

## Article

# Dealing with Gender Bias Issues in Data-Algorithmic Processes: A Social-Statistical Perspective

Juliana Castaneda <sup>1</sup>, Assumpta Jover <sup>2</sup>, Laura Calvet <sup>1</sup>, Sergi Yanes <sup>3</sup>, Angel A. Juan <sup>4,\*</sup>  
and Milagros Sainz <sup>3</sup>

- <sup>1</sup> Computer Science Department, Universitat Oberta de Catalunya, 08018 Barcelona, Spain; jcastanedaji@uoc.edu (J.C.); lcalvetl@uoc.edu (L.C.)  
<sup>2</sup> Department of Sociology and Social Anthropology, Universitat de València, 46022 Valencia, Spain; assumpta41185@gmail.com  
<sup>3</sup> GenTIC, Universitat Oberta de Catalunya, 08018 Barcelona, Spain; syanes@uoc.edu (S.Y.); msainzi@uoc.edu (M.S.)  
<sup>4</sup> Department of Applied Statistics and Operations Research, Universitat Politècnica de València, 03801 Alcoy, Spain  
\* Correspondence: ajuanp@eio.upv.es

**Abstract:** Are algorithms sexist? This is a question that has been frequently appearing in the mass media, and the debate has typically been far from a scientific analysis. This paper aims at answering the question using a hybrid social and technical perspective. First a technical-oriented definition of the algorithm concept is provided, together with a more social-oriented interpretation. Secondly, several related works have been reviewed in order to clarify the state of the art in this matter, as well as to highlight the different perspectives under which the topic has been analyzed. Thirdly, we describe an illustrative numerical example possible discrimination in the banking sector due to data bias, and propose a simple but effective methodology to address it. Finally, a series of recommendations are provided with the goal of minimizing gender bias while designing and using data-algorithmic processes to support decision making in different environments.

**Keywords:** algorithmic bias; gender bias; data science; artificial intelligence; decision making



**Citation:** Castaneda, J.; Jover, A.; Calvet, L.; Yanes, S.; Juan, A.A.; Sainz, M. Dealing with Gender Bias Issues in Data-Algorithmic Processes: A Social-Statistical Perspective. *Algorithms* **2022**, *15*, 303. <https://doi.org/10.3390/a15090303>

Academic Editor: Laurent Risser

Received: 20 July 2022

Accepted: 23 August 2022

Published: 27 August 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

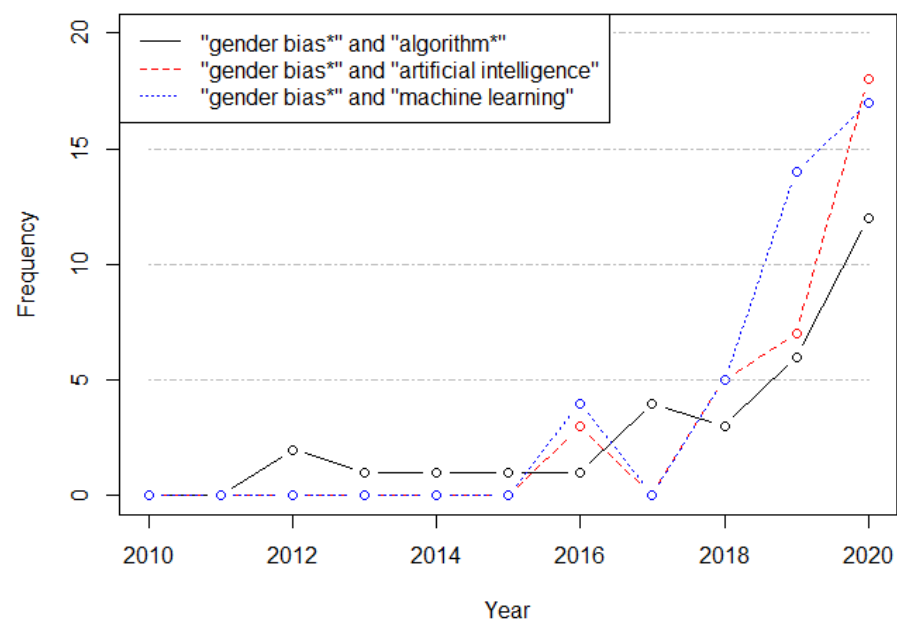


**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

As pointed out by some of the pioneers of machine learning, one of the main problems inherent to the logic of computer systems is their inability to dissociate the information they handle from the context in which they are created [1]. This generates the need for overcoming the dichotomy between the social and technical perspectives, which makes it difficult to address social issues such as the existence of bias in the development of data-algorithmic processes [2]. The conception of data-algorithmic processes as complex socio-technical systems, and not just as autonomous technical methods, contributes to add a social perspective to the debate [3].

Algorithmic-related bias refers to systematic and repeatable errors in a mathematical or computer system that lead to ‘unfair’ outputs, privileging one or more groups over others. Gender bias in data-algorithmic processes is a particular type of bias where one of the genders is discriminated. Some authors have associated the presence of gender bias with the under-representation of women in the design and production of artificial intelligence (AI) products and services [4–6]. The number of AI applications has been increasingly growing during the last decades, which cover a wide range of fields: from natural language generation to face recognition. At the same time, the concern regarding AI/machine learning (ML) and gender bias has also increased significantly (Figure 1).



**Figure 1.** Scopus-indexed articles for different gender-related terms.

For example, the widespread use of popular word embedding algorithms exhibiting stereotypical biases—including gender bias—in ML systems can thus amplify stereotypes in several contexts. For this reason, some methods have been developed to mitigate this problem [7]. Examples of methods for evaluating bias in text are the word embedding association test (WEAT) and the word embedding factual association test (WEFAT). These have implications not only for AI/ML, but also for other fields, such as Psychology, Sociology, and Human Ethics, since these methods raise the possibility that mere exposure to everyday language can account for the biases replicated by ML techniques [8]. This reinforces the importance of developing an interdisciplinary analysis on the presence of biases in data-algorithmic processes, and how these biases might guide decisions that do not represent the diversity and complexity of modern societies. Likewise, recent technological advances and the extensive use of algorithms raise ethical problems, particularly those prompted by algorithmic decision-making [9]. The potential biases in algorithm decision-making have encouraged several research on the effects of AI in the development of the different UNESCO's sustainable goals. Authors such as Tsamados et al. [10] and Taddeo and Floridi [11] formulate affirmation such as “algorithms are not ethically neutral”.

