus on groups. Moreover, this paper also reviews metrics that have recently emerged (generally customized for a specific case). Tools to audit algorithms have also been added to complete the answer to this research question.

As a result of this review, an interesting gap in the literature has been found. None of the revised metrics helps to detect variables or values that can cause unfairness. Therefore, it is necessary to know which groups are privileged and unprivileged to measure the differences between selected groups.

Applying these metrics to the entire dataset requires a high computational cost. Moreover, these metrics cannot be used with all variables because some of them, without being a discriminatory feature, can correctly segregate the samples.

C. RQ3: How to Mitigate Bias?

In the last question, the aim was to find techniques that would help mitigate biases. This is the question that most of the papers covered have focused on.

As a result of the review presented in this paper, mitigation techniques have been grouped into a taxonomy in Fig. 5. This taxonomy contains four main categories: Pre-training, Training, Post-training, and Fair Governance.

Most of the retrieved works focus on correcting data (pre-training) or improving learning (training). The pre-training algorithms include: Resampling, Fair Representation, and Re-weighting. The training algorithms consider: Regularization, Adversarial Training, and Emerging Approaches. Training algorithms are where the most varied solutions are developed. At this category, the main goal is to maintain accuracy while improving fairness. On the other hand, focusing on the output (post-training) is a less popular approach.

Some publications address their research to detecting transversal actions to mitigate unfairness. These actions attempt to improve the development ecosystem. The Fair Governance category contains actions that reduce bias when applied to: Teams (Diversity, Cross-disciplinary, and Training), Data (Collection, Documented, and Open Access), or Models (Documented, Open Access, and apply XAI techniques). This category can reduce bias in the Data Bias, Learning Bias, and Deployment Bias. Thus, the Fair Governance category is essential to have an impact on fairness.

IX. Conclusions and Future Work

This paper has established a systematic review to answer three research questions.

The first question focuses on understanding the origin of unfair results in AI models. As a result of the review, a comprehensive analysis of the type of biases that affect fairness was produced.

The second question explored equity metrics to detect discrimination in data or models. Studies show that quantifying this issue is very complex with the current state of the art. The review identifies two main factors needing improvement: obtaining generalized measures and automatically detecting sensitive features. The paper contributes to a compilation of the most popular and novel metrics to measure fairness.

The last question was aimed at obtaining information on how to mitigate the effects of bias in AI models. According to the extensive specialized literature reviewed, this mitigation is still a complex and imprecise task. More importantly, reducing bias can change learning and obtain undesirable results. Among others, the results could not represent reality when the algorithm avoids historical discrimination. A taxonomy that aggregates the different mitigation techniques depending on where they are applied is this paper's third and main contribution.

Future work should address the development of a fairness-by-design standard for developing AI models. In addition, the detection of feature bias should be automated at least for sensitive variables or their derivatives. Finally, an indicator of responsible AI development is needed beyond the use of performance metrics.

Acknowledgment

This work has been granted by the "EICACS (European Initiative for Collaborative Air Combat Standardisation)" project of the Horizon Europe programme of the European Commission, under grant agreement No. 101103669. The work of Paulo Novais is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project DSAIPA/AI/0099/2019.

References

[1] R. Kennedy, "The ethical implications of lawtech," in *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society*, 2021, pp. 198–207, Springer International Publishing.

[2] S. Camaréna, "Engaging with artificial intelligence (ai) with a bottom-up approach for the purpose of sustainability: Victorian farmers market association, melbourne australia," *Sustainability*, vol. 13, no. 16, 2021, doi: 10.3390/su13169314.

[3] S. Strauß, "Deep automation bias: How to tackle a wicked problem of ai?," *Big Data and Cognitive Computing*, vol. 5, no. 2, 2021, doi: 10.3390/bdcc5020018.

[4] A. Nadeem, O. Marjanovic, B. Abedin, "Gender bias in ai: Implications for managerial practices," in *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society*, 2021, pp. 259–270, Springer International Publishing.

