



Gender Bias in Natural Language Processing and Computer Vision: A Comparative Survey

MARION BARTL, School of Information and Communication Studies, University College Dublin, Dublin, Ireland and Insight Research Ireland Centre for Data Analytics, Dublin, Ireland

ABHISHEK MANDAL, School of Computing, Dublin City University, Dublin, Ireland and Insight Research Ireland Centre for Data Analytics, Dublin, Ireland

SUSAN LEAVY, School of Information and Communication Studies, University College Dublin, Dublin, Ireland and Insight Research Ireland Centre for Data Analytics, Dublin, Ireland

SUZANNE LITTLE, School of Computing, Dublin City University, Dublin, Ireland and Insight Research Ireland Centre for Data Analytics, Dublin, Ireland

Taking an interdisciplinary approach to surveying issues around gender bias in textual and visual AI, we present literature on gender bias detection and mitigation in NLP, CV, as well as combined visual-linguistic models. We identify conceptual parallels between these strands of research as well as how methodologies were adapted cross-disciplinary from NLP to CV. We also find that there is a growing awareness for theoretical frameworks from the social sciences around gender in NLP that could be beneficial for aligning bias analytics in CV with human values and conceptualising gender beyond the binary categories of male/female.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Computer vision**; • **General and reference** → **Surveys and overviews**; • **Social and professional topics** → **Gender**;

Additional Key Words and Phrases: Trustworthy AI, ethical AI, natural language processing, computer vision

ACM Reference Format:

Marion Bartl, Abhishek Mandal, Susan Leavy, and Suzanne Little. 2025. Gender Bias in Natural Language Processing and Computer Vision: A Comparative Survey. *ACM Comput. Surv.* 57, 6, Article 139 (February 2025), 36 pages. <https://doi.org/10.1145/3700438>

1 Introduction

Artificial Intelligence (AI) algorithms replicate and amplify existing gender inequalities in training data leading to the perpetuation of discrimination in society through the systems in which they

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant No. 12/RC/2289_P2.

Authors' Contact Information: Marion Bartl, School of Information and Communication Studies, University College Dublin, Dublin, Ireland and Insight Research Ireland Centre for Data Analytics, Dublin, Ireland; e-mail: marion.bartl@insight-centre.org; Abhishek Mandal, School of Computing, Dublin City University, Dublin, Ireland and Insight Research Ireland Centre for Data Analytics, Dublin, Ireland; e-mail: abhishek.mandal2@mail.dcu.ie; Susan Leavy, School of Information and Communication Studies, University College Dublin, Dublin, Ireland and Insight Research Ireland Centre for Data Analytics, Dublin, Ireland; e-mail: susan.leavy@ucd.ie; Suzanne Little, School of Computing, Dublin City University, Dublin, Ireland and Insight Research Ireland Centre for Data Analytics, Dublin, Ireland; e-mail: susanne.little@dcu.ie.



This work is licensed under a [Creative Commons Attribution International 4.0 License](https://creativecommons.org/licenses/by/4.0/).

© 2025 Copyright held by the owner/author(s).

ACM 0360-0300/2025/02-ART139

<https://doi.org/10.1145/3700438>

are deployed [88]. **Natural Language Processing (NLP)** and **Computer Vision (CV)** models are central AI components within systems such as social media platforms, news aggregators, search engines and many others. Bias within these systems can therefore amplify and perpetuate societal gender biases causing harm to historically disadvantaged genders and impeding efforts toward achieving gender equality [23, 80, 132]. Research in AI, including NLP and CV, is therefore seeking to address gender biases in different models [46, 111].

Gender bias in NLP is evident in examples including word embeddings that produce analogies such as “*man* is to *computer programmer* as *woman* is to *homemaker*” [19], sentiment analysis systems that give higher intensity scores to instances mentioning women [67], and co-reference resolution systems that fail to correctly identify neopronouns (*ze*, *xe*, etc.) and singular *they*, which is used as a gender-indefinite and non-binary pronoun [21, 27]. In CV, common gender biases include facial recognition systems having lower accuracy in detecting female faces [23], annotation models mislabelling gender of people depicted in traditionally gendered roles [140], visual datasets having stereotypical and biased annotations [15, 106, 138] and images depicting traditional gender roles [132]. Techniques for detecting and mitigating such biases focused on training data, intermediate representations, and the trained models themselves with promising results. However, within commercial systems the scale of adoption of methods for mitigating gender bias, such as in the case of Google Translate [70], is unclear. Furthermore, there may be issues regarding the reliability of the measurement of bias reduction [4].

