

# From Bias to Balance – A Systematic Literature Review on Gender Bias in AI

*Completed Research Paper*

**Frédéric Tronnier**

Goethe University  
Frankfurt am Main  
Frederic.tronnier@m-chair.de

**Sascha Löbner**

Goethe University  
Frankfurt am Main  
sascha.loebner@m-chair.de

**Alina Azanbayev**

EBS Universität  
Oestrich-Winkel  
alina.azanbayev@ebs.edu

**Max Lukas Walter**

Goethe University  
Frankfurt am Main  
maxlukas-walter92@outlook.de

## Abstract

*In this paper we provide a systematic review of gender bias in AI, focusing on bias detection, mitigation, and the challenges of addressing non-binary gender bias within textual, visual, and audio data. Underscoring the importance to address and deconstruct gender biases embedded in AI technologies, we explore the state of research on gender bias in AI and ML applications, debiasing and mitigation techniques for non-binary genders, and the specific challenges and trade-offs across different data types. Our findings reveal significant gaps in research, particularly in audio data and non-binary gender considerations, highlighting the need for more nuanced, data-type-specific approaches to promote inclusivity in AI systems. By offering a granular analysis of gender bias and proposing future directions for research, this work contributes to the broader understanding of creating inclusive and fair AI technologies.*

**Keywords:** Gender Bias; AI; Artificial Intelligence; ML; Generative AI; Machine Learning; Literature Review

## Introduction

In recent years, the rise of Artificial Intelligence (AI) has been predicted and longed for by media and scientists alike. With the emergence of ChatGPT by OpenAI, a first generative AI application could successfully achieve mainstream adoption. With the success of AI and machine learning (ML) technologies and applications, social issues and dangers were noted, in the form of bias in input data and output information. As broad as their potential and the possibilities, so too are the potential issues in the form of biases and their entry points at various stages of the development, processing and implementation lifecycle. Possible negative effects on a wide range of areas are correspondingly complex. On the one hand, entirely novel issues may arise due to new possibilities provided by the technologies, on the other hand existing social problems in the form of prejudice, misinformation or discrimination can be reinforced and amplified (Kearns & Roth, 2020).

Given that biases have long been inherent human traits, their presence in AI models poses a challenge that necessitates the identification of these biases and the thoughtful application of debiasing techniques,

including the consideration of potential trade-offs. Addressing and unlearning these biases in the creation of AI models is imperative to prevent machines from perpetuating existing societal biases, a concern that grows as AI increasingly “eats itself” by recycling and reinforcing these biases through its own generated data.

The existence of gender bias specifically, and the fact that it can be reinforced by ML applications, has already been demonstrated in several publications (Buolamwini & Gebru, 2018; Martin, 2022; Broussard, 2023). Adverse effects of gender bias occur not only at an individual level but extend towards communities and societies (Altman et al., 2018), with for instance recruiting solutions being biased against women (UNESCO, 2020). At the same time, first studies on gender bias in image data reveal the presence of potentially harmful and more nuanced biases against women in generative AI solutions that create image data (Zhou, 2024). It is therefore crucial to ensure that AI and ML applications are not biased towards specific individuals or groups (Nadeem et al., 2020)

Accordingly, methods for removing these biases at various stages of development, known as debiasing, are currently being researched (Meade et al., 2021; Zhao et al., 2018). However, new problem areas are constantly emerging in such a dynamic and expanding field, while affected groups of people also identify new potential sources of bias and feedback loops with regard to adapting social conditions, roles and norms, such as changing language or gender roles. For instance, non-binary gender roles may require more complex debiasing methods than binary gender roles (Manzini et al., 2019), raising the issue of potential trade-offs between said methods. While past AI applications and ML solutions focused strongly on the analysis of text data, emerging generative AI solutions are increasingly multi-modal. Generative AI solutions are able to assess, work with and create visual, audio and cross-category data which dramatically increases the potential for gender bias in less-researched data types (see Ulloa et al., 2022).

Considering that those topics have so far received limited attention, the objective of this work is therefore to provide a comprehensive overview of the existing literature on the detection and mitigation of gender bias in AI and ML applications. A distinct focus is put on debiasing techniques that are considered to be related to gender outside the binary framework and potential trade-offs of such techniques, across different and less-researched data types. The results aim to provide guidance and orientation for future development and extension approaches of bias-detection and mitigation techniques in generative AI and ML applications. Thus, this work answers the following research questions:

*RQ1: What is the current state of research on gender bias in artificial intelligence applications?*

*RQ2: How are literature-based debiasing techniques addressing non-binary gender bias?*

*RQ3: How does the current research landscape address gender bias across different data types and trade-offs in mitigation techniques of AI applications?*

This work is structured as follows: The second section provides background information and related work on the relationship between AI and gender bias, as well as on bias detection and debiasing methods. Section three then introduces the methodology, a systematic literature review, on the subject. In section four, the results of the literature review are presented, and research gaps are identified. This work concludes by stating its contributions, limitations and opportunities for future work.

## **Scientific Background**

### ***Artificial Intelligence and Gender Bias***

As the literature on AI is expanding, so is the number of scientific disciplines touching upon AI-related issues, such as gender bias. In this work, we follow Goodfellow et al. (2016) defining ML as the capability of AI systems to acquire their own knowledge autonomously by identifying patterns from raw data, as opposed to relying on pre-programmed, hard-coded knowledge. While such interdisciplinary research enriches our understanding of AI's technological, ethical, and normative implications, it also introduces complexities due to the lack of a common taxonomy. Issues involve inconsistencies and contradictions in the terminology of bias, even within the ML community (Hellström et al., 2020). Recognizing these

challenges, we aim to clarify terminology and advocate for frameworks that foster consensus and consistency within the scientific community.

Bias in the context of AI describes systematic and non-random errors in the data, algorithms, or the interpretation of results made by ML models that lead to unfair outcomes, such as privileging one arbitrary group of users over others (Friedman & Nissenbaum, 1996). Biases can enter the ML process in three different stages, the pre-, in-, and post-processing — each presenting unique challenges and mitigation opportunities (Shrestha & Das, 2022). Regarding the pre-processing stage, the main issue arises, apart from statistical leaning towards certain groups, through unrepresentative data. This occurs when the input data is already inherently biased, e.g. due to outdated views; or through data collection based on traditional role representation (Datta et al., 2018). The source of biases in the in-processing stage, involving the planning, coding and implementation, could be biased developers or the insufficient diversity in development teams resulting from a misrepresentation in hiring and the STEM field in general. Amplifying the problem, the share of women working in the Tech industry is low and on a downward trend (Blumberg et al., 2023). The last potential entry point for biases lies in the post-processing where the ML-algorithm itself can produce more biased data and amplify existing biases. This could lead to a so-called feedback loop (Kelleher et al., 2015).

Clarifying the concept and impact of gender bias in AI and ML is crucial, as existing definitions often lack precision (Blodgett et al., 2020; Devinney et al., 2022). At its core, computer technology operates on binary code, contrasting with the concept of social gender, which is non-binary and fluid, and not strictly tied to biological sex (Larson, 2017). Biases, while sometimes serving useful cognitive functions, can lead to discrimination, especially when entwined with gender preferences (Zimmer & Fahrenberg, 2014; Sun et al., 2019). Broussard (2023) points out that while binary coding is efficient for programming, it struggles to accommodate non-binary gender identities, leading to gender bias in technologies like automatic gender recognition (AGR). This discrepancy highlights a form of gender bias inherent in technology, where the bias does not stem from a developer's ill intent but the binary limitations of machine code in representing fluid social constructs like gender. It is worth noting however that bias in an ethical sense is neither positive nor negative. For instance, female patients exhibit a higher correlation with breast cancer compared to male patients. In this case, statistical biases reflect genuine prevalence differences between genders (Ullmann, 2022). However, the ethical significance of gender bias varies across different AI applications, suggesting a nuanced approach to normative definitions. Further, Keyes (2018) argues for careful consideration of gender-based decision-making in ML development, emphasizing the need to distinguish between sex and gender. Overcoming these biases towards inclusivity requires technological innovation and input from the LGBTQ+ community to ensure AI and ML applications reflect the diversity of gender identities.