In this context, our paper analyzes whether data-algorithmic processes are sexist –by first providing definitions of the algorithm concept, considering both social and technical perspectives–, and an introduction to the concept of data-algorithmic bias. In addition, examples of gender bias in AI applications, based on different fields such as natural language processing (NLP), speech recognition, decision management, and face recognition are explained. Afterwards, a review of works presenting methods to detect and mitigate gender bias in AI applications is offered, as well as a list of private initiatives and recommendations from international organizations. Likewise, a numerical example regarding discrimination due to data bias, a simple yet effective methodology to solve the issue, and a set of general recommendations for any AI practitioner interested in detecting and mitigating gender bias are also discussed.

The rest of the paper is structured as follows: Section 2 defines the concept of algorithm, both from a technical and a social scientist's point of view, and reviews an algorithm classification by function, implementation, and design paradigm. Next, Section 3 discusses

different definitions and classifications of algorithmic bias. Section 4 describes examples of gender bias in data-algorithmic processes for a wide range of fields, while Section 5 describes a few datasets with gender bias, which are diverse, popular, and freely-accessible. Section 6 presents private initiatives and recommendations from international organizations to address gender bias. An illustrative numerical example of discrimination due to data bias as well as a methodology to address this bias are presented in Section 7. Finally, Section 8 provides a series of recommendations to prevent, identify, and mitigate gender bias, while Section 9 draws a few conclusions from this work.

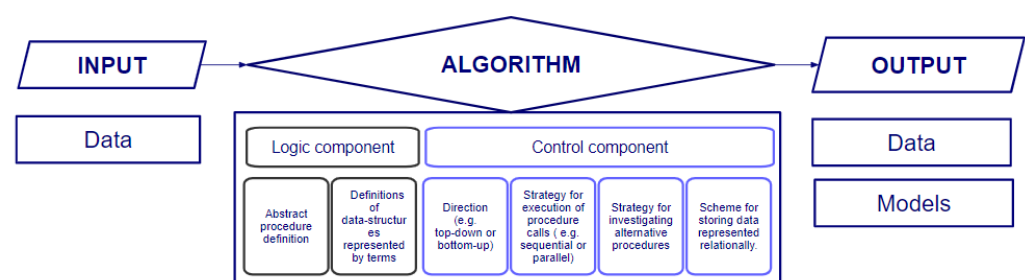
## 2. The Algorithm Concept

This section reviews the concept of algorithm, both from a technical perspective as well as from a more social one.

### 2.1. Algorithm Concept in Science and Engineering

Algorithm is a noun with several definitions, which vary according to the context. As defined by the Oxford dictionary, it is “a set of rules that must be followed when solving a particular problem”. According to the Cambridge dictionary, it is “a set of mathematical instructions that must be followed in a fixed order, and that, especially if given to a computer, will help to calculate an answer to a mathematical problem”.

Currently, algorithm is a term which has caught a lot of attention from the technical specialists, the social scientists, and the broader public due to the computer advances that have taken place over the last decades. According to [12], all interested parties are using the word in different ways. From an engineering perspective, an algorithm is defined as a system with two components, the logic and the control component [13]. The logic component prescribes the problem-solving knowledge it relies on. The control component prescribes how that knowledge is implemented, determining its efficiency. In addition, algorithms are typically related to an input and an output, which refer to the data on which the logic of the algorithm is executed and the results of the execution, respectively (Figure 2).



**Figure 2.** Decision-making process and decomposition of algorithms into their characteristics and components.

According to [14], the definition of algorithm is even extended to abstract machines and computer models in the Computer Science field, which makes it difficult for other fields to understand the real meaning of the algorithm concept. Technically, its logic can be as varied as the problems, contexts, systems, design, and everything that affects it, directly or indirectly. This variety in the interdisciplinary application of algorithms makes it difficult to find a standard classification of algorithms in the literature. However, they have been mainly classified by function [15], implementation, or design paradigms [16,17].

Many concepts related to algorithms are widely used today. Among the most common are ML, AI, deep learning, and black-box models [18]. AI is a concept generally understood as “that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function properly and with foresight in its environment” [19]. These algorithms are widely used in automated decision-making processes based on big

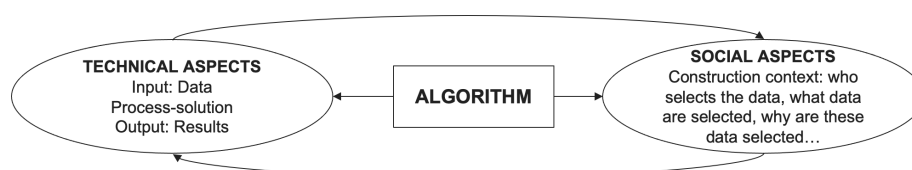
data. In some cases, it is not clear why a particular decision was made with a lack of transparency and a high risk of biases in the data-algorithm process due to human biases and data collection artifacts that are hidden during the data training process. This can lead to erroneous and unfair decisions [20]. This inconsistency in models has encouraged the development of a sub-field of ML known as ‘fairness’, which refers to the study of processes and techniques to ensure that biases in data and models do not generate unfavorable outcomes at the societal level, i.e., discrimination based on race, gender, disabilities, sexual orientation, religious beliefs, or political convictions [21].

## 2.2. Algorithm Concept in Social Sciences

Mathematics and engineering are not the only knowledge areas that have been interested in algorithms. One of the trends characterizing the current western societies is the increasing role of mathematics and computer science as influential mechanisms that shape and guide our behavior and the governance of society. This is what experts, such as Danaher et al. [22], have agreed to call “algocracy” or governance of the algorithms. Thus, algorithms cease to be understood as autonomous mathematical formulas, and begin to be conceptualized also in the context of their social impact.

From a social perspective, algorithms have been conceived, from their significance and power, as a particular form of rationality associated with a general mode of social ordering [23], or directly as a culture [24]. Beyond their diversity, what all these approaches have in common is the idea that talking about algorithms is not simply talking about technical materials. It also arises the need for new additional meanings of the algorithm concept, based on their applications in everyday life situations. Algorithms are not only created in a social context and by certain social groups –which respond to certain interests–, but they frequently intervene and operate in social reality. As pointed out by [25], an algorithm is usually employed in a context of trial and error processes, interactions, collaboration, discussion and negotiation among various intervening actors. All these actions show a common social and cultural background. Hence, to consider them as the expression of a pure mental effort (i.e., as a process of abstraction), might be an oversimplification of reality.