The connection between the domains of language and vision within AI has grown due to automatic labelling of images, multimodal visual-linguistic models such as VL-BERT [123], as well as generative models such as Dall-E [98] that take text as input to generate images. It is therefore important to understand how gender biases may be present in both sides of combined textual and visual models and how such biases may be mitigated.

In this article, we present a literature survey (as defined by Reference [50]) as an exploratory and non-systematic review of research on gender bias detection and mitigation in Natural Language Processing and Computer Vision. This survey includes models that use text and/or images as the primary data sources and applications in co-referencing, detection and classification of concepts relating to gender and personal descriptions (appearance, role, identity). A total of 587 papers was collected and 142 of them are included in this article, primarily chosen based on coverage of NLP and CV topics and focusing on recency with most publications from 2021–2022. Section 2 establishes the background concepts that relate to defining gender, how bias and fairness are defined in machine learning applications and how literature has identified and classified potential sources of bias. The objective is to highlight the methodologies and insights that can be shared across the two disciplines considering contrasting approaches to either addressing bias by identifying it in existing models (bias detection, Section 3) or attempting to resolve it through interventions (bias mitigation, Section 4).

2 Background and Foundational Concepts

In this section, we first present definitions of the concept of gender. We make a distinction between gender and sex, discuss gender as not a natural but rather a performative category, and distinguish between grammatical and social gender categories. Second, we present bias and fairness definitions in machine learning and present possible sources of bias within the machine learning lifecycle.

2.1 Conceptualising Gender

In this work, we differentiate between *sex* and *gender*, understanding *sex* to concern biological characteristics that form categorisations of *male*, *female*, and *intersex* [39]. *Gender* is understood to be a social category that is subject to change and fluctuation and operates on a spectrum. Social

gender pertains to someone's *gender identity* (how they experience their own gender), their *gender expression* (how they perform their gender and what gender roles they occupy), as well as their *perceived gender* (how a person is *gendered* by others) and how this influences their experience and performance of gender [66]. Gender is thus understood not as something that is pre-determined and static, but that comes into being through performance, both by the individual and the societal environment with which they interact [24]. Moreover, it intersects with other aspects of one's identity such as race, socioeconomic background, religion, ability, and nationality [39].

Gender can be performed both through language and visual indicators. In images and visual media, gender performance is through features such as a person's hairstyle, clothes, facial hair or their use of make-up, among others. Similarly, a person can perform their gender through language by introducing themselves with their preferred pronouns [39], or through their choice of words. However, such individual expressions of gender become aggregated and generalised by AI algorithms that learn predominant concepts of gender embedded within given datasets and therefore can be actively involved in "the production of gendered categories" [39].

It is important to not only distinguish between sex and gender, but, specifically when dealing with language, to also distinguish between the social and the linguistic category of gender [27, 33, 56]. Linguistic gender can either refer to the grammatical categorisation into different noun classes, as is the case for grammatical gender languages such as Italian, French or German. English, as a notional gender language [86], does not have grammatical gender, but it has referential gender, which is used to reference the social gender of a person or the sex of an animal [27]. Referential gender can, for example, be expressed through pronouns such as *he*, *she*, or *they*. English also has lexical gender, which refers to the fact that some words such as *boy* or *mother* carry gender information [2]. When talking about gender in a linguistic sense, we will use the words *feminine*, *masculine*, and *neuter*, as per linguistic convention [39].

2.2 Bias and Fairness Definitions in Machine Learning

Before going into detail on definitions of bias and fairness within the literature, we first establish some terminology around how the category of gender is broken down to measure fairness.

The terminology in this article regarding the sensitive attribute of *gender* follows the outline by Czarnowska et al. [35]. Gender is a **sensitive attribute**, that means people should not experience any discrimination based on their gender. The different genders that exist, such as *male*, *female*, *non-binary*, *agender*, or *gender-queer* are called **protected groups**. While the fundamentals remain the same for both NLP and CV, the expressions of gender will differ depending on the medium in question. In language, membership of certain protected groups is expressed or represented through **identity terms**. In visual media, gender identity is either represented explicitly through identity terms in annotations and labels, or implicitly where gender is *learnt* by models from information perceived in the visual medium.