With growing awareness, also the research on AI, and gender bias evaluation and mitigation methods continuously expands. There is a wide range of literature reviews (see e.g. Shrestha & Das, 2022, Sun et al., 2019; Orphanou et al., 2023) and ethical evaluations and frameworks where authors provide diverse perspectives and solution approaches with regard to gender bias (see e.g. Leavy, 2018; Hagendorff, 2020; Lütz, 2022; Blodgett et al., 2020). Masiero and Aaltonen (2020) for instance conduct a literature review specifically in the Information Systems (IS) domain, focusing on gender as a variable and gender imbalance in the IT industry. Nadeem et al. (2020) use Google Scholar exclusively as the source of their review and focus on factors that contribute and mitigate gender bias in AI. The literature becomes narrower the smaller the group of affected individuals. Literature reviews specifically focusing on the proposed definition of gender are provided by Fabbrizzi et al. (2021) and Cao & Daumé III (2020). Subramanian et al. (2021) examine whether different debiasing methods can cancel each other out and ways to prevent this. This literature review therefore differentiates itself from prior work, in its focus, on differing data types and the non-binary spectrum, as well as the comprehensive and systematic search across multiple academic domains.

### ***Bias Detection and Mitigation***

In the interest of the identification and mitigation of gender-bias, different techniques and strategies for debiasing have emerged. Subsequently, we review those methods starting with identifying potential biases and then exploring methods for their mitigation.

*Bias detection* is critical, with strategies differing by system and application. In the context of AI within our paper, natural language processing (NLP) models, encompassing key areas like machine translation (MT), sentence completion, sentiment analysis, and more, play a crucial role for extracting information from human language (Sun et al., 2019). Word embeddings, which are representations of words in vector spaces where smaller distance represents a higher semantic similarity often reflect societal biases present in the data they are trained on. For example, certain occupations or roles may be more closely associated with one gender than another in the training data, leading to biased representations in the embedding space. Underlying biases are revealed when gender-associated words (“doctor” or “nurse”) align too closely with specific pronouns (“he” or “she”) (Bolukbasi et al., 2016). In human-computer interaction (HCI), particularly within user-centered design, incorporating user feedback and evaluation methods such as user testing and participatory AI during development stages aids in detecting biases during the development or testing stage (Birhane et al., 2022).

The next logical step is the application of *bias mitigation* techniques. Data augmentation, which involves expanding or adjusting the training data, is a primary method used in the pre-processing stage to balance datasets or annotate data for clearer algorithm interpretation, known as gender tagging (Cao & Daumé III, 2020). Legal frameworks, such as the European Commission's Artificial Intelligence Act (2021), offer in-processing mitigation by setting legally binding guidelines for AI and ML regulation, emphasizing gender equality (Lütz, 2022). In contrast to the ethical guidelines of the High-Level Expert Group on AI (2019) the Artificial Intelligence Act (European Commission, 2021) when coming into force would be legally binding eliminating possible moral hazard problems. Having stated that, the importance of ethical frameworks and guidelines like the expert groups' should not be underestimated, as they lay the groundwork for further research and discussion on responsible AI developments. Ethical guidelines also serve educational purposes, another key aspect of countermeasures (Marassi, 2023). Technological countermeasures include adversarial training, where two separate algorithmic components, the predictor and the adversary follow opposing objectives of generating output correctly and modeling a protected attribute, which should not be possible in a “good” model (Zhang et al., 2018). The overall goal is to maximize the algorithm's performance while minimizing the adversary's prediction (Mehrabi et al., 2021; Verma & Rubin, 2018). Additionally, in the post-processing stage like bias fine-tuning is used to adjust the weights of a trained model before using it for a new dataset or task (Parraga et al., 2022). Debiasing methods, such as hard and soft debiasing, address biases in word embeddings by modifying the representation of gender-neutral and gendered words (Bolukbasi et al., 2016). Iterative nullspace projection (INLP), combining the merits of hard debiasing (Bolukbasi et al., 2016) and adversarial approaches (Chouchane et al., 2023; Zhang et al., 2018), further refines this by identifying and removing bias-related subspaces in data (Ravfogel et al., 2020). The last option we present, fairness gerrymandering aims at improving overall fairness across groups, acknowledging that trade-offs may be necessary against the inconvenience for a smaller part of the group (Kearns & Roth, 2020). These approaches collectively contribute to reducing gender bias in AI systems.

## Methodology

We conduct a systematic literature review following the guidelines and best practices provided by vom Brocke et al. (2009). The taxonomy of the systematic literature review, as illustrated in Table 1, is based on Cooper (1988). The main focus lies on research outcomes and their applications, namely debiasing methods, in theory and practice. While integration is the primary goal, criticism in the form of ethical evaluation is also an essential motivation of this work. Since studies are categorized according to their research outcomes, e.g., bias detection and/or mitigation techniques, a conceptual organization form is used. Given the ethical considerations in the results section of this work, we take on an espousal of position perspective. This work aims to provide support and guidance for researchers and implementers of the discussed methods. The coverage approach is representative and pivotal, meaning a sample of relevant papers is analyzed, with particularly influential and groundbreaking works forming the basis for this review. As this work focuses on existing literature in the informatics and information systems (IS) domain, it is not exhaustive but offers a representative sample of the respective domains.

We follow the PRISMA 2020 checklist (Page et al., 2021) to report on our literature identification process. As information sources, the databases consulted to identify relevant studies were AIS eLibrary, Web of Science, ACM Portal, ACL Anthology, IEEE Xplore Digital Library, PubMed, Google Scholar and Springer Link. In order to facilitate the inclusion and exclusion of studies, Google Scholar was searched repeatedly

using the Publish or Perish software (Harzing, 2007). The term “gender bias” was used in every search query and served as the basis for the keyword-based search process and as a thematic prerequisite for the inclusion of a study. Further search inputs contained the following keywords: “gender bias in artificial intelligence”; “gender bias in machine learning”; “mitigating gender bias”; “detecting gender bias”; “gender bias countermeasures”; “debiasing techniques”; “non-binary gender bias”; “debias trade-offs”; “binary vs non-binary debiasing”. Logical operators like “AND” and “OR” were used to express various combinations of the mentioned search terms. This review comprises secondary sources in order to include studies that had a broad impact on the field of study and to ensure the provided definitions represent interdisciplinary consensus. We include papers from January 2018 to January 2024. The search itself was performed in January 2024.

Findings were excluded based on the following criteria: First, articles with zero citations were excluded, as those are likely to lack relevance or quality. Second, articles with no specified source or publisher as well as non-peer-reviewed articles were excluded due to a perceived lack of academic rigor. Third, articles with entirely different scope than AI or gender bias were excluded. Lastly, duplicates and publications before 2018 are excluded. An exception was made for Bolukbasi et al. (2016), which was included due to its significant impact in debiasing word embeddings). We report the identification of studies following the PRISMA 2020 flow diagram (Page et al., 2021) in Figure 1.