[5] S. Parsheera, "A gendered perspective on artificial intelligence," in *2018 International Telecommunication Union Kaleidoscope: Machine Learning*

for a 5G Future, Santa Fe, Argentina, November 26-28, 2018, 2018, pp. 1–7, IEEE.

[6] T. K. Gilbert, Y. Mintz, "Epistemic therapy for bias in automated decision-making," in *Proceedings of the 2019 Conference on AI, Ethics, and Society*, New York, NY, USA, 2019, p. 61–67, Association for Computing Machinery.

[7] D. A. da Silva, H. D. B. Louro, G. S. Goncalves, J. C. Marques, L. A. V. Dias, A. M. da Cunha, P. M. Tasinaffo, "Could a conversational ai identify offensive language?," *Information*, vol. 12, no. 10, 2021, doi: 10.3390/info12100418.

[8] C. Zhao, C. Li, J. Li, F. Chen, "Fair meta-learning for few-shot classification," in *2020 IEEE International Conference on Knowledge Graph, Online, August 9-11, 2020*, 2020, pp. 275–282, IEEE.

[9] R. S. Baker, A. Hawn, "Algorithmic bias in education," *International Journal of Artificial Intelligence in Education*, pp. 1052–1092, 12 2021, doi: 10.1007/s40593-021-00285-9.

[10] R. R. Fletcher, A. Nakeshimana, O. Olubeko, "Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health," *Frontiers in Artificial Intelligence*, vol. 3, 4 2021, doi: 10.3389/frai.2020.561802.

[11] M. Loi, A. Ferrario, E. Viganò, "Transparency as design publicity: explaining and justifying inscrutable algorithms," *Ethics and Information Technology*, vol. 23, pp. 253–263, 9 2021, doi: 10.1007/s10676-020-09564-w.

[12] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, S. Staab, "Bias in data-driven artificial intelligence systems—an introductory survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, 5 2020, doi: 10.1002/widm.1356.

[13] D. Pessach, E. Shmueli, "Improving fairness of artificial intelligence algorithms in privileged- group selection bias data settings," *Expert Systems with Applications*, vol. 185, 12 2021, doi: 10.1016/j.eswa.2021.115667.

[14] D. Pessach, E. Shmueli, "A review on fairness in machine learning," *Association for Computing Machinery: Computing Surveys*, vol. 55, feb 2022, doi: 10.1145/3494672.

[15] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, "A survey on bias and fairness in machine learning," *Association for Computing Machinery: Computing Surveys*, vol. 54, no. 6, pp. 115:1–115:35, 2021, doi: 10.1145/3457607.

[16] S. Khenissi, B. Mariem, O. Nasraoui, "Theoretical modeling of the iterative properties of user discovery in a collaborative filtering recommender system," in *Proceedings of the 14th Conference on Recommender Systems*, New York, NY, USA, 2020, p. 348–357, Association for Computing Machinery.

[17] A. Peña, I. Serna, A. Morales, J. Fiérrez, "Faircvtest demo: Understanding bias in multimodal learning with a testbed in fair automatic recruitment," in *International Conference on Multimodal Interaction, Virtual Event, The Netherlands, October 25-29, 2020*, 2020, pp. 760–761, Association for Computing Machinery.

[18] J. L. Davis, A. Williams, M. W. Yang, "Algorithmic reparation," *Big Data and Society*, vol. 8, 2021, doi: 10.1177/20539517211044808.

[19] A. M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, C. Mooney, "Current challenges and future opportunities for xai in machine learning- based clinical decision support systems: A systematic review," *Applied Sciences (Switzerland)*, vol. 11, 6 2021, doi: 10.3390/app11115088.

[20] K. Sokol, "Fairness, accountability and transparency in artificial intelligence: A case study of logical predictive models," in *Proceedings of the 2019 Conference on AI, Ethics, and Society*, New York, NY, USA, 2019, p. 541–542, Association for Computing Machinery.