Definitions of bias and fairness within the literature mostly concern difference, pertaining to the different and more favourable treatment of one protected group compared with another. Bias describes the presence of difference while fairness describes the absence of difference. Moreover, as Blodgett et al. [17] and Green [51] point out, with respect to social biases, this difference is not free of value judgements, meaning that it expresses historical and/or current unequal and discriminatory treatment of one protected group over the other. ML models are trained to generalise so they normalise the most common traits and therefore, bias detection and mitigation is necessarily a normative undertaking [17].

For machine learning systems, Mehrabi et al. [88] define fairness as "the absence of any prejudice or favouritism toward an individual or group based on their inherent or acquired characteristics." Following Hutchinson and Mitchell [62] and Chouldechova and Roth [32], Czarnowska et al. [35]

distinguish between two types of fairness: *group fairness* and counterfactual, or *individual fairness*. Group fairness is achieved when performance for the groups in question reaches the same statistical score, such as an F1 measure. Individual fairness is achieved when changing the protected group, does not influence model output [35]. As an example, a facial detection algorithm that works well on an image of a man should work equally well on an image of a woman in the same context. Fairness metrics, used to assess whether a model exhibits differences in performance between protected groups, therefore measure unfairness.

Similar but related terminology concerns bias, which can be linked or equated with unfairness. Work on bias has covered a wide range of applications. However conceptualisations of bias in these works have often been inconsistent or unclear [46, 63]. This observation led Blodgett et al. [17] to take a step back from an all-purpose bias definition and instead advocate for researchers to be clear about conceptualisation of bias in their own research. Blodgett et al. [17] moreover state that researchers should provide information about harms biases may cause, who would be affected by those harms, and why certain biases are classified as harmful, making their normative reasoning explicit.

In addressing the damaging effects of social biases embedded in machine learning models, Barocas et al. [6] and Crawford [34] proposed a framework of harms. They differentiate between *allocational* and *representational* harms. Allocational harms relate to differences in the allocation of resources or opportunities through a machine learning model, for example whether a resume filtering system accepts or rejects a candidate based on a social bias in the model [17]. Allocational harms could thus arise if group fairness is not given, since the performance of a model would be different for the two protected groups. Representational harms, however, arise from misrepresentation of protected groups, which can include stereotyped or denigrating representations, representations imposed upon a protected group by a third party [6], or omission of a protected group, thereby inhibiting their societal participation and recognition [28, 38].


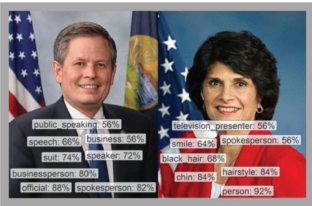

2.3 Sources of Bias

Biases within a model can have several points of origin within the machine learning pipeline. These are the *input data*, training or ground truth *labels*, *intermediate representations* such as word embeddings, the *model*, as well as the overall *research design*. Following Shah et al. [111], Hovy and Prabhume [61], Mehrabi et al. [88], and Fabrizzi et al. [46], we will go into further detail on each of these below. An overview with examples for each type of bias discussed can be found in Table 1.

Data. In a machine learning system training datasets contain selection bias, meaning that decisions about which data to include in the training sample will influence the model [111]. Selection bias is a necessary feature of data, since in most cases the training dataset is a sample of the entirety of available data. One clear example of selection bias would be a face recognition system's training data mostly containing pictures of light-skinned people, while the system's target population includes people of all skin colours [23, 46].

Even very large training sets, such as the Common Crawl Corpus used for training the **large language model (LLM)** GPT-3 [22] containing text data from the openly accessible Internet, suffers from selection bias due to greater Internet access in more developed countries over-representing their perspective [11]. Given the scale of datasets sourced from the Internet, termed "web-scale," it is often assumed that they mitigate or avoid selection bias. However, they are still influenced by economic considerations that dictate who, where and how Internet content is created and this is exacerbated for digital images. Besides this example of geographic bias, another example of selection bias concerns demographic bias, which relates to the demographic groups included in