Table 1. Taxonomy of the Literature Review with focus areas in grey, following Cooper (1988)				
Characteristics	Categories			
Focus	research outcomes	research methods	theories	applications
Goal	integration	criticism		central issues
Organization	historical	conceptual		methodological
Perspective	neutral representation		espousal of position	
Audience	Specialized scholars	general scholars	practitioners/politicians	general public
Coverage	exhaustive	exhaustive & selective	representative	central/pivotal

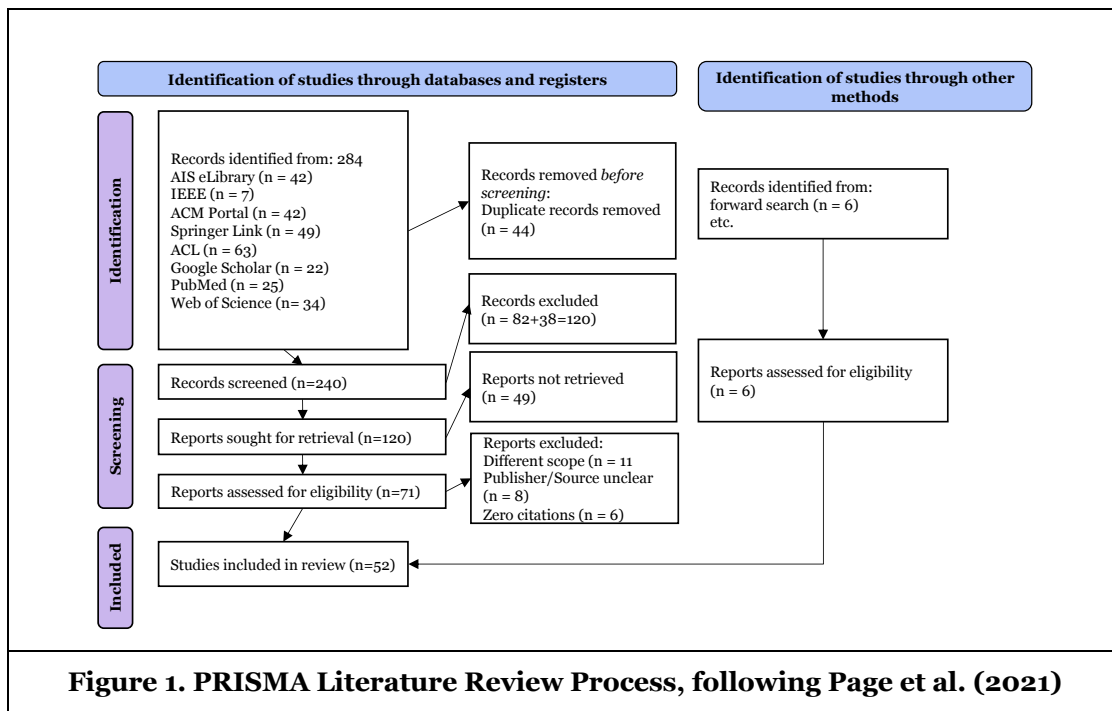


Figure 1. PRISMA Literature Review Process, following Page et al. (2021)

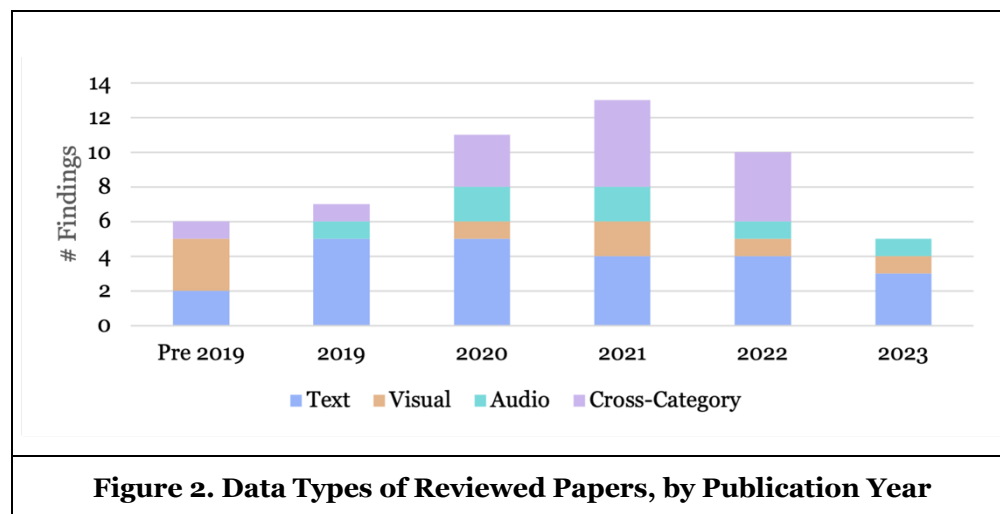
With respect to reporting biases, we acknowledge the potential presence of biases in both the identified literature, as well as for the performed literature review itself. The stated goal of this review is to integrate but also criticize current research on the topic, particularly for potential lack of research on non-binary gender. Moreover, we knowingly did not take a neutral perspective, following the Cooper (1988) taxonomy for literature reviews. These factors are included in the limitations section of this work.

A total of 52 studies emerged as final hits that are included in this work. Studies were examined and then categorized according to the type of data, audio, visual, text or cross-category, that is being processed and considered in the study. If a paper's focus lies in a cross-categorical application, e.g. algorithmic debiasing methods that can be used on several data-types, it falls in the "cross-categories" category. We hereby follow the concept-centric approach by Webster and Watson (2002).

## Results

Table 2 provides an overview on the occurrence of the derived categories *bias detection*, *mitigation*, *non-binary*, *trade-offs* as well as *ethics & fairness*, and presents their distribution among the relevant literature. The results are grouped by the data-types *audio*, *text*, *visual* and *cross-category*. If a paper touches on a category or only mentions or discusses it superficially, we indicate this with an "x". If the main focus of the article falls into one category, this is marked with an "X". A paper can focus on multiple categories, which are then all marked with a capital "X". Among the identified articles, 23 discuss text data, 8 visual data, 7 audio data and 14 cross-category data. Within the categories, most research was carried out with regard to bias detection and mitigation. While non-binary gender bias is investigated for all data categories, we identify a significant research gap for audio data. The same is apparent for categories *trade-offs* and *ethics & fairness*. A reason for this might be the accessibility of text and image data compared to audio data. Nevertheless, the 8 articles on visual data cover more categories than the 7 articles on audio data.

Figure 2 depicts the increase, and subsequent decrease of research on gender biases over time. In all years, biases in text data made up the highest number of publications, with an equal number of articles on cross-category data in 2021. The emergence of new applications such as ChatGPT for text creation and Midjourney or Stable Diffusion for image creation tasks could lead to more research on the biases for the respective data types in the future.