[21] K. Gade, S. C. Geyik, K. Kenthapadi, V. Mithal, A. Taly, "Explainable AI in industry: practical challenges and lessons learned: implications tutorial," in *Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, 2020, p. 699, Association for Computing Machinery.

[22] A. Stevens, P. Deruyck, Z. V. Veldhoven, J. Vanthienen, "Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva," in *2020 IEEE Symposium Series on Computational Intelligence Canberra, Australia, December 1-4, 2020*, 2020, pp. 1241–1248, IEEE.

[23] R. Bellamy, K. Dey, M. Hind, S. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. Varshney, Y. Zhang, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. PP, 09 2019, doi: 10.1147/JRD.2019.2942287.

[24] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. A. Specter, L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *5th IEEE International Conference on Data Science and Advanced Analytics, Turin, Italy, October 1-3, 2018*, 2018, pp. 80–89, IEEE.

[25] E. Mutlu, O. O. Garibay, "A quantum leap for fairness: Quantum bayesian approach for fair decision making," in *Human-Computer Interaction (HCI) International 2021 - Late Breaking Papers: Multimodality, eXtended Reality, and Artificial Intelligence*, 2021, pp. 489–499, Springer International Publishing.

[26] J. Stoyanovich, B. Howe, S. Abiteboul, G. Miklau, A. Sahuguet, G. Weikum, "Fides: Towards a platform for responsible data science," in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, New York, NY, USA, 2017, Association for Computing Machinery.

[27] C. Addis, M. Kutar, "AI management an exploratory survey of the influence of GDPR and FAT principles," in *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, Leicester, United Kingdom, August 19-23, 2019*, 2019, pp. 342–347, IEEE.

[28] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020, doi: https://doi.org/10.1016/j.inffus.2019.12.012.

[29] V. Bhargava, M. Couceiro, A. Napoli, "Limeout: An ensemble approach to improve process fairness," in *European Conference on Machine Learning and Knowledge Discovery in Databases 2020 Workshops*, 2020, pp. 475–491, Springer International Publishing.

[30] S. Wachter, B. Mittelstadt, C. Russell, "Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai," *Computer Law and Security Review*, vol. 41, 7 2021, doi: 10.1016/j.clsr.2021.105567.

[31] S. Abbasi-Sureshjani, R. Raumanns, B. E. J. Michels, Schouten, V. Cheplygina, "Risk of training diagnostic algorithms on data with demographic bias," in *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, 2020, pp. 183–192, Springer International Publishing.

[32] M. DeCamp, C. Lindvall, "Latent bias and the implementation of artificial intelligence in medicine," *Journal of the American Medical Informatics Association*, vol. 27, pp. 2020–2023, 12 2020, doi: 10.1093/jamia/ocaa094.

[33] Y. Zheng, S. Wang, J. Zhao, "Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models," *Transportation Research Part C: Emerging Technologies*, vol. 132, p. 103410, 2021, doi: https://doi.org/10.1016/j.trc.2021.103410.

[34] V. V. Vesselinov, B. S. Alexandrov, D. O'Malley, "Nonnegative tensor factorization for contaminant source identification," *Journal of Contaminant Hydrology*, vol. 220, pp. 66–97, 2019, doi: https://doi.org/10.1016/j.jconhyd.2018.11.010.

[35] O. J. Akintande, "Algorithm fairness through data inclusion, participation, and reciprocity," in *Database Systems for Advanced Applications*, 2021, pp. 633–637, Springer International Publishing.

[36] S. Park, H. Ko, "Machine learning and law and economics: A preliminary overview," *Asian Journal of Law and Economics*, vol. 11, 8 2020, doi: 10.1515/ajle-2020-0034.

[37] F. Marcinkowski, K. Kieslich, C. Starke, M. Lünich, "Implications of AI (un-)fairness in higher education admissions: the effects of perceived AI (un-)fairness on exit, voice and organizational reputation," in *Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, 2020, pp. 122– 130, Association for Computing Machinery.