Table 1. Overview of Bias Categories in Relation to Gender Bias in NLP and CV

Bias	Examples
<p>Selection/sampling bias Biases introduced as a result of the process by which instances are included in a dataset</p>	<p>NLP Using Wikipedia as a pre-training dataset, which has a strong male skew with regard to subjects and authors [131]</p> <p>CV  Using web crawling to retrieve images related to cooking results in images of mainly women, example from Reference[140]</p>
<p>Geographic bias Introduced due to origin or background of the data collector/creator [46, 112]</p>	<p>NLP Geographical bias toward countries most connected to the internet arising from scraping training data from the internet [11]</p> <p>CV Visual datasets overwhelmingly containing depictions of Western countries [133]</p>
<p>Demographic/population bias Bias caused by exclusion of demographic groups and demographic diversity [61]</p>	<p>NLP Misunderstanding of today’s young people’s speech as a result of common taggers trained on Penn-Treebank, whose training data are journalistic texts from the 1980s [59]</p> <p>CV Facial recognition software showing lower accuracy on darker-skinned women [23]</p>
<p>Labelling bias Annotations or labels used to identify subjects in data causing bias due to errors or human biases</p>	<p>NLP Annotations of hate speech/microaggressions being subject to human annotators’ perception of the threshold for hate speech [61]</p> <p>CV  Bias being generated from automated labelling of images of men and women [107]¹</p>
<p>Semantic bias Bias evident in pre-trained representations that contain semantic information</p>	<p>NLP $\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$ [19]</p> <p>CV  Bias in pre-trained models can be passed on in multi-stage models or systems [82]</p>
<p>Amplification Learning gendered differences and correlations that are exploited at prediction time over-amplifying the connection</p>	<p>NLP ML translation systems changing the perceived demographics of output sentences, making them sound older and more male [61]</p> <p>CV Generative models producing highly stereotypical images, e.g., men for CEO and women for housekeeper [82]</p>

(Continued)

Table 1. Continued

Framing bias

Gender-based differences in how subjects are written about [131] or how they are presented due to image capture techniques such as angle, focus and cropping

CV



Automatic cropping software cropping the mid portion of images of women [16]²

the dataset or how much demographic variation a dataset contains [61]. As an example, Hovy and Prabhume [61] mention that the training data for commonly used part-of-speech-taggers is mostly comprised of newswire data from the 1980s and therefore may not perform as well on data from other demographics such as the speech of young people in 2023.

Annotation. Label bias primarily arises through errors made by the annotators. For a visual example, please refer to Table 1. Label bias can have various causes [46, 61, 111]. First, annotators might not be working diligently, or they could lack the domain expertise needed for a certain annotation task. Second, even well-informed annotators may make mistakes if there are multiple possible labels but the annotation guide does not account for this possibility [95]. One example in CV would be the labelling of facial datasets for the category of *race*, which, if not based on the subjects' own perception or on scientifically based scales such as the Fitzpatrick skin type classification system [23], can lead to varying labels, since racial categories are not universally defined [46]. This also relates to the fact that demographic differences and differences in authors' and annotators' social viewpoints might lead to flawed annotations. As an example, Cao and Daumé [27] mentioned that some annotators have trouble with identifying the correct referent of singular *they* and neopronouns in sentences for which coreference information needs to be annotated. Last, Hovy and Prabhume [61] specifically mentioned the case of crowdsourced annotations, which are popular due to their low cost and scalability. Demographic variation between annotators and limited opportunities for annotator training can lead to undesired labelling.

Framing. One type of bias of particular relevance to CV is *framing bias*. Framing bias relates to how images are captured and composed to convey certain meanings that underline a difference or disparity [46]. An example are advertisements that draw focus on or depict only certain parts of women's bodies, while depicting men's bodies in full, contributing to the objectification of women [46] (see Table 1). Fabbrizzi et al. [46] mention that search engines for image retrieval are particularly prone to such framing bias. In the context of NLP, this can relate to gender-based bias in how people are represented in text. For instance, in Wikipedia, articles about women were shown to have a different structure than those about men [131].

Intermediate Representations. In NLP especially, other sources of bias are intermediate, pre-trained representations of words, such as word embeddings or contextualised word representations obtained through large language models [61, 111]. This kind of bias is also called *semantic bias*, since intermediate representations capture semantic and grammatical information about words. As such, they have also been shown to capture societal attitudes, biases, and stereotypes present in their training data [25, 71, 90]. Intermediate representations are obtained through unsupervised learning, usually using a large dataset. This dataset contains its own selection bias

¹Image © Schwemmer et al. [107].

²Image © Birhane et al. [16].

(discussed above), however, since the intermediate representations are then used independently of their training data, they are an additional source of bias. There have been various research efforts to remove or mitigate bias from intermediate word representations [87]. However, “biases are usually masked, not removed by these [debiasing] methods” [61].