**Table 2. Classification of Literature by Data Type and Focus**

		Detection	Mitigation	Non-Binary	Trade-Offs	Ethics & Fairness
Text	Leavy et al. (2020)	<b>X</b>	<b>X</b>			x
	Cao and Daumé (2020)	<b>X</b>	x	x		x
	Leavy (2019)	<b>X</b>				
	Fleisig and Fellbaum (2022)	x	<b>X</b>			
	Haris et al. (2023)	<b>X</b>				
	Prates et al. (2019)	<b>X</b>	x			x
	Caton and Haas (2020)	x	x	x	x	<b>X</b>
	Lu et al. (2020)	<b>X</b>	x			
	Vig et al. (2020)	<b>X</b>	x	x		
	Savoldi et al. (2021)	<b>X</b>	<b>X</b>	x	x	
	Manzini et al. (2019)	x	<b>X</b>	x	x	x
	Stanovsky et al. (2019)	<b>X</b>				
	Subramanian et al. (2021)		<b>X</b>		<b>X</b>	x
	Kaneko et al. (2022)	x	<b>X</b>		<b>X</b>	x
	Dev et al. (2021)		x	<b>X</b>	x	x
	Leavy (2018)	<b>X</b>	x			
	Bolukbasi et al. (2016)		<b>X</b>			
	Nozza et al. (2022)	<b>X</b>		x		x
	Felkner et al. (2023)	<b>X</b>	x	x	x	
	Ovalle et al. (2023)	<b>X</b>		x		x
Kaneko and Bollegala		<b>X</b>		x		
Görizt et al. (2022)	<b>X</b>	x				
Jussupow et al. (2021)	<b>X</b>	x				
Visual	Atay et al. (2021)	<b>X</b>	<b>X</b>			
	Kafkalias et al. (2022)	<b>X</b>				
	Wang, Qinami et al. (2020)		<b>X</b>	x		
	Wang, Zhao et al. (2018)		<b>X</b>			
	Aka et al. (2021)	<b>X</b>				
	Zhao, Andrews et al. (2023)	x	<b>X</b>	x	x	x
	Keyes (2018)	x	x	<b>X</b>		<b>X</b>
	Buolamwini and Gebru (2018)	<b>X</b>	x		x	x
Audio	Ngueajio and W. (2022)	<b>X</b>	<b>X</b>			x
	Chouchane et al. (2023)		<b>X</b>	x		
	Yasmin et al. (2021)		<b>X</b>	x		
	Bailey and Plumbley (2020)	x	<b>X</b>			x
	Gorrostieta et al. (2019)		<b>X</b>			x
	Costa-jussà et al. (2020)	<b>X</b>	x			
	Garnerin et al. (2021)		<b>X</b>			
Cross-Category	Mehrabi et al. (2021)	x	x			<b>X</b>
	Cirillo et al. (2020)	x	<b>X</b>	<b>X</b>		x
	Fletcher et al. (2021)	x	x			<b>X</b>
	Orphanou et al. (2021)	x	<b>X</b>			x
	Feldman and Peake (2021)	x	<b>X</b>			x
	Sun, Gaut et al. (2019)	x	<b>X</b>	x		
	Blodgett et al. (2020)	<b>X</b>	<b>X</b>	x		
	Albert and Delano (2022)		x	<b>X</b>		
	Guitierrez (2021)	x	<b>X</b>	x		x
	Yang et al. (2020)		<b>X</b>		x	x
	Alabdulmohsin et al. (2022)		<b>X</b>	x	x	x
	Hamidi et al. (2018)		<b>X</b>	x		x
	Nadeem et al. (2020)	x	<b>X</b>		x	x
Masiero and Aaltonen (2020)	x	x			x	
<b>Σ: total count/central focus</b>	<b>38/22</b>	<b>45/ 28</b>	<b>21/4</b>	<b>13/2</b>	<b>27/4</b>	

## Bias Detection

Bias detection methods were the second-most researched category in the reviewed literature. In particular, for text data, bias detection is found to be at the center of the analyzed research articles.

A comprehensive overview and framework of sources, concepts, and bias-effects in MT is provided by Savoldi et al. (2021). Other studies focus on the assessment of gender bias in MT by providing evidence of biased translation tools, such as Google Translate (Prates et al., 2020). By translating sentences containing occupational aspects of gender-neutral languages into English, strong tendencies towards stereotype replication are demonstrated. Through subsequent comparison of the outcomes with real-world female participation in the occupational fields, e.g. STEM, education and corporate, inaccurate depictions of the real world are demonstrated. Google Translate is not the only application in MT that contains and reproduces gender bias. Stanovsky et al. (2019) test five additional MT models and prove that at the time of testing, all models inhibited affinities for stereotypical translations concerning gender, e.g. occupations, descriptions of looks and societal roles. Table 3 depicts a variety of datasets that are used to study gender bias in different applications.

<b>Dataset</b>	<b>Application</b>	<b>Reference</b>
BiasBios	Occupation classification	Kaneko, Bollegala et al. (2022)
WinoQueer	Large Language Models	Felkner et al. (2023)
	Machine Translation	Rudinger et al. (2018), Stanovsky et al. (2019)
WinoMT	Machine Translation	Stanovsky et al. (2019)
WinoST	Machine Translation	Costa-jussa et al. (2022)
Equit Evaluation Corpus	Sentiment Analysis	Kiritchenko and M. (2018), Sun et al. (2019)
Librispeech	Speech Recognition	Panayotov et al. (2015), Ngueajio and Washington (2022)
COCO	Bias Amplification	Zhao et al. (2022)
CelebA	Bias Amplification	Zhao et al. (2022)
imSitu	Bias Amplification	Zhao et al. (2022)

With regard to NLP models, Leavy (2018; 2019; 2020) emphasized the importance of diversity and gender theory in AI. A description of gender-bias detection through text analytics is outlined by Leavy (2018). Different detection methods are presented, ranging from straightforward approaches like counting mentions of women and/or men in texts (gender is treated as a binary variable) to more complex ones. This includes identifying biased metaphors and stereotypes, which requires linguistic proficiency in NLP-related tasks, like word embeddings or coreference resolution annotation, as discussed by (Cao & Daumé III, 2020) and (Lu et al., 2020).

A widely used method for analyzing textual data is the creation and analysis of corpora. Examples of such corpora are newspaper coverage of politicians (Leavy, 2019), comparisons 19th-century texts and modern-day newspaper articles (Leavy et al., 2020) or blockbuster-movie scripts (Haris et al., 2023). Besides the detection and downstream debiasing, text analysis additionally provides information about the performance of the algorithms used to scan the data (Leavy 2019). Lu et al. (2020) provide a benchmark to measure and mitigate gender bias in several NLP tasks, by breaking associations between gendered and gender-neutral words through augmentation training in datasets. This counterfactual data augmentation (CDA) process is described as creating datasets consisting of matched pairs that differ in only the targeted concept, i.e. gender (Lu et al., 2020). This way, biases can be measured based on the difference in output of different instances in the same matched pair. CDA is then used to create a dataset consisting of matched pairs, where the pairs have been switched. The datasets generated through this process are then used to train coreference resolution and language models and reveal inherent gender biases.

Regarding the process of data augmentation in coreference resolution, Cao & Daumé III (2020) developed two new datasets to measure human-introduced gender bias in crowdwork-sourced annotation and the

performance of several coreference resolution systems. Per their account, human annotation can incorporate gender bias into training data due to a mix of potentially vague annotation guidelines and the unique perspectives of the annotators themselves. Based on the detection of gender-based stereotypes in human annotation, careful consideration of the expertise and personality of crowd workers is emphasized. Mediation analysis, as shown by Vig et al. (2020), is an analysis approach that allows the interpretation of the flow of information through different internal model components of large language models (LLMs), making it possible to narrow down the components that contain and replicate bias. Göritz et al. (2022) developed a gender language analyzer for textbook data. Their solution is able to automatically generate improvement suggestions, linking the article to the bias mitigation section. Lastly, further research uses experimental setups to study user evaluations of AI systems. The results demonstrate that biases in AI could reinforce social inequality if users already possess stereotypes (Jussupow et al. 2021).

Buolamwini & Gebru (2018) show that there are significant disparities in the accuracy of facial recognition (FR) technology regarding skin color and gender. Commercial FR systems perform less accurately the darker the skin tone, and if the person is female. Recent studies reveal that image-generating models, like the ones mentioned, show a clear gender bias concerning the representation of women in male-dominated occupations. Atay et al. (2021) use ML algorithms to evaluate gender bias in FR models by training the models on images of 24 individuals and comparing the accuracy rates of the tested algorithms. Their findings serve as a good example of the difference between mathematical- and social fairness: While a higher rate of female subjects in the training data increases the accuracy of some of the tested algorithms, a reduction of male subjects would not negatively affect accuracy (Atay et al., 2021). This illustrates how conceptualizations of fairness based on pure logic can fail; e.g. reweighting the training data to a 50/50 split between groups would not result in real-life fairness in this case. As mentioned, defining biased ground truth labels by annotation can be critical in the development process. An approach to finding human-induced bias is to supervise the annotation processes, e.g. through crowdworking under experimental conditions, as conducted by Cao & Daumé III (2020) and Kafkalias et al. (2022). While the former study uses this approach to verify the critical role of annotators, the latter additionally aims to draw conclusions about the annotator's demographic and personal status (Kafkalias et al., 2022). Aka et al. (2021) demonstrate a solution that does not rely on ground truth labels by using the normalized pointwise mutual information (nPMI) metric to implement an open-sourced bias detection tool.

With focus on Automatic Speech Recognition (ASR) systems, Ngueajio & Washington (2022) discuss a broader range of impacted traits besides gender, like race or disabilities. A common metric to evaluate the performance of ASR systems is measuring the difference between ground truth-labeled test transcripts and the actual predicted transcripts. Further mentioned performance measurement metrics for ASR systems are also used for bias evaluation in speech translation by Costa-jussa et al. (2020).