[38] D. Solans, F. Fabbri, C. Calsamiglia, C. Castillo, F. Bonchi, "Comparing equity and effectiveness of different algorithms in an application for the room rental market," in *Proceedings of the 2021 Conference on AI, Ethics, and Society*, New York, NY, USA, 2021, p. 978–988, Association for Computing Machinery.

[39] S. Hajian, F. Bonchi, C. Castillo, "Algorithmic bias: From discrimination discovery to fairness- aware data mining," in *Proceedings of the 22nd*

International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 2125–2126, Association for Computing Machinery.

[41] C. M. Madla, F. K. H. Gavins, H. A. Merchant, M. Orlu, S. Murdan, A. W. Basit, "Let's talk about sex: Differences in drug therapy in males and females," *Advanced Drug Delivery Reviews*, vol. 175, p. 113804, 2021, doi: https://doi.org/10.1016/j.addr.2021.05.014.

[42] G. Currie, K. E. Hawk, "Ethical and legal challenges of artificial intelligence in nuclear medicine," *Seminars in Nuclear Medicine*, vol. 51, pp. 120–125, 2021, doi: https://doi.org/10.1053/j.semnuclmed.2020.08.001.

[43] G. Starke, E. D. Clercq, B. S. Elger, "Towards a pragmatist dealing with algorithmic bias in medical machine learning," *Medicine, Health Care and Philosophy*, vol. 24, pp. 341–349, 9 2021, doi: 10.1007/s11019-021-10008-5.

[44] A. M. Fejerskov, "Algorithmic bias and the (false) promise of numbers," *Global Policy*, vol. 12, pp. 101– 103, 7 2021, doi: 10.1111/1758-5899.12915.

[45] C. Panigutti, A. Perotti, A. Panisson, P. Bajardi, D. Pedreschi, "Fairlens: Auditing black-box clinical decision support systems," *Information Processing & Management*, vol. 58, p. 102657, 2021, doi: https://doi.org/10.1016/j.ipm.2021.102657.

[46] S. Kino, Y.-T. Hsu, K. Shiba, Y.-S. Chien, C. Mita, I. Kawachi, A. Daoud, "A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects," *SSM-Population Health*, vol. 15, p. 100836, 2021, doi: https://doi.org/10.1016/j.ssmph.2021.100836.

[47] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, A. Tzovara, "Addressing bias in big data and ai for health care: A call for open science," *Patterns*, vol. 2, p. 100347, 2021, doi: https://doi.org/10.1016/j.patter.2021.100347.

[48] L. Xu, "The dilemma and countermeasures of ai in educational application," *Pervasive Health: Pervasive Computing Technologies for Healthcare*, pp. 289–294, 12 2020, doi: 10.1145/3445815.3445863.

[49] W. Holmes, K. Porayska-Pomsta, K. Holstein, E. Sutherland, T. Baker, S. B. Shum, O. C. Santos, M. T. Rodrigo, M. Cukurova, I. I. Bittencourt, K. R. Koedinger, "Ethics of ai in education: Towards a community-wide framework," *International Journal of Artificial Intelligence in Education*, 2021, doi: 10.1007/s40593-021-00239-1.

[50] A. Peña, I. Serna, A. Morales, J. Fierrez, "Faircvtest demo: Understanding bias in multimodal learning with a testbed in fair automatic recruitment," *2020 International Conference on Multimodal Interaction*, pp. 760–761, 10 2020, doi: 10.1145/3382507.3421165.

[51] A. Ortega, J. Fierrez, A. Morales, Z. Wang, M. de la Cruz, C. L. Alonso, T. Ribeiro, "Symbolic ai for xai: Evaluating lfit inductive programming for explaining biases in machine learning," *Computers*, vol. 10, 11 2021, doi: 10.3390/computers10110154.