Model. Models themselves can textit amplify existing biases. To make a prediction, a model may rely on “spurious correlations [...] or statistical irregularities in the data” [61]. Then, even though there may be only a small distance between data points in the training data, the model has made a discrimination that makes the difference more pronounced [111]. If the difference between data points is related to a protected group, then models can over-amplify the difference between the two groups. For example, Zhao et al. [140] found that an image labelling system over-amplified a difference of 33% more women associated with the activity of cooking in the training set to 68% at prediction time. Furthermore, Hovy and Prabhumoye [61] identified that bias amplification can also be caused by the fact that models are designed to always make a prediction, regardless of available evidence. They gave the example of the translation of the Hungarian genderless pronoun *ő* referring to the words *doctor* and *nurse* into a translation based on stereotypes (“he is a doctor, she is a nurse”), rather than presenting all available syntactically and semantically correct options to the user.

Research Design. Bias in the research design relates to biases within the field of research, which seep into research designs. As an example, Hovy and Prabhumoye [61] mentioned the overwhelming focus on English and other Indo-European languages within NLP, which creates a research culture in which it is more lucrative and prestigious to work on English, further amplifying this bias. Within CV and image datasets the recent domination of mobile phone cameras and the prevalence of image sources such as Flickr, Twitter and stock photography tend to create bias in both content and the source of the photography.

3 Bias Detection

Before attempting to mitigate gender bias in an AI system methodologies for measuring, and therefore detecting, bias need to be developed. In this section, we first discuss bias detection in NLP, distinguishing between *task-agnostic* metrics, used to assess word embeddings and language models, and *task-specific* metrics, used in downstream applications [4]. In the next section on bias detection in CV, we distinguish between *data-centric* methods, which specifically target training datasets, and *model-centric* methods applied to measure gender bias in trained models. We then discuss bias detection in visual-linguistic models and conclude with a comparative examination of gender bias detection in NLP and CV.

3.1 Natural Language Processing

In this section, we first present several measures for quantifying bias in NLP. *Task-agnostic* metrics, which target intermediate representations non-specific to a task, are discussed in Section 3.1.1 and *task-specific* metrics, i.e., those that measure bias for a specific downstream task, are discussed in Section 3.1.2. Section 3.1.3 presents research on the handling of gender beyond the male-female binary within gender bias research, which is increasingly acknowledged as a scientific gap within the NLP community. Finally, Section 3.1.4 discusses some overall limitations of current bias measures in NLP.

3.1.1 Task-agnostic Metrics. Task-agnostic bias metrics target models that are pre-trained to be used as input representations in later tasks, therefore these metrics measure semantic bias (cf. Table 1). As these metrics are applied to the trained model they are independent of the domain

Table 2. Overview of Task-agnostic Gender Bias Detection Methods in NLP, All Measuring Semantic Bias

Acronym	Name	Models tested	Method	Ref.
WEAT	Word Embedding Association Test	GloVe	cosine similarity	[26]
ILPBS	Increased Log Probability Bias Score	BERT	change in likelihood due to presence of gendered word	[71]
DisCo	Discovery of Correlations	BERT ALBERT	fills of masked word significantly associated with gender	[137]
SeT	Sensitivity Test	BERT RoBERTa	minimal change to last layer of model	[28]
HONEST	Hurtfulness of Language Model Sentence Completion	BERT GPT-2	number of times sentence is completed with a hurtful word	[92, 93]
CrowS-Pairs	Crowdsourced Stereotype Pairs Benchmark	BERT RoBERTa ALBERT	percentage of higher likelihood of stereotyped over anti-stereotyped sentence	[91]
StereoSet / CAT	StereoSet / Context Association Test (intersentence & intrasentence)	BERT RoBERTa GPT-2 XLNet	percentage of higher likelihood of stereotyped association and percentage of higher likelihood of meaningful sentence continuation	[90]

The *Models tested* column refers to the models that were tested in the original papers.

and therefore of the dataset. These models are either LLMs or their predecessor word embeddings. Some of the methodology that was first developed to show how pre-trained word embeddings capture social biases were later adapted for LLMs [53], however, other methods were tailored to the LLMs' context-dependent structure and training objective as language models. An overview of the methods discussed in this section can be found in Table 2.

Embedding Association Tests. One of the most commonly used frameworks for gender bias detection in NLP applications is the **Word Embedding Association Test (WEAT)**. The test was adapted by Caliskan et al. [26] from the Implicit Association Test used in psychological research [52]. The WEAT measures associations between identity terms that express gender, such as *he*, *she*, and so on, and positive or negative terms, or terms relating to fields with a stereotyped gender-connotation, such as family life or natural sciences. The metric used here is the distance between the terms' vector representations. While the WEAT has been praised for drawing on literature outside of NLP and thereby presenting an inter-disciplinary approach that grounds word embedding associations in human cognition [17], it has also been criticised for over-estimating bias [45].