## **Bias Mitigation**

Gender bias in text-based AI and ML applications has long surpassed the point at which it has been a niche research topic. With ongoing expansion of AI into society, more people experience the negative effects that can arise through biased applications and their societal effects. Accordingly, bias mitigating is identified to be the most researched category in this work. Blodgett et al.'s (2020) approach satisfies the need for mitigation strategies by providing recommendations on how biases can be conceptualized and standardized in order to analyze and mitigate them interdisciplinarily. The authors criticize studies in the field of bias mitigation in NLP systems for shortcomings, such as vaguely stated motivations. Despite the criticism, they propose a debiasing method for pre-trained word embeddings, which preserves gender-related information, called wanted bias. This method includes categorizing the information into four groups: feminine, masculine, gender-neutral, and stereotypical. Only words falling into the stereotypical set, identified as potentially hurtful, are removed. Subsequently, an encoder is employed to transform the initial, biased word embeddings into debiased embeddings. These debiased embeddings are then used to train a decoder, which learns to reconstruct the original word embeddings.

Bolukbasi et al. (2016) introduced the hard- and soft-debias approach, which was later extended by Manzini et al. (2019), with their approach offering multiclass debiasing. Additional methods outside the binary-attribute framework are evaluated by Subramanian et al. (2021) who focus on fairness gerrymandering of groups. Fairness gerrymandering describes an increase of bias against a subgroup while the overall fairness

of a model is increased. Sun et al. (2019) extensively summarize strategies for mitigating bias in textual data and word embeddings and the adjustment of algorithms. The authors identify three approaches to mitigate bias in training datasets: data augmentation, gender tagging, and bias fine-tuning. The work of Kaneko et al. (2022) differentiates between intrinsic and extrinsic measures in order to evaluate correlations between them. Intrinsic measures solely reveal bias contained in masked language models (MLMs) without following its development in downstream tasks, such as NLP systems. Since the authors find weak correlations between both measures, they strongly emphasize not relying on intrinsic bias measurements to determine if an MLM can be applied in a downstream task. A more promising solution can be to combine intrinsic and extrinsic bias measurements to avoid debiasing MLMs, only to let them learn social biases later in the application process.

Further navigation through bias-mitigation techniques leads to visual data applications and their debiasing. Balancing datasets, e.g. by under- or overfitting, can be an effective debiasing method for the training data, as discussed in Atay et al. (2021) and Wang et al. (2018). If, however, balanced datasets are trained on bias-amplifying models, gender bias can be reintroduced. A similar disparity between debiasing effects in different stages of the development process has been discussed in relation to NLP models (Kaneko et al., 2022). Wang et al. (2018) base their argumentation about the insufficiencies of rebalancing datasets as a debiasing method on difficulties in the transfer and preservation between development stages in visual recognition tasks. Using adversarial debiasing, the authors manage to remove gender-implying image space, e.g. faces, without losing further information.

Wang et al. (2020) propose a benchmark for bias-mitigation methods in visual recognition tasks. They test various adversarial training approaches using this benchmark and find that the best performance is achieved by a domain-independent classifier, which aims to distinguish identical attributes across different groups. This approach prioritizes fairness through awareness rather than blindness, which characterizes traditional adversarial approaches (Zhao et al., 2022). They address both previously described studies and proposed debiasing approaches (Wang et al., 2018; 2020). The paper's aim is to provide a bias mitigation technique for multi-attribute classification tasks. Similar to Wang et al.'s (2018) findings, balancing datasets is proven to be insufficient in the multi-attribute setting, besides being highly impractical. In contrast to Wang et al.'s (2020) findings, the best mitigation method for multi-attribute bias mitigation is dataset and metric-dependent. The authors find that mitigating single attribute bias can lead to unintentionally increasing multi-attribute bias (Zhao et al., 2022).

For audio data, the majority of identified articles at least touch bias mitigation. In an attempt to improve the recognition of transgender individuals' voices through gender recognition technology, Yasmin et al. (2022) combine human-extracted features extracted using mathematical concepts like rough-set theory and machine-extracted features and find that the resulting recognition system shows promising results. With higher accuracy in gender detection technology, protecting individuals' privacy becomes increasingly important. Even though authors like Keyes (2018) suggest avoiding AGR technology when possible since gender is insufficient as a variable in many applications (think recommendation systems that use gender as a reference point for product recommendations); developing gender-inclusive systems is a way of improving fairness in ML. Still, the fact that individuals who identify outside the binary gender spectrum are being discriminated against, and the need for systems that protect soft biometrical information is high. Chouchane et al. (2023) show how biometrical information can be protected through gender concealment, which is performed by an adversarial auto-encoder (AAE) that encodes gender-revealing audio data in order to prevent gender classification algorithms from picking up on gender-related information. By introducing noise via a Laplace mechanism, the amount of revealed gender information can be adjusted to users' needs by tuning the noise insertion up or down (Chouchane et al., 2023). Audio data inhibits a variety of information about its source, such as indications about emotional states and even mental health problems (Bailey & Plumbley, 2020; Gorrostieta et al., 2019). Rebalancing the audio data and, instead of processing it, using raw audio as input can mitigate gender bias in depression detection (Bailey & Plumbley, 2020). This can prevent serious consequences that wrong assessments of an individual's mental state can bring along. A similar result as Atay et al. (2021) found for visual data is shown for speech recognition training data. The (re)-allocation of gender can positively impact the performance of ASR applications, but only concerning female voices when the proportion of female voices in the training corpus is increased (Garnerin et al., 2021).

## Non-Binary

Awareness of possible entry points for biases into the AI development process is essential for their mitigation. Although many illustrations of bias entry exist, e.g. feedback loops, few studies have endeavored to identify non-binary bias entry points (Cao & Daumé III, 2020). Of the 150 papers analyzed by the authors, concerning NLP systems that mention “gender”, 92,8% assume binary social gender. In comparison, the papers reviewed in this study show an increase in considering social gender to be non-binary; out of 48 reviewed studies, 20 recognized this definition. The fact that even more papers recognize this definition but only refer to it and continue to make binary assumptions illustrates an incapacity for action.

Further challenges arise from non-inclusive language. By questioning non-binary and AI-familiar individuals, Dev et al. (2021) found a wide range of harms the participants experience, through named entity recognition-, coreference resolution- and MT-systems. Exemplary harms include misgendering, exclusion of applicants’ resumes because the systems do not recognize non-binary names or offensive translations in MT, using terms like “ladyboy”. Accordingly, Keyes (2018) names automatic gender recognition (AGR) systems “Misgendering Machines”. The analysis of 58 studies from 1995 to 2017 about, or including AGR technology shows that 94.8% of papers treated gender as binary (92.9% in gender-focused, 96.7% in non-gender-focused papers). They provide design recommendations for AGR research, emphasizing the need for more expertise and gender awareness amongst human-computer interaction researchers and developers.

For the application of healthcare, Cirillo et al. (2020) find that also the (socio-cultural) gender of robots has to be considered since it has a strong impact on patient behavior and interaction. They find that the awareness of (biological) sex of patients and differences in gender of patients and robots can positively impact the quality of healthcare and biomedical application and improve ethical decision-making of AI. The existence of genderless and gender explicit robots has opened an ongoing debate on similarities between human and robot genders. While an alignment of the different approaches towards gender bias mitigation, outlined in this work, is to be hoped for in the long run, universal standards and broader training on inequality and gender-related topics are recommended. This could be realized with on-the-job training (especially in the field of ML-development).