While a social scientist is usually unable to fully understand the mathematical details inside an algorithm, a computer scientist might not always be fully aware of the social and cultural context in which her algorithms will be implemented. These background divergences generate barriers in the way both communities communicate with each other. In order to establish a common ground between these communities, this paper enriches the technical definition of an algorithm with a social one. Hereby, this paper offers a broad view of the algorithm concept which goes beyond the idea of an autonomous mathematical entity to consider it as a complex socio-technical system. Within this common perspective, the concept of “algorithmic culture” is born as one that embraces algorithms simultaneously both as computer technology and as formative tools for the social domain. To understand how algorithms, society, and culture are intertwined, Draude et al. [2] identify two overlapping but distinguishable levels of entanglement. First, the level of social inequalities that are reproduced by combining practical purpose and automated decision-making. Secondly, the level of cultural work of algorithms in sorting and ranking. Figure 3 represents a more social view of the algorithm concept, one which highlights the importance of technical and social aspects in their composition, while assuming that both aspects are articulated in a co-constitutive manner.



**Figure 3.** Socio-technical definition of the algorithm concept.

One of the concerns that has guided social science interest in algorithms is the conviction that they are likely to produce and reproduce social inequalities. The socio-technical perspective is clear in this regard: algorithms and their effects are contextualized within the human world [2]. On this basis, algorithms should be divided not according to what they “are” but according to what they “do”. Thus, considering algorithms as socio-technical systems implies also considering the possible existence of biases in information systems, specifically in historical data employed to train algorithms. This social perspective aims at analyzing the possible existence of discrimination factors in the data-algorithmic process. As stated by Wellner and Rothman [26], an AI algorithm is likely to be considered neutral, and possible bias are usually linked to the training dataset. Furthermore, depending on the specific feedback mechanism, biased results might be used as new inputs, thus deepening the bias. The social perspective calls for a technological revolution, in which machine learning not only “teaches” the algorithm to identify an object and classify it, but also “educates” it to reflect certain social values [27]. As Wellner and Rothman [26] also state, the challenge of “educating” an algorithm to identify a gender bias is analogous to the one of training an algorithm to be ethical or fair. The importance of users increases when we realize that fairness is not an easily definable parameter to introduce into a given system. It is complex and changes over time and place, but users of a system can sometimes detect it more effectively than developers. Therefore, developers must design systems so that users can easily report biases.

### 3. Data-Algorithmic Bias: Definitions and Classifications

Bias is a concept used in statistics and ML to refer to the assumptions made by a specific model [28]. Viewing algorithms as socio-technical systems implies, however, a broader understanding of bias in information systems. Friedman and Nissenbaum [29] use the term bias to refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others. According to these authors, a system discriminates unfairly if it denies an opportunity, or assigns an undesirable outcome to an individual, on grounds that are unreasonable or inappropriate. In the same vein, Ntoutsis et al. [30] define bias as the inclination or prejudice of a decision made by an AI system that is in favor of, or against to, an individual or group, especially in a way that is considered unfair. We find general agreement in the scientific literature when it comes to defining computer biases on the basis of the injustices they are likely to generate. Despite this, neither unfair discrimination alone gives rise to bias unless it occurs systematically, nor does systematic discrimination generate bias unless it is accompanied by an unfair outcome [29]. Since AI relies heavily on data generated by humans –or collected through human-created systems–, it is not strange that any bias that exists in humans is also reflected in our systems. As algorithms are part of existing biased institutions and structures, they are susceptible to reproduce and amplify these biases by favoring those phenomena and aspects of human behavior that are easily quantifiable over those that are difficult or even impossible to measure [30].

Friedman and Nissenbaum [29] were among the first to offer a classification of the types of biases related to algorithms from a socio-technical perspective. Their analysis was based on an analysis of seventeen computer systems from various fields, including banking, commerce, computer science, education, medicine, and law. These authors categorize three types of biases in their work. The first is the pre-existing bias, i.e., when computer systems incorporate biases that exist independently to the creation of the system. The second type of bias described is the technical bias. In contrast to pre-existing bias, technical bias arises from problem solving in technical design. Technical bias can be identified in computer tools, in the de-contextualization of algorithms, in the generation of pseudo-random numbers, or in the formalization of human constructs. Finally, the third type of bias is the emergent bias, which arises only in a context of use, usually some time after a design is completed, as a result of changes in social knowledge, population, or cultural values. Emergent bias may be caused by new knowledge in society that cannot be incorporated into the system



design. It can also be a bias that arises when the population using the system differs in some significant dimension from the population assumed as users during the design stage.

Mehrabi et al. [31] present a classification that considerably expands these types of bias. This work offers a particularly complete and updated definition of the biases affecting AI applications. However, its main contribution is the classification of biases beyond the different phases of the algorithmic process in which they fall, i.e.,: in the data, in the creation of the algorithms, or in the interaction with the users. These authors analyze biases in ML from the phenomenon of the feedback loop. This feedback loop is understood as a situation in which the trained ML model makes decisions that produce results, and these same results affect future data to be collected for subsequent training rounds or models. Aggregation bias occurs when false conclusions are drawn for one subgroup based on the observation of a different one or, in general, when false assumptions about a population affect the outcome and definition of the model. Temporal bias arises from differences between populations and behaviors over time [32]. Likewise, social bias occurs when other people's actions or content coming from them affect our judgment [33]. From algorithms to interaction with users, Mehrabi et al. [31] situate four other biases: (i) the popularity bias –or overexposure of the most popular articles [34,35]; (ii) the ranking bias that correlates with the idea that the best ranked results are the most relevant and important –which will attract more clicks than others; (iii) the evaluation bias, which includes the use of inappropriate and disproportionate benchmarks for application evaluation; and (iv) the emergent bias, which arises as a result of a change in population, cultural values or societal knowledge, usually some time after the completion of the design [29]. As can be seen, this last categorization encompasses and extend all the biases we have discussed in previous classifications. Moreover, this last proposal goes one step further by illustrating the data-algorithmic process as a feedback loop composed by: (i) the algorithm; (ii) the data feeding the algorithm; and (iii) the interaction with the user that is affected by the algorithm's outcome.

#### 4. Examples of Gender Bias

This section reviews scientific works that study examples of gender biases in data-algorithmic processes. It is structured in four subsections describing applications in natural language processing and generation, speech recognition, decision management, and face recognition.

##### 4.1. Natural Language Processing and Generation

Most of the research looking at gender biases in AI and algorithms has been conducted in the field of computational linguistics. The goal is to avoid biases in the training of those algorithms involved in the development of voice assistants.