[52] S. K. Misra, S. Das, S. Gupta, S. K. Sharma, "Public policy and regulatory challenges of artificial intelligence (ai)," *Advances in Information and Communication Technology*, vol. 617, pp. 100–111, 2020, doi: 10.1007/978-3-030-64849-7_10.

[53] S. Pundhir, U. Ghose, V. Kumari, "Legitann: Neural network model with unbiased robustness," in *Proceedings of International Conference on Communication and Artificial Intelligence*, 2021, pp. 385– 397, Springer Singapore.

[54] S. Sharma, A. H. Gee, D. Paydarfar, J. Ghosh, "Fair- n: Fair and robust neural networks for structured data," in *Proceedings of the 2021 Conference on AI, Ethics, and Society*, New York, NY, USA, 2021, p. 946–955, Association for Computing Machinery.

[55] J. Kang, H. Tong, "Fair graph mining," in *The 30th International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, 2021, pp. 4849–4852, Association for Computing Machinery.

[56] J. Bobadilla, R. Lara-Cabrera, Á. González- Prieto, F. Ortega, "DeepFair: Deep learning for improving fairness in recommender systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, p. 86, 2021, doi: 10.9781/ijimai.2020.11.001.

[57] N. Martinez-Martin, Z. Luo, A. Kaushal, E. Adeli, A. Haque, S. S. Kelly, S. Wieten, M. K. Cho, D. Magnus, L. Fei-Fei, K. Schulman, A. Milstein, "Ethical issues in using ambient intelligence in health-care settings," *The Lancet Digital Health*, vol. 3, pp. e115–e123, 2 2021, doi: 10.1016/S2589-7500(20)30275-2.

[58] S. K. Kane, A. Guo, M. R. Morris, "Sense and accessibility: Understanding people with physical disabilities' experiences with sensing systems," in *The 22nd International Conference on Computers and Accessibility, Virtual Event, Greece, October 26-28, 2020*, 2020, pp. 42:1–42:14, Association for Computing Machinery.

[59] A. Paviglianiti, E. Pasero, "VITAL-ECG: a de-bias algorithm embedded in a gender-immune device," in *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT, Roma, Italy, June 3-5, 2020*, 2020, pp. 314–318, IEEE.

[60] R. Mark, "Ethics of public use of ai and big data: The case of amsterdam's crowdedness project," *The ORBIT Journal*, vol. 2, no. 2, pp. 1–33, 2019, doi: https://doi.org/10.29297/orbit.v2i1.101.

[61] C. E. Kontokosta, B. Hong, "Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions," *Sustainable Cities and Society*, vol. 64, p. 102503, 2021, doi: https://doi.org/10.1016/j.scs.2020.102503.

[62] V. D. Badal, C. Nebeker, K. Shinkawa, Y. Yamada, K. E. Rentscher, H.-C. Kim, E. E. Lee, "Do words matter? detecting social isolation and loneliness in older adults using natural language processing," *Frontiers in Psychiatry*, vol. 12, 11 2021, doi: 10.3389/fpsyt.2021.728732.

[63] B. Richardson, D. Prioleau, K. Alikhademi, J. E. Gilbert, "Public accountability: Understanding sentiments towards artificial intelligence across dispositional identities," in *IEEE International Symposium on Technology and Society, Tempe, AZ, USA, November 12-15, 2020*, 2020, pp. 489–496, IEEE.

[64] D. Muralidhar, "Examining religion bias in AI text generators," in *Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, 2021, pp. 273– 274, Association for Computing Machinery.

[65] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, pp. 12592–12594, 6 2020, doi: 10.1073/pnas.1919012117.

[66] E. Puyol-Antón, B. Ruijsink, S. K. Piechnik, S. Neubauer, S. E. Petersen, R. Razavi, A. P. King, "Fairness in cardiac mr image analysis: An investigation of bias due to data imbalance in deep learning based segmentation," *Medical Image Computing and Computer Assisted Intervention*, vol. 12903 LNCS, pp. 413–423, 2021, doi: 10.1007/978- 3-030-87199-4_39.