## Ethics, Fairness and Trade-Offs

Since gender bias mitigation aims to increase fairness in the long run, most reviewed papers address ethics or fairness in some way. The decision to include a study in this subgroup or not was based on the way fairness was approached. For instance, Caton & Haas (2020), Mehrabi et al. (2021) and Fletcher et al. (2020) are included due to wide-ranging and detailed insights into fairness, while Keyes’ (2018) paper addresses a lack of fairness and aims to implement ethical considerations into the discussed field of AGR. Caton & Haas (2020) collect notable projects in ML fairness, such as debiasing tools and fairness measurement metrics. They also discuss the fairness dilemma of balancing fairness and accuracy and different fairness notions, such as individual and group fairness. Fletcher et al. (2020) discuss ethical implications of binary diagnose classification through health data. Here, depending on the tested disease and the risk of stigma or the prevention of spreading, it can be more beneficial to either minimize false positive or false negative results. In such scenarios, ethical considerations need to be supported by law.

The literature becomes even sparser when potential problems between different debiasing approaches due to inconsistency in definitions of gender and handling of gender bias are considered. Subramanian et al. (2021) argue that fair models need to account for intersectional bias, which can arise from grouping sensitive attributes in independent groups, ignoring interdependencies between the groups and their subgroups. They find that INLP is the most realistic and promising in the presence of intersectional bias. Further papers address problems in multi-attribitional debiasing without having their primary focus on trade-offs between groups (Alabdulmohsin et al., 2022; Buolamwini & Gebru, 2018; Felkner et al., 2023; Kaneko et al., 2022; Manzini et al., 2019; Yang et al., 2020). In the audio-data subcategory, no studies mention possible trade-offs between different debiasing approaches. Alabdulmohsin et al. (2022) suggest a debiasing method for multiclass datasets that is based on a reduction-to-binary approach (R2B). This is approached by extending demographic parity to the multiclass setting. In this case, the different properties of binary- and non-binary classification algorithms are used to achieve an overall increase in fairness.

## Discussion and Identification of Research Gaps

When thinking about ethical frameworks for technologies that have such a drastic impact on our society as AI does, one will soon find that it is not enough to write guidelines, such as the ones introduced in an earlier section, and hope for compliance solely based on morals. The absence of consequences can lead to a lack of incentives to follow ethical guidelines, which can result in the failure of many initiatives (Hagendorff, 2020). Research on gender bias and other fairness-related topics in AI is therefore to be seen as a meaningful step in the right direction. Ignoring clear and recognized gender definitions provided by medicine, social studies, and governmental institutions, though, is obviously no reasonable approach. Keeping in mind that most bias in AI is not intentionally but reflects societal conditions instead, the recent boom of AI is a chance to identify and effectively work against biases on all levels. Through the analysis of the current base of scientific literature on the subject, we elicit the following research trends and gaps:

1) *Sector specific regulations*: The exploration of more technical or sector-specific regulations, including specific mandates to uphold gender equality norms in addressing algorithmic discrimination (Lütz, 2022).

2) *Definition of gender and sex*: Clear definitions and a general understanding of the differences between gender and sex are essential for successful bias mitigation across all affected gender groups and sub-groups in AI research and the development of MI algorithms.

3) *Fairness trade-offs*: The same holds for bias itself, while the necessity of a clear and precise definition and framework is discussed in section one, fairness gerrymandering raises ethical questions about which groups to focus on in debiasing approaches and how to justify any preferences made (Ferrara, 2023). In terms of algorithmic fairness, fundamentally incompatible approaches are no exception but rather inevitable (Friedler et al., 2021). Depending on the use case and desired outcomes, seemingly contradictory notions of fairness are necessary to ensure that fairness can be aimed for under variable conditions, e.g., diverse worldviews, cultures, or genders.

4) *Overcoming binary frameworks*: Representative bias mitigation requires all potentially harmed individuals to be acknowledged, at least, to be protected ideally. Looking at the distribution of papers about debiasing methods in the gender bias context, it is striking not only how many studies make no mention of genders besides the binary but even more how many do acknowledge the social and fluid nature of gender but still focus on technological solutions that work solely in the binary framework. A knowledge-based approach is vital on a technological level, regarding the understanding of algorithms, their applications and potential risks, as well as on an ethical-social level, with focus on inequality, discrimination and bias. However, this goes beyond algorithmic information theory. A broader awareness within the society needs to be created in order to generate unbiased and fair datasets.

5) *Inclusivity in ML design*: An often-missing core value that needs to be implemented further in the whole ML development process is diversity. Like bias, more diversity would influence the whole circle by creating non-discriminatory data. The feedback loop could, therefore, be used to debias the whole process. Also involving disadvantaged or discriminated individuals during auditing can generate technologies that mitigate contradictions of debiasing methods because the systems are trained on a more realistic and inclusive representation of the real world.

6) *Participatory standards*: Participatory AI, as a way to create fair ML models, combines several necessary preconditions for bias mitigation by the involvement of affected groups. But as described by Birhane et al. (2022) there is no consensus on a minimum set of standards to be used for evaluation of participatory mechanisms. Involving end users in auditing or development processes implements real-time detection of gender bias in algorithms and data, and thus addresses discrimination in real-world scenarios.

7) *Data type coverage*: While we found many studies taking text data into account, research efforts for voice and visual data have to be intensified. This can be accelerated by the provision of large data sets across industries to the research community to better study non-binary bias detection in real world scenarios.

## Limitations and Future Work

This paper is not without limitations. Although we aimed to provide a representative sample of academic literature on gender bias in AI, there might exist further literature that could not be assessed due to time-, and accessibility-constraints. Based on the focus of this work, research in informatics and IS-related databases was considered, with other disciplines not being investigated. Future studies could therefore investigate knowledge created in other disciplines and for technologies other than AI. Further, the use of

inclusive language throughout the study is aspired, but cannot be guaranteed due to its continuous change and improvement. The same goes for inclusion of all groups affected by gender bias in AI. The nature of bias as a psychological concept prevents a guarantee for a bias-free work, especially with taxonomy categories like “criticism” and “espousal of position” being stated, as illustrated in Table 1.

The findings of this paper identified several gaps in academic literature that provide opportunities for future work. We advocate for the development of standardization of term-definitions and consensus about fairness metrics in the domain. Through analyzing the literature with the aim of answering the proposed research questions, an inconsistency throughout disciplines and subfields of research concerning standardization of term-definitions and consensus about fairness metrics, as well as uniform aims of bias mitigation are shown. Besides providing an overview on ongoing research on this topic, this study aims to shed a light on the need for clearer and more productive research through broader frameworks and closer attention to- and involvement of impacted individuals.

## **Conclusion and Contributions**

This work contributes to the body of research on AI, by providing a systematic review of the current state of gender bias in AI applications and ML technologies. We thoroughly assess bias detection and mitigation approaches for various data types, finding that the literature provides a wide range of gender bias detection, and mitigation techniques, answering *RQ1*. The review emphasizes a significant focus on bias detection and mitigation in AI literature. We find that there exist mitigation techniques for all information types with notable gaps in research, particularly in audio data and non-binary gender considerations. Methodological approaches vary by data type, from text analytics and corpus creation in NLP to image analysis in facial recognition and structured approaches in audio data processing.

In answering *RQ2*, we find that there exist debiasing techniques specifically for non-binary gender bias. In comparison to the total number of studies evaluated, the proportion is however rather low. Trade-offs between binary and non-binary debiasing methods could be identified, which could pose serious new challenges for researchers and developers in the future.

Regarding *RQ3*, we pursue a pioneering effort in the granular analysis of gender bias across multiple data types. By meticulously examining bias detection and mitigation strategies within audio, visual, textual data, and considering the underrepresented non-binary gender perspectives, we contribute a novel, multidimensional understanding of gender bias in AI. Our detailed comparison across data types reveals unique biases and challenges inherent to each, particularly the existing trade-off in mitigation techniques. With our work, we contribute to uncovering potential impediments to develop more inclusive AI systems, highlighting the critical importance of nuanced, data-type-specific approaches.