Gender bias is susceptible of being exhibited in multiple parts of an NLP system, including training data, resources, pre-trained models (for example, word embeddings), and the algorithms themselves. One of the first works to point out gender stereotypes within NLP was the one carried out by [36] on Word2Vec. For these authors, the blind application of ML runs the risk of amplifying the biases already present in the data. Such a risk confronts us with word embedding, a popular framework for representing text data as vectors that has been used in many NLP and ML tasks. From here, they define gender bias as the correlation between the magnitude of the projection in the gender subspace of an embedded word representing a gender-neutral word and the bias rating of that word, as rated by crowd workers. In their research, they identified problematic stereotypical representations of men and women. [8] adopt the core concept of the implicit association test (IAT), which is used in psychology to measure bias in word embedding. The authors confirm in their work that there are gender biases found through the IAT test in the GloVe and Word2Vec embedding tools, highlighting the stereotypical association of masculinity and femininity with sciences and arts respectively. The previous studies show that word embedding learns from large

corpus of text available online, and that the presence of gender biases in them might be a reflection of underlying biases in the society.

#### 4.2. Speech Recognition

Speech recognition is another form of AI likely to exhibit gender bias. Tatman [37] evaluates the accuracy of subtitles automatically generated by YouTube in two genders and five dialects of English. The dialect and gender of speakers were controlled using videos uploaded as part of the “accent tag challenge”, where speakers explicitly identify their linguistic background. The results show robust differences in accuracy across both gender and dialect, with lower accuracy for female and Scottish speakers. According to this research, these disparities exist because of the way we have structured our data analysis, databases, and machine learning methods. Similar to how cameras are customized to photograph white faces, audio analysis struggles with higher-pitched voices. The underlying reason may be that the databases have a lot of data on white males and less data on female and minority voices. Tatman and Kasten [38] compare the accuracy of two automatic speech recognition (ASR) systems –Bing Speech and YouTube’s automatic captions– across gender, race and four dialects of American English. According to their results, there is an inaccuracy of ASR systems in dealing with socio-linguistic variations.

#### 4.3. Decision Management

Nowadays, more and more decisions about loans, grants or study applications are partially automated based on models relying on historical data. Different studies have evidenced the existence of biases. According to Dastin [39], a well-known hiring tool preferred male candidates over female ones. Likewise, some credit services seemed to offer smaller credit lines to women than to men. It is entirely possible for data-algorithmic processes to discriminate by gender even when they are programmed to be “blind” to that variable. Just as in the first example given, a blind algorithm could end up biased against a gender if it relies on inputs (data) that correlate with this variable.

#### 4.4. Face Recognition

Many of the algorithms used in image or voice recognition applications are based on gender stereotypes. These, in turn, rely on the physical appearance of the people who are supposed to use them. Hence, the data-algorithmic process might reproduce false beliefs about what the physical attributes that define people should be like according to their biological sex, ethnic or cultural background, or sexual orientation. Many times, the dominant cultural pattern corresponds to a white male belonging to an affluent group. A similar criticism is found in most of the works dedicated to this topic [40]. Hence, Buolamwini and Gebru [41] argue that most biases in face recognition applications are located in the datasets. According to these authors, the most commonly used datasets for benchmark tests are mostly male (77%) and of Caucasian origin (between 79% and 86% are white faces). For this reason, they classify white males better than females or people from other ethnicity, which could generate both gender and race biases.

### 5. Datasets with Gender Bias

The UCI Machine Learning Repository (Available online: <https://archive.ics.uci.edu/ml/index.php>, accessed on 19 July 2022) is a well-known collection of databases, domain theories, and data generators. The machine learning community use them for educational purposes and the empirical analysis of algorithms. We focus on the dataset called ‘Adult dataset’, also known as ‘Census Income dataset’, which is based on census data from 1994. The prediction task posed is to determine whether a person makes over 50K a year. The dataset contains 48,842 observations split into a training set and a test set with 32,561 and 16,281 observations, respectively. There are 51 papers that cite this dataset.

In the training set, there is information regarding 10,771 females and 21,790 males (33.08% vs 66.92%). Thus, we find imbalance in the class classification, which may results

in models that have poor predictive performance, specifically for the minority class. Only 10.95% of females belong to the class '>50K', while this percentage reaches 30.57% for males. Based on this historical dataset, it can be concluded that a higher percentage of males have a higher income.

A classification tree is built using the training set and predictions are made for the test set [42,43]. The independent variables employed are: age (continuous), sex (categorical), workclass (categorical), occupation (categorical), education (continuous), hours per week (continuous), and native country (categorical). The classification tree is a popular and easy-to-interpret method. The accuracy is 80.84%, considering all the individuals; 76.70% for males and 89.12% for females. The sensitivity (defined as the percentage of individuals in the category '<50K' for which the prediction is correct) is 91.79% for males and 100% for females. The specificity (defined as the percentage of individuals in the category '>50K' for which the prediction is correct) is 41.46% for males and 0% for females. Thus, while the accuracy reached by the classification method is relatively high, the values of specificity are low, extremely low for females. The method always predicts lower income, '<=50K', for females. The use of this algorithm for decision-making, for example, linked to granting mortgages or allowing rent could have serious negative social impacts.

We may find examples of dataset with gender bias in other AI fields. A case in point is COCO, a large-scale object detection, segmentation, and captioning dataset. The process to build the dataset is detailed in [44], which is a work with more than 26K citations in Google Scholar. Recent authors have pointed out that the occurrence of men in image is significantly higher than women and the gender disparity reaches high values for specific contexts [45]. For instance, 90% of surfboard images only contain male players. The dataset imSitu [46] constitutes another example. imSitu supports situation recognition, the problem of producing a concise summary of the situation an image depicts. This resource exhibits gender bias. For instance, women are represented as cooking twice as often as men [47].

## 6. Initiatives to Address Gender Bias

This section describes a few recent, relevant, and representative private initiatives to mitigate gender bias as well as related recommendations of international organizations.

### 6.1. Private Initiatives

Google translate announced gender-specific translations in 2018. This feature provides options for both feminine and masculine translations when translating queries that are gender-neutral in the source language. IBM introduced in 2018 AI Fairness 360, an extensible open-source library with techniques developed by the research community to help detect and mitigate bias in machine learning models throughout the AI application life-cycle. The package includes a set of metrics for datasets and models to test for biases, explanations for these metrics, and algorithms to mitigate bias in datasets and models. Now the library is available in both Python and R. Facebook launched in 2018 a tool called the Fairness Flow, a diagnostic tool that enables its teams to analyze how some types of AI models and labels perform across different groups.

More recently, Google has published ML-fairness-gym, a set of components for building simulations that explore the potential long-run impacts of deploying machine learning-based decision systems in social environments. Thus, they allow the study of the effects of fairness in the long run [48]. These tools are shared in GitHub and run with Python 3.