[67] A. Rajabi, O. O. Garibay, "Towards fairness in ai: Addressing bias in data using gans," in *Human- Computer Interaction (HCI) International 2021 - Late Breaking Papers: Multimodality, eXtended Reality, and Artificial Intelligence*, Cham, 2021, pp. 509–518, Springer International Publishing.

[68] Y. Zhang, J. Sang, "Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing," in *Proceedings of the 28th International Conference on Multimedia*, New York, NY, USA, 2020, p. 4346–4354, Association for Computing Machinery.

[69] A. K. Singh, A. Blanco-Justicia, J. Domingo-Ferrer, D. Sánchez, D. Rebollo-Monedero, "Fair detection of poisoning attacks in federated learning," in *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence*, 2020, pp. 224–229, IEEE.

[70] C. Fan, M. Esparza, J. Dargin, F. Wu, B. Oztekin, A. Mostafavi, "Spatial biases in crowdsourced data: Social media content attention concentrates on populous areas in disasters," *Computers, Environment and Urban Systems*, vol. 83, p. 101514, 2020, doi: https://doi.org/10.1016/j.compenvurbsys.2020.101514.

[71] R. Chiong, Z. Fan, Z. Hu, F. Chiong, "Using an improved relative error support vector machine for body fat prediction," *Computer Methods and Programs in Biomedicine*, vol. 198, p. 105749, 2021, doi: https://doi.org/10.1016/j.cmpb.2020.105749.

[72] W. Lee, H. Ko, J. Byun, T. Yoon, J. Lee, "Fair clustering with fair correspondence distribution," *Information Sciences*, vol. 581, pp. 155–178, 2021, doi: https://doi.org/10.1016/j.ins.2021.09.010.

[73] W. Zhang, A. Bifet, X. Zhang, J. C. Weiss, W. Nejdl, "Farf: A fair and adaptive random forests classifier," in *Advances in Knowledge Discovery and Data Mining*, 2021, pp. 245–256, Springer International Publishing.

[74] C. G. Harris, "Mitigating cognitive biases in machine learning algorithms for decision making," in *Companion Proceedings of the Web Conference 2020*, New York, NY, USA, 2020, p. 775–781, Association for Computing Machinery.

[75] M. A. U. Alam, "Ai-fairness towards activity recognition of older adults," *Pervasive Health: Pervasive Computing Technologies for Healthcare*, pp. 108–117, 12 2020, doi: 10.1145/3448891.3448943.

[76] Y. Zhang, A. Ramesh, "Learning fairness-aware relational structures," *Frontiers in Artificial Intelligence and Applications*, vol. 325, pp. 2543–2550, 8 2020, doi: 10.3233/FAIA200389.

[77] S. Ahmed, S. A. Athyaab, S. A. Muqtadeer, "Attenuation of human bias in artificial intelligence: An exploratory approach," in *2021 6th International Conference on Inventive Computation Technologies*, 2021, pp. 557–563, IEEE.

[78] Y. Hou, H. Hong, Z. Sun, D. Xu, Z. Zeng, "The control method of twin delayed deep deterministic policy gradient with rebirth mechanism to multi-dof manipulator," *Electronics (Switzerland)*, vol. 10, 4 2021, doi: 10.3390/electronics10070870.

[79] K. Xivuri, H. Twinomurinzi, "A systematic review of fairness in artificial intelligence algorithms," in *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society*, Cham, 2021, pp. 271–284, Springer International Publishing.

[80] S. Sharma, Y. Zhang, J. M. Ríos Aliaga, D. Bouneffouf, V. Muthusamy, K. R. Varshney, "Data augmentation for discrimination prevention and bias disambiguation," in *Proceedings of the Conference on AI, Ethics, and Society*, New York, NY, USA, 2020, p. 358–364, Association for Computing Machinery.