Overall, it can be seen that many researchers from various fields have recognized the issues that lead to and arise with gender bias and are ready to invest time and thought into further solution approaches. Even though a gap between binary- and non-binary referencing literature has to be acknowledged, the trend seems to include a broader spectrum of individuals in the development process of new, and the mitigation process of biased systems. Looking at the distribution of studies between data categories, an imbalance in favor of text data is noted. This literature gap, particularly for gender bias in audio and visual data poses opportunities for further research. The emergence of generative AI that processes and outputs both, visual and audio data, will clearly amplify the need for such research. Similarly, current gaps in academic literature could be identified for fairness gerrymandering for data-types other than text data, as well as greater consideration of trade-offs between individual debiasing methods. This work therefore contributes to the basis of knowledge on gender bias in AI by identifying several multiple opportunities for future research, and eliciting factors that need to be considered when working with AI.

## **Acknowledgments**

The project FIIPS-at-Home was funded by the German Federal Ministry of Education and Research under the funding code 16KISA068K. Responsibility for the contents lies with the authors. The project was funded by the European union -- NextGeneration EU. The views expressed are those of the authors and do not necessarily reflect those of the European Union or the European Commission.

## References

- Aka, O., Burke, K., Bauerle, A., Greer, C., & Mitchell, M. (2021). Measuring Model Biases in the Absence of Ground Truth. *AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*
- Alabdulmohsin, I., Schrouff, J., & Koyejo, O. (2022). A Reduction to Binary Approach for Debiasing Multiclass Datasets. In *Neural Information Processing Systems (NeurIPS)*.
- Albert, K., & Delano, M. (2022). Sex trouble: Sex/gender slippage, sex confusion, and sex obsession in machine learning using electronic health records. *Patterns*, 3(8), 100534.
- Atay, M., Gipson, H., Gwyn, T., & Roy, K. (2021). Evaluation of Gender Bias in Facial Recognition with Traditional Machine Learning Algorithms. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–7). IEEE. <https://doi.org/10.1109/SSCI50451.2021.9660186>
- Bailey, A., & Plumbley, M. D. (2020). *Gender Bias in Depression Detection Using Audio Features*. University of Surrey. <http://arxiv.org/pdf/2010.15120.pdf>
- Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the People? Opportunities and Challenges for Participatory AI. *EAAMO'22: Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, Article No. 6.
- Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. [doi.org/10.18653/v1/2020.acl-main.485](https://doi.org/10.18653/v1/2020.acl-main.485)
- Blumberg, S., Krawina, M., Mäkelä, E., & Soller, H. (2023). *Women in Tech: The Best Bet to Solve Europe's Talent Shortage*. McKinsey.
- Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *30th Conference on Neural Information Processing Systems*.
- Brocke, J., Simons, A., Niehaves, B., Reimer, K., Plattfaut, R., Cleven, A. (2009). RECONSTRUCTING THE GIANT: ON THE IMPORTANCE OF RIGOUR IN DOCUMENTING THE LITERATURE SEARCH PROCESS. (2009). *ECIS 2009 Proceedings*. 161. <https://aisel.aisnet.org/ecis2009/161>
- Broussard, M. (2023). *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*. MIT Press.
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *PMLR: Vol. 18. Conference on fairness, accountability, and transparency*
- Cao, Y. T., & Daumé III, H. (2020). Toward Gender-Inclusive Coreference Resolution. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4568–4595). Association for Computational Linguistics.
- Caton, S., & Haas, C. (2020). Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, Article 3616865. Advance online publication. <https://doi.org/10.1145/3616865>
- Chouchane, O., Panariello, M., Zari, O., Kerenciler, I., Chihaoui, I., Todisco, M., & Önen, M. (2023). Differentially Private Adversarial Auto-Encoder to Protect Gender in Voice Biometrics. In *IH&MMSec '23: Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security (Chair)*.
- Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S., & Mavridis, N. (2020). Sex and Gender Differences and Biases in Artificial Intelligence for Biomedicine and Healthcare. *NPJ Digital Medicine*, 3(81).
- Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society*, 1, Article No. 104, 104–126. <https://doi.org/10.1007/BF03177550>
- Costa-jussa, M., Gonen, H., Hardmeier, C., & Webster, K. (Eds.) (2021). *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics.
- Datta, A., Datta, A., Makagon, J., Mulligan, D. K., & Tschantz, M. C. (2018). Discrimination in Online Advertising: A Multidisciplinary Inquiry. In *PMLR: Vol. 18. Conference on fairness, accountability, and transparency* (pp. 20–34).
- Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J., & Chang, K.-W. (2021). Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1968–1994).
- Devinney, H., Björklund, J., & Björklund, H. (2022). Theories of “Gender” in NLP Bias Research. In *ACM Conference on Fairness 2022* (pp. 2083–2102). <https://doi.org/10.1145/3531146.3534627>

- European Commission. (2021). *Proposal for a Regulation of the European Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM(2021) 206 final)*.
- Fabbrizzi, S., Papadopoulous, S., Ntoutsis, E., & Kompatsiaris, I. (2021). A Survey on Bias in Visual Datasets. *Computer Vision and Image Understanding*, 103552(223).
- Felkner, V., Chang, H.-C. H., Jang, E., & May, J. (2023). WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the Association for Computational Linguistics*
- Ferrara, E. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, 6(1), 3. <https://doi.org/10.3390/sci6010003>
- Fletcher, R. R., Nakeshimana, A., & Olubeko, O. (2020). Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. *Frontiers in Artificial Intelligence*.
- Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (Im)possibility of fairness: Different Value Systems Require Different Mechanisms For Fair Decision Making. *Communications of the ACM*, 64(4), 136–143. <https://doi.org/10.1145/3433949>
- Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on information systems (TOIS)*, 14(3), 330-347.
- Garnerin, M., Rossato, S., & Besacier, L. (2021). Investigating the Impact of Gender Representation in ASR Training Data: a Case Study on Librispeech. In M. Costa-jussa, H. Gonen, C. Hardmeier, & K. Webster (Eds.), *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
- Gorrostieta, C., Lotfian, R., Taylor, K., Brutti, R., & Kane, J. (2019). Gender De-Biasing in Speech Emotion Recognition. In *Proc. Interspeech 2019* (pp. 2823–2827). ISCA.
- Göriz, L., Stattkus, D., Beinke, J., Thomas, O. (2022). To Reduce Bias, You Must Identify It First! Towards Automated Gender Bias Detection. *ICIS 2022 Proceedings*. 10. [https://aisel.aisnet.org/icis2022/data\\_analytics/data\\_analytics/10](https://aisel.aisnet.org/icis2022/data_analytics/data_analytics/10)
- Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1), 99–120. <https://doi.org/10.1007/s11023-020-09517-8>
- Haris, M. J., Upreti, A., Kurtaran, M., Ginter, F., Lafond, S., & Azimi, S. (2023). Identifying gender bias in blockbuster movies through the lens of machine learning. *Humanities and Social Sciences Communications*, 10(10), Article 94. <https://doi.org/10.1057/s41599-023-01576-3>
- Harzing, A. W. (2007). *Publish or Perish* (Version 8) [Computer software]. <https://harzing.com/resources/publish-or-perish>
- Hellström, T., Dignum, V., & Bensch, S. (2020). Bias in Machine Learning--What is it Good for?.
- High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*. European Commission. <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines.1.html>
- Jussupow, E., Meza Martínez, M., Maedche, A., Heinzl, A. (2021). Is This System Biased? – How Users React to Gender Bias in an Explainable AI System. *ICIS 2021 Proceedings*. 11. [https://aisel.aisnet.org/icis2021/hci\\_robot/hci\\_robot/11](https://aisel.aisnet.org/icis2021/hci_robot/hci_robot/11)
- Kafkalias, A., Herodotou, S., Theodosiou, Z., & Lanitis, A. (2022). Bias in Face Image Classification Machine Learning Models: The Impact of Annotator's Gender and Race. *IFIP Advances in Information and Communication Technology, Artificial Intelligence Applications and Innovations: 18th IFIP WG 12.5 International Conference*. Springer.
- Kaneko, M., & Bollegala, D. (2019). Gender-preserving Debiasing for Pre-trained Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Symposium conducted at the meeting of Association for Computational Linguistics.
- Kaneko, M., Bollegala, D., & Okazaki, N. (2022). Debiasing isn't enough! - On the Effectiveness of Debiasing MLMs and their Social Biases in Downstream Tasks. *Proceedings of the 29th International Conference on Computational Linguistics*, 1299–1310. <http://arxiv.org/pdf/2210.02938.pdf>
- Kearns, M., & Roth, A. (2020). *The ethical algorithm. The science of socially aware algorithm design*. Oxford University Press.
- Kelleher, J. D., MacNamee, B., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. MIT Press.
- Keyes, O. (2018). The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1–22.
- Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. arXiv preprint arXiv:1805.04508.