### 6.2. International Organizations

The UNESCO remarks the need for a human-centred AI, points out that AI contributes to widening existing gender gaps (in particular, gender biases and stereotyping are being reproduced because women are underrepresented in the industry), and is currently elaborating the first global standard-setting instrument on the ethics of AI in the form of a recommendation (Available online: <https://en.unesco.org/artificial-intelligence/ethics>, accessed on 19 July 2022). The preliminary study on the ethics of AI defines 'inclusiveness'



as a generic principle for the development, implementation and use of AI: “AI should be inclusive, aiming to avoid bias and allowing for diversity and avoiding a new digital divide”. Similarly, the Council of Europe recommendation on the human rights impacts of algorithmic systems proposes a set of guidelines for both States and public and private sector actors (Available online: <https://search.coe.int/cm>, accessed on 19 July 2022). The Council recommends a precautionary approach monitoring socio-technical developments to protect human rights. It highlights that datasets often contain bias and may stand in as a proxy for classifiers such as gender, race, religion, political opinion, or social origin, and points out the importance of enhancing public awareness and discourse.

Similarly, the council of the OECD on AI provides a set of internationally-agreed principles and recommendations. The recommendations for policy-makers include: investing in AI R&D, fostering a digital ecosystem for AI, providing an enabling policy environment for AI, building human capacity and preparing for labour market transition, and international co-operation for trustworthy AI. The value-based principles are: inclusive growth, sustainable development and well-being, human-centred values and fairness, transparency and explainability, robustness, security and safety, and accountability.

## 7. An Illustrative Numerical Example

In order to illustrate some of the previously described concepts, this section introduces a numerical example that aims at: (i) showing a typical example of discrimination due to data bias, which lead us to a wrong model despite using logistic regression, which is a well-tested, gender-agnostic, and race-agnostic machine learning algorithm; and (ii) propose a simple yet effective methodology to solve the issue. Table 1 contains 92 observations related to bank users’ applications for a financial credit during the last month. The following notation has been used: ‘G’ refers to gender, which can be *Male* (M) or *Female* (F); ‘R’ refers to race, which can be *White* (W) or *Other* (O); ‘S’ refers to the risk *Score* (S) obtained by the applicant after a series of tests, where the higher the score, in a scale from 0 to 120, the more risk is assumed by the financial entity; finally, ‘A?’ refers to whether the application has been approved by the bank (Y) or not (N).

**Table 1.** Row data for the example.

#	G	R	S	A?	#	G	R	S	A?
1	M	O	75	Y	47	M	W	65	Y
2	M	O	70	Y	48	F	O	35	N
3	F	O	55	Y	49	M	O	55	Y
4	F	O	25	Y	50	M	O	80	Y
5	M	O	60	Y	51	M	O	55	Y
6	M	O	50	Y	52	F	W	85	Y
7	M	O	65	N	53	F	W	60	Y
8	M	W	25	Y	54	F	O	65	Y
9	M	W	20	Y	55	M	W	67	Y
10	M	W	77	Y	56	M	O	60	N
11	F	W	55	N	57	M	W	65	Y

**Table 1.** *Cont.*

12	M	W	60	Y	58	F	O	75	N
13	F	O	62	N	59	M	W	35	Y
14	M	W	70	Y	60	F	O	25	Y
15	M	W	45	Y	61	M	O	70	N
16	M	W	40	Y	62	F	O	65	N
17	F	O	40	Y	63	F	O	51	Y
18	F	O	45	Y	64	M	W	75	Y
19	F	W	35	Y	65	M	W	73	Y
20	M	W	80	Y	66	M	O	79	N
21	M	O	45	Y	67	M	O	92	Y
22	M	O	58	Y	68	M	O	60	Y
23	M	O	85	Y	69	M	W	85	N
24	F	W	30	Y	70	M	O	95	Y
25	M	O	75	N	71	M	W	85	Y
26	M	W	95	Y	72	F	W	84	N
27	F	O	85	Y	73	M	W	95	Y
28	M	O	77	N	74	M	O	97	Y
29	F	O	94	N	75	M	O	90	Y
30	M	O	90	Y	76	F	O	80	N
31	M	O	99	N	77	M	W	90	Y
32	M	W	70	Y	78	M	O	97	N
33	F	O	65	N	79	M	W	93	Y
34	F	W	103	Y	80	M	O	100	Y
35	M	O	90	Y	81	M	W	113	Y
36	M	W	25	Y	82	M	W	100	Y
37	M	W	60	Y	83	M	W	65	Y
38	F	O	45	Y	84	M	O	105	Y
39	M	W	60	Y	85	M	O	99	N
40	M	W	0	Y	86	F	W	107	Y
41	F	W	65	Y	87	M	O	120	N
42	F	W	70	Y	88	F	W	90	N
43	M	W	60	Y	89	M	W	82	Y
44	M	W	60	Y	90	M	O	105	Y
45	M	W	65	Y	91	M	O	65	N
46	F	W	60	Y	92	M	W	107	Y

One could assume that the acceptance or rejection of the credit application should be mainly based on the score that the candidate has achieved after a series of rigorous tests –and, possibly, on some other variables–, but not on the actual gender or race of the candidate. Contingency tables can provide some insights on how the acceptance of the credits has been distributed by gender and race. Figures 4 and 5 show that while women represent around 30% of the sample, they only get around 26% of the accepted credits. This difference, however, does not seem to be statistically significant: for a standard  $\alpha$  value

of 0.05, the Pearson Chi-Square test results in a  $p$ -value of 0.079 (and, a similar  $p$ -value is obtained with the Likelihood Ratio Chi-Square test). In the case of race, however, things are different: while non-white applicants represent around 51% of the sample, they only get around 41% of the approved credits. This results in both groups (white and other) being significantly different in terms of their odds to get a credit (both the Pearson and the Likelihood Chi-Square tests reflect an extremely low value of 0.001 for the associated  $p$ -value).