[81] A. Pandey, A. Caliskan, "Disparate impact of artificial intelligence bias in ridehailing economy's price discrimination algorithms," in *Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, 2021, pp. 822–833, Association for Computing Machinery.

[82] S. Udeshi, P. Arora, S. Chattopadhyay, "Automated directed fairness testing," in *Proceedings of the 33rd International Conference on Automated Software Engineering*, New York, NY, USA, 2018, p. 98–108, Association for Computing Machinery.

[83] N. V. Berkel, J. Goncalves, D. Hettiachchi, S. Wijenayake, R. M. Kelly, V. Kostakos, "Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study," *Proceedings of the Association for Computing Machinery on Human-Computer Interaction*, vol. 3, 11 2019, doi: 10.1145/3359130.

[84] W. Fleisher, "What's fair about individual fairness?," in *Proceedings of the 2021 Conference on AI, Ethics, and Society*, New York, NY, USA, 2021, p. 480–490, Association for Computing Machinery.

[85] O. Aka, K. Burke, A. Bauerle, C. Greer, M. Mitchell, "Measuring model biases in the absence of ground truth," in *Proceedings of the 2021 Conference on AI, Ethics, and Society*, New York, NY, USA, 2021, p. 327–335, Association for Computing Machinery.

[86] S. Verma, J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*, New York, NY, USA, 2018, p. 1–7, Association for Computing Machinery.

[87] D. Fan, Y. Wu, X. Li, "On the fairness of swarm learning in skin lesion classification," *Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning"*, p. 120–129, 2021, doi: 10.1007/978-3-030-90874-4_12.

[88] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, M. B. Zafar, "A unified approach to quantifying algorithmic unfairness," *Proceedings of the 24th International Conference on Knowledge Discovery and Data Mining*, jul 2018, doi: 10.1145/3219819.3220046.

[89] L. Liang, D. E. Acuna, "Artificial mental phenomena: Psychophysics as a framework to detect perception biases in ai models," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 403–412, 1 2020, doi: 10.1145/3351095.3375623.

[90] P. Saleiro, B. Kuester, A. Stevens, A. Anisfeld, L. Hinkson, J. London, R. Ghani, "Aequitas: A bias and fairness audit toolkit," *CoRR*, vol. abs/1811.05577, 2018.

[91] H. J. P. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, M. Madaio, "Fairlearn: Assessing and improving fairness of AI systems," *Computing Research Repository*, vol. abs/2303.16626, 2023, doi: 10.48550/arXiv.2303.16626.

[92] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, "The what-if tool: Interactive probing of machine learning models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 56–65, 2020, doi: 10.1109/TVCG.2019.2934619.

[93] C. Wilson, A. Ghosh, S. Jiang, A. Mislove, L. Baker, J. Szary, K. Trindel, F. Polli, "Building and auditing fair algorithms: A case study in candidate screening," in *Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 2021, p. 666–677, Association for Computing Machinery.

[94] A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz- Marí, L. Gómez-Chova, G. Camps-Valls, "Fair kernel learning," in *Machine Learning and Knowledge Discovery in Databases*, 2017, pp. 339–355, Springer International Publishing.

[95] P. Smith, K. Ricanek, "Mitigating algorithmic bias: Evolving an augmentation policy that is non-biasing," in *2020 IEEE Winter Applications of Computer Vision Workshops*, 2020, pp. 90–97, IEEE.

[96] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2015, p. 259–268, Association for Computing Machinery.

[97] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 3, 06 2014, doi: 10.1145/3422622.

[98] B. H. Zhang, B. Lemoine, M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 Conference on AI, Ethics, and Society*, New York, NY, USA, 2018, p. 335–340, Association for Computing Machinery.

[99] C. Weber, "Engineering bias in ai," *IEEE Pulse*, vol. 10, pp. 15–17, 1 2019, doi: 10.1109/MPULS.2018.2885857.