- Larson, B. (2017). Gender as a Variable in Natural-Language Processing: Ethical Considerations. In D. Hovy, S. Spruit, M. Mitchell, E. M. Bender, M. Strube, & H. Wallach (Eds.), *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*.
- Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In E. Abraham (Ed.), *2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering (GE)* (pp. 14–16). IEEE. <https://doi.org/10.1145/3195570.3195580>
- Leavy, S. (2019). Uncovering gender bias in newspaper coverage of Irish politicians using machine learning. *Digital Scholarship in the Humanities*, 34(1), 48–63. <https://doi.org/10.1093/lc/fqy005>
- Leavy, S., Meaney, G., Wade, K., & Greene, D. (2020). Mitigating Gender Bias in Machine Learning Data Sets. In L. Boratto, S. Faralli, M. Marras, & G. Stilo (Eds.), *Communications in Computer and Information Science: Vol. 1245. Bias and Social Aspects in Search and Recommendation: First International Workshop, BIAS 2020* (1st ed., pp. 12–26). Springer.
- Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2020). Gender bias in neural natural language processing. In V. Nigam, T. Ban Kirigin, C. Talcott, J. Guttman, S. Kuznetsov, B. Thau Loo, & M. Okada (Eds.), *Lecture Notes in Computer Science. Logic, Language, and Security*. Springer.
- Lütz, F. (2022). Gender equality and artificial intelligence in Europe. Addressing direct and indirect impacts of algorithms on gender-based discrimination. *ERA Forum*, 23(1), 33–52.
- Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.
- Marassi, L. (2023). Assessing User Perceptions of Bias in Generative AI Models: Promoting Social Awareness for Trustworthy AI. In *Proceedings of the 2023 Conference on Human Centered Artificial Intelligence: Education and Practice* (pp. 46–46).
- Martin, K. (Ed.). (2022). *Ethics of data and analytics: Concepts and cases* (1st ed.). CRC Press.
- Masiero, S., Aaltonen, A. (2022). Gender Bias in Information Systems Research: A Literature Review. *AISWN International Research Workshop on Women, IS and Grand Challenges*.
- Meade, N., Poole-Dayana, E., & Reddy, S. (2021). An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *arXiv preprint arXiv:2110.08527*.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 1–35. [doi.org/10.1145/3457607](https://doi.org/10.1145/3457607)
- Nadeem, A., Abedin, B., Marjanovic, O. (2020). Gender Bias in AI: A Review of Contributing Factors and Mitigating Strategies. *ACIS 2020 Proceedings*. 27. <https://aisel.aisnet.org/acis2020/27>
- Ngueajio, M. K., & Washington, G. (2022). Hey ASR System! Why Aren't You More Inclusive? Automatic Speech Recognition Systems' Bias and Proposed Bias Mitigation Techniques. A Literature Review. *HCI*, 13518, 421–440. <https://doi.org/10.48550/arXiv.2211.09511>
- Orphanou, K., Otterbacher, J., Kleanthous, S., Batsuren, K., Giunchiglia, F., Bogina, V., Tal, A. S., Hartman, A., & Kuflik, T. (2023). Mitigating Bias in Algorithmic Systems—A Fish-eye View. *ACM Computing Surveys*, 55(5), 1–37. <https://doi.org/10.1145/3527152>
- Ovalle, A., Goyal, P., Dhamala, J., Jagers, Z., Chang, K.-W., Galstyan, A., Zemel, R., & Gupta, R. (2023). “I’m fully who I am”: Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. In *ACM Digital Library, Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1246–1266). Association for Computing Machinery.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *International journal of surgery*, 88, 105906.
- Parraga, O., More, M. D., Oliveira, C. M., Gavenski, N. S., Kupssinskü, L. S., Medronha, A., Moura, L. V., Simões, G. S., & Barros, R. C. (2022). Debiasing Methods for Fairer Neural Models in Vision and Language Research: A Survey. *ACM Computing Surveys - Special Issue on Trustworthy AI*, Advanced online publication. <http://arxiv.org/pdf/2211.05617.pdf>
- Prates, M., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, 32(10), 6363–6381.
- Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*
- Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9, 845–874.

- Shrestha, S., & Das, S. (2022). Exploring gender biases in ML and AI academic research through systematic literature review. *Frontiers in Artificial Intelligence*, 5, 976838.
- Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1679–1684).
- Subramanian, S., Han, X., Baldwin, T., Cohn, T., & Frermann, L. (2021). Evaluating Debiasing Techniques for Intersectional Biases. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*
- Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.
- Ullmann, S. (2022). Gender Bias in Machine Translation Systems. In A. Hanemaayer (Ed.), *Social and Cultural Studies of Robots and AI. Artificial Intelligence and Its Discontents: Critiques from the Social Sciences and Humanities* (1st ed. 2022, pp. 123–146). Springer.
- Ulloa, R., Richter, A. C., Makhortykh, M., Urman, A., & Kacperski, C. S. (2022). Representativeness and face-ism: Gender bias in image search. *New Media & Society*, 0(0). DOI: 14614448221100699
- UNESCO. (2020). UNESDOC Digital library, Artificial intelligence and gender equality: key finding of UNESCO's global dialogue. Available at: <https://unesdoc.unesco.org/ark:/48223/pf0000374174>
- Verma, S., & Rubin, J. (2018). Fairness definitions explained. In Y. Brun, B. Johnson, & A. Meliou (Eds.), *2018 ACM/IEEE International Workshop on Software Fairness* (pp. 1–7). IEEE.
- Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). Investigating Gender Bias in Language Models Using Causal Mediation Analysis. *Neural Information Processing Systems*, 33, 12388–12401.
- Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., & Ordonez, V. (2018). Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. *IEEE/CVF International Conference on Computer Vision*, 5309–5318. 10.1109/ICCV.2019.00541
- Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., & Russakovsky, O. (2020). Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Webster, J. and Watson, R.T. (2002). Analyzing the Past to Prepare For the Future: Writing a Literature Review. *MIS Quarterly*, 26 (2).
- Yang, F., Cisse, M., & Koyejo, S. (2020). Fairness with Overlapping Groups. *NIPS-20: Proceedings of the 34th Conference on Neural Information Processing Systems*, 4067–4078.
- Yasmin, G., DAS, A. K., Nayak, J., Vimal, S., & Dutta, S. (2022). A rough set theory and deep learning based predictive system for gender recognition using audio speech. *Soft Computing*, 1–24.
- Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. In J. Furman (Ed.), *ACM Conferences, Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335–340). ACM. <https://doi.org/10.1145/3278721.3278779>
- Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of Association for Computational Linguistics*.
- Zhao, D., Andrews, J. T. A., & Xiang, A. (2022). Men Also Do Laundry: Multi-Attribute Bias Amplification. *Proceedings of the 40th International Conference on Machine Learning*.
- Zhou, M., Abhishek, V., Dardenger, T., Kim, J., & Srinivasan, K. (2024). Bias in Generative AI. *arXiv preprint arXiv:2403.02726*.
- Zimmer, A., & Fahrenberg, J. (2014). Heuristik. In M. A. Wirtz (Ed.), *Dorsch - Lexikon der Psychologie* (18th ed., p. 691). Hogrefe