#### Tabulated Statistics: Accepted?, Gender

Rows: Accepted?		Columns: Gender		
		Female	Male	All
No	Count	10	12	22
	% of Row	45.45	54.55	100.00
Yes	Count	18	52	70
	% of Row	25.71	74.29	100.00
All	Count	28	64	92
	% of Row	30.43	69.57	100.00
Cell Contents:		Count % of Row		
Pearson Chi-Square = 3.081, DF = 1, P-Value = 0.079				
Likelihood Ratio Chi-Square = 2.946, DF = 1, P-Value = 0.086				

**Figure 4.** Contingency table of acceptance by gender.

#### Tabulated Statistics: Accepted?, Race

Rows: Accepted?		Columns: Race	
	Other	White	All
No	18 81.82	4 18.18	22 100.00
Yes	29 41.43	41 58.57	70 100.00
All	47 51.09	45 48.91	92 100.00
Cell Contents:		Count % of Row	

Pearson Chi-Square = 10.928, DF = 1, P-Value = 0.001  
Likelihood Ratio Chi-Square = 11.660, DF = 1, P-Value = 0.001

**Figure 5.** Contingency table of acceptance by race.

As a next step, we have used the data in Table 1 to generate a logistics regression model, which aims at predicting the outcome of new users' applications based on their score, gender, and race. Figure 6 displays the deviance table for the obtained model. Notice that the  $p$ -value associated with the regression model is 0.000, which means that at least one of the predictive variables employed (score, gender, and race) can help to explain the acceptance decision process. Actually, all three variables have  $p$ -values lower than  $\alpha = 0.005$ , which means that all these variables play an important role in our model.

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	3	19.084	6.3614	19.08	0.000
Score	1	5.039	5.0388	5.04	0.025
Gender	1	5.060	5.0605	5.06	0.024
Race	1	11.510	11.5100	11.51	0.001
Error	88	82.130	0.9333		
Total	91	101.214			

**Figure 6.** Deviance table for the logistics regression model.

Figure 7 shows the logistics regression model obtained. This model can be used to predict the probability that a new user's application is accepted,  $P(\text{Yes})$ , based on the composed expression that employs exponential functions,  $\exp(\cdot)$ , of a new variable  $Y'$ , which is defined as a piecewise function of the recorded variables. Hence, for instance, if the new candidate is a white female,  $Y' = 3.695 - 0.02945 \cdot \text{Score}$ , and so on.

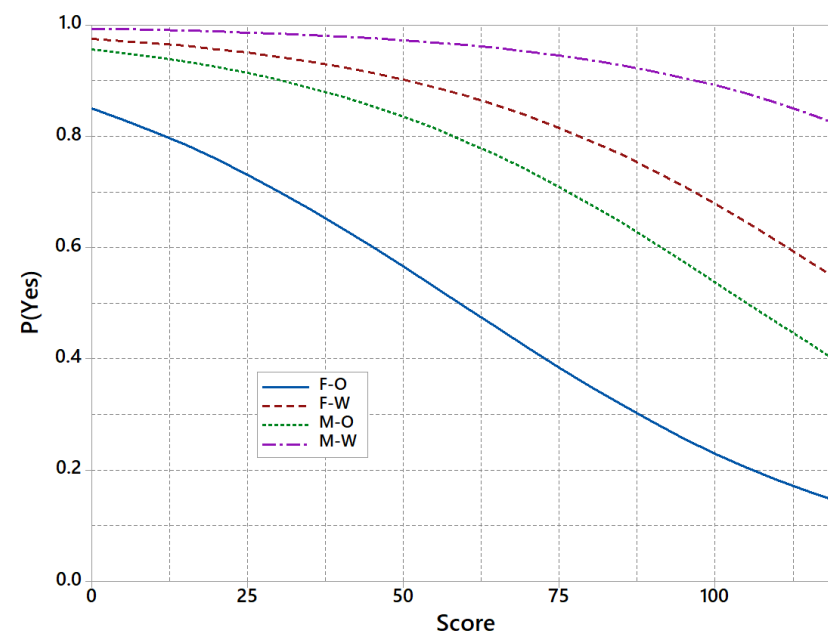
Regression Equation

$$P(\text{Yes}) = \exp(Y') / (1 + \exp(Y'))$$

Gender Race  
 Female Other  $Y' = 1.738 - 0.02945 \text{ Score}$   
 Female White  $Y' = 3.695 - 0.02945 \text{ Score}$   
 Male Other  $Y' = 3.100 - 0.02945 \text{ Score}$   
 Male White  $Y' = 5.057 - 0.02945 \text{ Score}$

**Figure 7.** Logistic regression model.

Despite we found no significant gender differences when applying for a credit, the model is still proposing different coefficients for each gender. This is even worse in the case of race, where significant differences were found regarding the application rate of success. Figure 8 shows how the probability of having the application accepted varies with the risk score but also depends on the group to which a new customer belongs. Hence, for the same score a non-white female has significantly less options than a white male.

**Figure 8.** Biased model for  $P(\text{Yes})$  vs Score by group.

Since the problem is not in the algorithm, it must be in the raw data. One easy way to correct this bias problem in the original data is to create a new model that does not consider

either race or gender, but just score. Still, if we suspect that the score data might also be somewhat biased by gender or race, then a better alternative could be to simply randomize the values inside these columns, so higher scores are not more frequent in white men (or any other gender-race combination) and the model cannot assign significantly different probabilities based on the gender or race of the applicant. Figure 9 shows the adjusted model, which makes use of the randomized gender and race variables. Despite the model is still employing slightly different independent terms for each gender-race combination, this is just the result of a random assignment, so these differences are not going to be significant in any case. Notice that, in fact, the variability of the independent terms in the adjusted model is much lower than in the original one.

Regression Equation

$$P(\text{Yes}) = \exp(Y') / (1 + \exp(Y'))$$

R_Gender	R_Race	Y'
Female	Other	$Y' = 2.405 - 0.01687 \text{ Score}$
Female	White	$Y' = 2.858 - 0.01687 \text{ Score}$
Male	Other	$Y' = 1.979 - 0.01687 \text{ Score}$
Male	White	$Y' = 2.432 - 0.01687 \text{ Score}$

Figure 9. Adjusted logistic regression model.

Finally, Figure 10 displays, for the adjusted model, how the probability of getting the application accepted varies with the score. Notice that differences among groups have been reduced, and reduced to random effects. In other words, fixed a risk score, the model assigns approximately the same probabilities to a non-white female and to a white male. Hence, it makes sense to use only one model (e.g., the F-O or the M-W ones) to make predictions regardless of the gender and race.