[100] N. T. Lee, "Detecting racial bias in algorithms and machine learning," *Journal of Information, Communication and Ethics in Society*, vol. 16, pp. 252– 260, 11 2018, doi: 10.1108/JICES-06-2018-0056.

[101] N. McDonald, S. Pan, "Intersectional ai: A study of how information science students think about ethics and their impact," *Proceedings of the Association for Computing Machinery on Human-Computer Interaction*, vol. 4, 10 2020, doi: 10.1145/3415218.

[102] P. K. Bharti, T. Ghosal, M. Agrawal, A. Ekbal, "How confident was your reviewer? estimating reviewer confidence from peer review texts," in *Document Analysis Systems - 15th International Workshop, 2022, La Rochelle, France, May 22-25, 2022, Proceedings*, vol. 13237 of *Lecture Notes in Computer Science*, 2022, pp. 126–139, Springer.

Rubén González Sendino

Rubén González completed his Master's degree in Artificial Intelligence from the Universidad Politécnica de Madrid in 2017. He is currently pursuing a PhD at the same institution. Over the years, he has gained extensive experience in innovation departments and has actively collaborated on European Horizon 2020 projects. He possesses expertise in deep neural networks, natural language processing, causal algorithms, and computer vision. Beyond the technical intricacies, he is deeply committed to explicability and responsible artificial intelligence.

Emilio Serrano

Emilio Serrano received the M.Sc. degree in computer science (2006) and the Ph.D. degree, with European mention and Extraordinary Ph.D. Award in artificial intelligence (2011), from the University of Murcia, Spain. He has also been a Visiting Researcher with The University of Edinburgh, the University of Oxford, and the National Institute of Informatics in Tokyo. He is currently an Associate Professor with the Department of Artificial Intelligence, Universidad Politécnica de Madrid (UPM). His main research line is the Social and Explainable Artificial Intelligence for Smart Cities. His scientific production includes more than 80 publications. He lectures deep learning for natural language processing and social network analysis among other courses. He has been principal investigator in four educational innovation projects in data science, participated in several European and National funding programs (6 European projects), and supervised two Ph.D theses.

Javier Bajo

Dr. Javier Bajo, full professor at the Department of Artificial Intelligence, Computer Science School at Universidad Politécnica de Madrid (UPM), holds (since 03/05/2019) the position of Director of the UPM AI.nnovation Space Research Center in Artificial Intelligence. He was Director of the Department of Artificial Intelligence (20/05/2016-19/10/2017) at UPM, Secretary of the PhD in Artificial Intelligence at UPM (23/06/2016-19/10/2017) and Coordinator of the Research Master in Artificial Intelligence at UPM (18/02/2013 - 20/05/2016). He also holds the position of Director of the Data Center at the Pontifical University of Salamanca (13-10/2010 - 08-11-2012), with 21 employees. His main lines of research are Social Computing and Artificial and Hybrid Societies; Intelligent Agents and Multiagent Systems, Ambient Intelligence, Machine Learning. He has supervised 11 Ph.D thesis, participated in more than 50 research projects (in most of them as principal investigator) and published more than 300 articles in recognized journals (81 JCR papers) and conferences. His h-index is 39. He is founder of the PAAMS series of conferences and is an IEEE, ACM and ISIF member.

Paulo Novais

Dr. Paulo Novais is a Full Professor of Computer Science at the Department of Informatics, the School of Engineering, the University of Minho (Portugal) and a researcher at the ALGORITMI Centre in which he is the leader of the research group ISLab - Synthetic Intelligence lab. He is the director of the PhD Program in Informatics and co-founder and Deputy Director of the Master in Law and Informatics at the University of Minho. His main research objective is to make systems a little more smart, reliable and sensitive to human presence and interaction. He is the coordinator of the Portuguese Intelligent Systems Associate Laboratory (LASI). He has supervised 20 PhD thesis, participated in several research projects sponsored by Portuguese and European public and private Institutions, and published more than 350 articles in recognized journals and conferences.