With the emergence of LLMs, the WEAT was further adapted, as word representations in LLMs are not singular, like in traditional word embedding models, but are dependent on sentence context. May et al. [85] developed the **Sentence Embedding Association Test (SEAT)** and created "semantically bleached," i.e., very simple, sentence templates into which the target and attribute terms were embedded to extract the respective vector representations of the words. The authors tested their methodology on a variety of LLMs, but found discrepancies in the results, leading them to question whether the concepts tested (e.g., gender or pleasantness) can be represented within simple sentences and their association measured using cosine similarity.

Avoiding the problem of sentence templates, Guo and Caliskan [53] also adapted the WEAT for use in LLMs. For their so-called **Contextualized Embedding Association Test (CEAT)**, they extracted 10,000 sentences containing stimuli (target/attribute words) from a Reddit Corpus, compute the WEATs to obtain effect sizes for all pairings and then use a random effects model, which is used in meta analysis, to analyze the distribution of effect sizes. They found presence of all tested bias categories in all of the tested LLMs (GPT, GPT-2, BERT, ELMo) but also found some

negative results, indicating that “some WEAT stimuli tend to occur in stereotype-incongruent contexts more frequently” [53].

Increased Log Probability Bias Score (ILPBS). Kurita et al. [71] presented a method of measuring associations between identity terms and stereotypical terms within masked language models, such as BERT. The *log probability bias score* measures word likelihood in varying contexts. Similar to May et al. [85], they also created templates. However, instead of using semantically bleached contexts, their templates contain both an identity term and a stereotyped attribute term. The difference in association between counterfactual identity terms in sentences with the same attribute using an LLM is measured to create the score. Kurita et al. [71] found statistically significant bias scores where their adaptation of WEAT did not provide significant results. The authors interpret this as proof that gender bias assessment methods used on standard word embeddings cannot be simply translated to LLMs.

Discovery of Correlations (DisCo). Webster et al. [137] took a slightly different template-based approach. Instead of measuring the associations between pre-defined terms, their method aimed to discover terms correlated with gender (DisCo). They created two template variants for DisCo, one that uses first names, e.g., “[NAME] studied [BLANK] at college,” and another that uses nouns that contain gender information, e.g., “The [NOUN] likes to [BLANK]” [137]. The model was then asked to fill in the blank slot. The researchers used the χ^2 measure to see whether the three most likely proposals were significantly correlated with the associated gender of [PERSON] or [NAME], i.e., whether the proposed fill was gender-dependent, indicating model bias. Using DisCo, the researchers found that in both BERT and ALBERT, first names are more likely to generate fills with what they termed gendered correlations, than gendered nouns. Furthermore they showed that LLMs with similar accuracy do not necessarily show the same gendered correlations.

ABC Stereotype Model/Sensitivity Test (SeT). Another common criticism of research on bias and fairness in NLP is that techniques are not sufficiently grounded in theory from outside of the field, such as psychology, sociology, feminist theory [17]. Cao et al. [28] conducted a study asking participants to report stereotypes held by the general population and based their research on Koch et al.’s [68] **Agency Beliefs Communion (ABC)** stereotype model from social psychology theory. Cao et al. [28] also presented their own methodology for measuring stereotyping in LLMs: the SeT. The SeT measures how much model weights would need to change to arrive at predicting an anti-stereotypical trait for a given group, e.g., “Men are *kind*.” Compared to both CEAT [53] and the Log Probability Bias Score [71], the SeT showed better alignment with the ways in which humans tend to stereotype. However, overall human and model judgements only showed moderate correlation.

Open-ended Language Generation. Bias can also be measured through open-ended language generation. Sheng et al. [114] created sentence templates with placeholders for identity terms in contexts related to respect for a person and occupations. They used two measures to assess bias within the generated sentence completions by the LLM: sentiment as well as regard. Nozza et al. [92] also create templates that contain identity terms in different contexts, and they additionally presented the HONEST score, which uses the HURTLex lexicon of harmful language [9], to measure how often a language model’s top candidates for completing a sentence contain toxic language. They applied the HONEST score to BERT and GPT-2 models in six languages and found, for example, that sentence templates containing a female subject were completed with a reference to promiscuity 9% of the time. Nozza et al. [93] extended this research to LGBTQ+-related identity terms and

measured the HONEST score as well as toxicity on the sentence level using the Perspective API.³ They found that the completed sentences by the LLMs queried are classified to be harmful 13% of the time. Furthermore, Dhamala et al. [41] created the BOLD measures and dataset, which use sentences from Wikipedia that contain mentions of protected groups to measure bias in open-ended language generation. Akyürek et al. [3] took a critical perspective on using open-ended language generation as a measure for bias. They demonstrated that bias measures are highly dependant on experimental design, including factors like model parameters, which can influence whether or not bias reaches a harmfulness threshold.