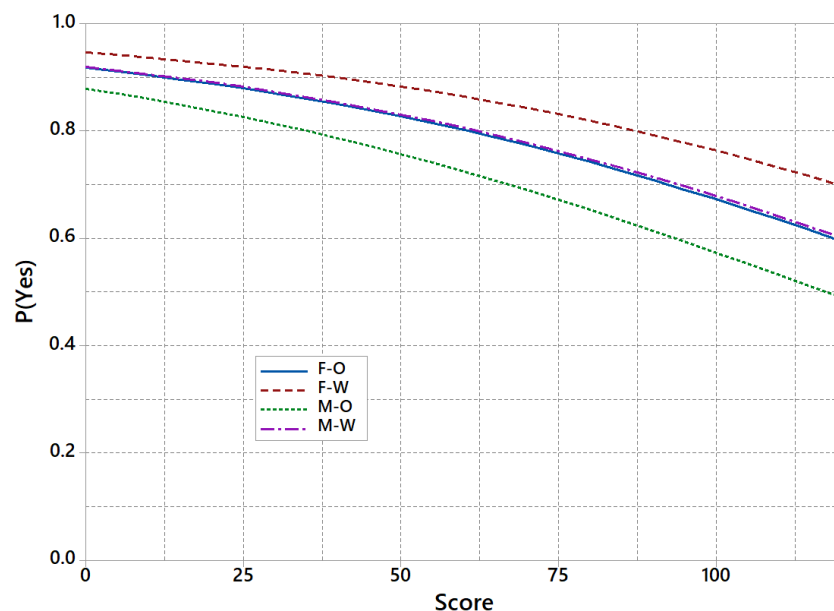


Figure 10. Adjusted model for  $P(\text{Yes})$  vs  $\text{Score}$  by group.

By using a randomized assignment of gender and race values, this model has also fixed any potential bias of data in the score variable. In addition, the curves are now less disperse, with no one being permanently too close to one and no one approaching too much to zero for large risk scores. This can be seen as a positive externality of the adjustment process, and one that makes the credit assignment more socially responsible.



Finally, notice that the proposed approach can also be extended to many other machine learning algorithms, in particular the supervised ones.

## 8. Recommendations to Prevent, Identify, and Mitigate Gender Bias

The importance of ethics in algorithm processes has been explored and discussed for a long time. However, the topic of integrating the gender dimension in data-algorithmic processes is increasingly getting more attention as the number of AI applications grows exponentially as well as relevant examples of gender bias in AI are identified in the scientific literature. In addition, the training datasets do not always contain representative public demographics limiting the integration of vulnerable groups amplifying the gender biases. Aiming at contributing to reduce the gender bias, we provide a list of recommendations for AI practitioners. They are classified into 3 categories, depending on whether they aim to prevent, identify, or mitigate gender bias.

- Preventing gender bias: (i) configure a reasonable representation of both genders among each category of experts working in the design, implementation, validation, and documentation of algorithms; (ii) set a reasonable gender distribution among each category of experts working in the extraction/collection, pre-processing, and analysis of data; (iii) incorporate at least one expert in data-algorithmic bias to the group; and (iv) train all staff (male/female/non-bi) in gender bias (and approaches to prevent, avoid, detect, and correct it).
- Identifying gender bias: (i) be transparent regarding the composition of the working group (gender distribution and expertise in ethics and data-algorithmic bias), the strategies implemented to mitigate bias, and the results of the tests implemented to detect potential bias; (ii) assess and publish the limitations regarding gender bias; (iii) improve interpretability of ‘black-box’ models; and (iv) analyze periodically the use and results of the algorithms employed.
- Mitigating gender bias: (i) avoid to reuse data and pre-trained models with gender bias that cannot be corrected; (ii) apply methods to get a balanced dataset if needed [49], as well as to measure accuracy levels separately for each gender; (iii) assess different fairness-based measures to choose which ones are more suitable in a particular case; (iv) test different algorithms (and configurations of parameters) to find which one outperforms the others (benchmark instances or datasets with biases are available in the literature to assess new algorithms); (v) modify the dataset to mitigate gender bias relying on specific-domain experts; (vi) document and store previous experiences where bias has been detected in a dataset and how it has been mitigated (as commented before, gender bias tend to be recurrent in some specific fields); and (vii) implement approaches to remove unwanted features related to gender from intermediate representations in deep learning models.

## 9. Conclusions

Algorithms are becoming increasingly employed for high-stakes decision-making in a wide range of fields (from financial loans to universities admissions or hiring practices). In addition, a data-driven culture is being established in an increasing number of companies and governments. As the number of AI applications grows, as well as their capabilities and relevance, it is important to assess the potential data-algorithmic biases. While this is not a new concept, there are plenty of examples of AI applications where this issue is not studied, thus ignoring the potential consequences. From all the types of data-algorithmic biases, this work focuses on gender bias. We have discussed examples of gender bias in different AI fields (natural language processing, face recognition, and decision management processes such as recruitment, among others). Some of the main sources of this bias are the under-representation of women in the design and development of AI products and services, as well as the use of datasets with gender bias [50,51]. The latter issue would probably be minimized with the incorporation of statisticians in the AI development teams, since these experts can help to avoid using biased datasets during the algorithm training process. We

have reviewed the scientific works aiming to mitigate this type of bias and have pointed out some private initiatives to deal with it in specific applications. Furthermore, an illustrative numerical example is provided. This example proposes a simple yet effective methodology to identify and correct possible gender bias in many machine learning algorithms. Finally, we have proposed a list of general recommendations for any AI practitioner.

The development and use of AI applications is increasing across companies from a wide range of fields and governments. Hence, it is of vital importance to detect and mitigate gender bias in data-algorithmic processes, which may have huge impacts for women and society in general. The multiple sources of gender bias, as well as the particularities of each type of algorithm and dataset, makes removing bias a particularly difficult challenge. Because of the difficulty of addressing this issue and the potential impacts that may have, it becomes necessary the adoption of an interdisciplinary approach, as well as the close cooperation among companies and governments.

**Author Contributions:** Conceptualization, J.C. and M.S.; methodology, A.J. and A.A.J.; software, A.A.J.; validation, S.Y. and M.S.; formal analysis, A.A.J.; writing—original draft preparation, A.J., J.C. and L.C.; writing—review and editing, A.A.J. and M.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work has been partially funded by the Spanish Ministry of Culture and Sports (02/UPR/21).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Pearl, J. *Probabilistic Reasoning in Intelligent Systems*; Kaufmann: San Mateo, CA, USA, 1988.
- Draude, C.; Klumbyte, G.; L’ucking, P.; Treusch, P. Situated algorithms: A sociotechnical systemic approach to bias. *Online Inf. Rev.* **2019**, *44*, 325–342.
- Seaver, N. What should an anthropology of algorithms do? *Cult. Anthropol.* **2018**, *33*, 375–385.
- Photopoulos, J. Fighting algorithmic bias. *Phys. World* **2021**, *34*, 42.
- Ahmed, M.A.; Chatterjee, M.; Dadure, P.; Pakray, P. The Role of Biased Data in Computerized Gender Discrimination. In *Proceedings of the 2022 IEEE/ACM 3rd International Workshop on Gender Equality, Diversity and Inclusion in Software Engineering (GEICSE)*, Pittsburgh, PA, USA, 20 May 2022; pp. 6–11.
- Kuppler, M. Predicting the future impact of Computer Science researchers: Is there a gender bias? *Scientometrics* **2022**, 1–38. <https://doi.org/10.1007/s11192-022-04337-2>.
- Brunet, M.E.; Alkalay-Houlihan, C.; Anderson, A.; Zemel, R. Understanding the origins of bias in word embeddings. In *Proceedings of the International Conference on Machine Learning*, June 9–15 2019, Long Beach, CA, USA; pp. 803–811.
- Caliskan, A.; Bryson, J.J.; Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **2017**, *356*, 183–186.
- Mittelstadt, B.D.; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data Soc.* **2016**, *3*, 1–21.
- Tsamados, A.; Aggarwal, N.; Cows, J.; Morley, J.; Roberts, H.; Taddeo, M.; Floridi, L. The ethics of algorithms: Key problems and solutions. *AI Soc.* **2022**, *37*, 215–230.
- Taddeo, M.; Floridi, L. The debate on the moral responsibilities of online service providers. *Sci. Eng. Ethics* **2016**, *22*, 1575–1603.
- Gillespie, T. Algorithm. In *Digital Keywords*; Princeton University Press: Princeton, NJ, USA, 2016; Chapter 2, pp. 18–30.
- Kowalski, R. Algorithm = logic + control. *Commun. ACM* **1979**, *22*, 424–436.
- Moschovakis, Y.N. What is an Algorithm? In *Mathematics Unlimited—2001 and Beyond*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 919–936.
- Sedgewick, R.; Wayne, K. *Algorithms*; Addison-Wesley Professional: Boston, MA, USA, 2011.
- Brassard, G.; Bratley, P. *Fundamentals of Algorithmics*; Prentice-Hall, Inc.: Englewood Cliffs, NJ, USA, 1996.
- Skiena, S.S. *The Algorithm Design Manual*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020.
- Mohri, M.; Rostamizadeh, A.; Talwalkar, A. *Foundations of Machine Learning*; MIT Press: Cambridge, MA, USA, 2018.
- Nilsson, N.J. *The Quest for Artificial Intelligence*; Cambridge University Press: Cambridge, UK, 2009.

20. Pedreschi, D.; Giannotti, F.; Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F. Meaningful explanations of black box AI decision systems. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 9780–9784.
21. Oneto, L.; Chiappa, S. Fairness in machine learning. In *Recent Trends Learn from Data*; Springer: Cham, Switzerland, 2020; pp. 155–196.
22. Danaher, J.; Hogan, M.J.; Noone, C.; Kennedy, R.; Behan, A.; De Paor, A.; Felzmann, H.; Haklay, M.; Khoo, S.M.; Morison, J.; et al. Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data Soc.* **2017**, *4*, 2053951717726554.
23. Beer, D. Power through the algorithm? Participatory web cultures and the technological unconscious. *New Media Soc.* **2009**, *11*, 985–1002.
24. Seaver, N. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data Soc.* **2017**, *4*, 2053951717738104.
25. Kitchin, R. Thinking critically about and researching algorithms. *Inf. Commun. Soc.* **2017**, *20*, 14–29.
26. Wellner, G.; Rothman, T. Feminist AI: Can we expect our AI systems to become feminist? *Philos. Technol.* **2020**, *33*, 191–205.
27. Ihde, D. *Technosystem: The Social Life of Reason* by Andrew Feenberg. *Technol. Cult.* **2018**, *59*, 506–508.
28. Mitchell, T.M. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997.
29. Friedman, B.; Nissenbaum, H. Bias in computer systems. *ACM Trans. Inf. Syst. (TOIS)* **1996**, *14*, 330–347.
30. Ntoutsis, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdil, W.; Vidal, M.E.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; et al. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1356.
31. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–35.
32. Olteanu, A.; Castillo, C.; Diaz, F.; Kiciman, E. Social data: Biases, methodological pitfalls, and ethical boundaries. *Front. Big Data* **2019**, *2*, 13.
33. Baeza-Yates, R. Bias on the web. *Commun. ACM* **2018**, *61*, 54–61.
34. Introna, L.; Nissenbaum, H. Defining the web: The politics of search engines. *Computer* **2000**, *33*, 54–62.
35. Prates, M.O.; Avelar, P.H.; Lamb, L.C. Assessing gender bias in machine translation: A case study with Google translate. *Neural. Comput. Appl.* **2020**, *32*, 6363–6381.
36. Bolukbasi, T.; Chang, K.W.; Zou, J.Y.; Saligrama, V.; Kalai, A.T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 4349–4357.
37. Tatman, R. Gender and dialect bias in YouTube’s automatic captions. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, Valencia, Spain, 4 April 2017; pp. 53–59.
38. Tatman, R.; Kasten, C. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 934–938.
39. Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics*; Auerbach Publications: Boca Raton, FL, USA, 2018; pp. 296–299.
40. Ensmenger, N. Beards, sandals, and other signs of rugged individualism: Masculine culture within the computing professions. *Osiris* **2015**, *30*, 38–65.
41. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; pp. 77–91.
42. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
43. Therneau, T.; Atkinson, B. *Rpart: Recursive Partitioning and Regression Trees*; R package Version 4.1-15; 2019. Available online: <https://cran.r-project.org/web/packages/rpart/index.html> (accessed on 19 July 2022).
44. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
45. Tang, R.; Du, M.; Li, Y.; Liu, Z.; Zou, N.; Hu, X. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference*, Ljubljana, Slovenia, 19–23 April 2021; pp. 633–645.
46. Yatskar, M.; Zettlemoyer, L.; Farhadi, A. Situation recognition: Visual semantic role labeling for image understanding. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5534–5542.
47. Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; Chang, K.W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv* **2017**, arXiv:1707.09457.
48. D’Amour, A.; Srinivasan, H.; Atwood, J.; Baljekar, P.; Sculley, D.; Halpern, Y. Fairness is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 525–534.
49. Kaur, H.; Pannu, H.S.; Malhi, A.K. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 1–36.

- 
50. Panteli, N.; Urquhart, C. Job crafting for female contractors in a male-dominated profession. *New Technol. Work. Employ.* **2022**, *37*, 102–123.
  51. Tiainen, T.; Berki, E. The re-production process of gender bias: A case of ICT professors through recruitment in a gender-neutral country. *Stud. High. Educ.* **2019**, *44*, 170–184.