Challenge Datasets. Measuring gender bias in NLP overall is dependant on datasets that contain identity terms for which different behaviours are measured. These challenge datasets are designed specifically to assess shortcomings with respect to a certain (social) variable. Challenge datasets are mentioned under several different names throughout the literature. Blodgett et al. [18] referred to “benchmark datasets,” thereby drawing the connection to performance-measuring benchmarks. Stanczak and Augenstein [119] called them “probing datasets” while Sun et al. [124] referred to “Gender Bias Evaluation Test sets.” Bowman [20] specifically mentioned adversarial datasets, for which annotators were asked to create cases that make a model fail.

Two challenge datasets to assess social biases in LLMs are CROWS-PAIRS [91] and STEREOSET [90]. CROWS-PAIRS consists of minimal sentence pairs, one of which contains a stereotype and one which does not. By measuring the likelihood that a given language model gives to either sentence, social biases can be assessed. STEREOSET is slightly more comprehensive; it features both intra- and inter-sentence settings. In the intra-sentence setting, sentences have a gap and multiple words to fill them: a stereotypical, a non-stereotypical and an unrelated filler. Measuring the likelihood of each of these fillers can provide indication of model bias. The inter-sentence setting is similarly structured, only here the likelihood of three possible sentence continuations is being measured. Despite being more comprehensive, the authors of CROWS-PAIRS noted that the STEREOSET has a lower annotator validation rate than CROWS-PAIRS [91].

3.1.2 Task-specific Metrics. In this section, we present research on detecting gender bias in specific NLP tasks. While task-agnostic metrics are mostly architecture-dependent (e.g., suitable for masked and/or causal language models), task-specific metrics are less tied to a specific architecture, because different model architectures can be used for the same task. Instead, methodologies are often dependent on challenge datasets, which allows researchers to test model performance with regard to a protected variable, gender in our case. Therefore, in addition to the discussion of several works on task-specific gender bias, we provide a non-exhaustive overview of these datasets in Table 3.

Coreference Resolution. Challenge datasets for detecting gender bias exist for a broad variety of downstream NLP applications. For example, the WINOBIAS dataset [141] targets pronoun resolution in pro- and anti-stereotypical sentences, such as “The physician hired the secretary because he was highly recommended” [141]. Similar datasets, which also target binary gender bias in coreference resolution, are WINOGENDER [100] and GAP [136]. Cao and Daumé [27] then developed the GICOREF dataset, which contains challenging coreference cases with non-binary and neopronouns, as well as gender-fluid cases, in which pronoun use changes while still referring to the same person.

Occupation Classification. De-Arteaga et al. [36] filtered almost 400,000 professional biographies from the Common Crawl Corpus and used this dataset to assess gender bias in occupation

³<https://www.perspectiveapi.com/>, Accessed: 27 April 2024.

Table 3. Datasets for Measuring Gender Bias in Specific NLP Tasks

Name	Data	Source	Size	Ref.
Coreference resolution				
WinoBias	two gender-stereotypical occupations paired with one pronoun (m/f)	sentences constructed by annotators following Winograd scheme, occupations from U.S. Department of Labor Survey	3,168 sentences	[141]
WinoGender	minimal pair sentences with one occupation and one human participant that differ in pronoun (m/f/n)	handcrafted sentence templates following Winograd scheme, occupations from Reference [26]	720 sentences	[100]
Gendered Ambiguous Pronoun (GAP)	two same gender named entities (m/f) with a corresponding pronoun	Wikipedia	8,908 ambiguous pronoun-name pairs	[136]
Maybe Ambiguous Pronouns (MAP)	two named entities with a corresponding pronoun, with controlled levels of gender information	Wikipedia	1,830 ambiguous pronoun-name pairs	[27]
GICoref	naturally occurring text up to 1,000 words about individuals with queer gender identities	Wikipedia, LGBTQ periodicals, Archive Of Our Own ⁴	95 documents	
Occupation classification				
Bias in Bios	short online biographies with occupation mentions	Common Crawl, BLS Standard Occupation Classification	397,340 biographies	[36]
Sentiment analysis				
Equity Evaluation Corpus (EEC)	sentences with gendered noun phrases/first names (m/f) and emotion phrases	handcrafted sentence templates, names from Reference [26], emotional state words from Roget's Thesaurus	8,640 sentences	[67]
Gender-Occupation Dataset	sentences with gendered noun phrases (m/f) and occupations with varying levels of gender participation	handcrafted sentence templates, occupations from U.S. Current Population Survey 2018	800 sentences	[14]
machine translation				
WinoMT	occupations and gendered pronouns (m/f/n) in (anti-)stereotypical settings; EN, ES, FR, IT, RU, UK, HE, AR, DE	concatenation of WinoBias and WinoMT	3,888 sentences	[120]
Occupations Test Dataset	gender-tags added; EN-ES, EN-DE sentences with pronouns/proper names (m/f), the word "friend" and occupations; EN-ES sentences with first-person singular	occupations from U.S. Bureau of Labor Statistics through Prates et al. [96]	1,000 sentences	[44]
Arabic Parallel Gender Corpus	references (m/f/ambiguous) + m sentences reinflected as f and vice versa	OpenSubtitles 2018 corpus [76]	12,348 sentences	[55]
MuST-SHE	sentences with first-person singular references (m/f/ambiguous), adjectives and nouns	handcrafted templates	226,175 sentence pairs	
MuST-SHE	multimodal instances with gender information in audio or text, annotated for speaker gender and gendered linguistic items; EN-FR, EN-IT	subset of the TED talk-based MuST-C corpus [42]	2,136 triplets (audio, transcript, translation)	[12]

m = male, f = female, n = neutral.

classification. They measured gender bias using the difference in true positive rate for male and female biographies per occupation in two settings: (1) gender markers contained in the biographies left as is or (2) they were "scrubbed." De-Arteaga et al. [36] found a significant gender gap that was correlated with statistics of gender participation in the workforce. When "scrubbing" gender information, the gender gap was reduced, but the accuracy of the classifier remained stable.

Sentiment Analysis. One of the first to study biases in **sentiment analysis (SA)** systems were Kiritchenko and Mohammad [67]. They created the **Equity Evaluation Corpus (EEC)** that consists of sentences designed to target race and gender biases within an SA system. Using this corpus to evaluate 219 openly available SA systems, the authors found that around three quarters of the systems consistently attribute higher sentiment intensity to identity terms related to historically disadvantaged protected groups, such as women and Black people [67]. Based on

⁴Organisation for Transformative Works, a nonprofit open source repository for fanfiction and other fanworks contributed by users, <https://archiveofourown.org>. Accessed: 27 April 2024.

this approach, Bhaskaran and Bhallamudi [14] created another EEC that is designed to expose gendered occupational stereotypes in SA systems. They found differing sentiment scores for male and female identity terms as well as more negative sentiment toward lower-earning versus higher-earning jobs. Addressing limitations posed by handcrafted templates, Asyrofi et al. [5] created BIASFINDER, a system that automatically generates templates that differ in identity terms of the same protected group, and for which different transformer-based SA systems predict differing sentiment. They call these sentence templates “**bias-uncovering test sets**” (BTC). On average, their SA systems find around 8,000 of these in an IMDB movie review dataset [79] and 24,000 in the Twitter SENTIMENT140 dataset [1].

Machine Translation. Another application of NLP for which bias is measured is **machine translation (MT)**. Gender bias in MT becomes evident, for example, when translating from a non-gender marking language to a gender-marking language, in which the choice of grammatical gender for an originally gender-neutral word is based on stereotypes or societal gender roles. For example, the phrase “The doctor and the nurse” would be translated into German as “Der Arzt_{masc} und die Krankenschwester_{fem}.” While it could be argued that this translation simply reflects numbers of male and female participation in the respective professions, Prates et al. [96] established that Google Translate, for instance, has a tendency to create male-default translations and to over-amplify men’s participation in STEM fields. In a similar study, Cho et al. [31] showed a similar male skew for gender-neutral pronoun translation from Korean to English. Besides using occupation words, they also demonstrated this skew for gender-neutral pronoun translation in formal/informal contexts, and contexts containing words that carry positive or negative sentiment. In addition, Stanovsky et al. [120] illustrates MT systems’ tendency to ignore morpho-syntactic contextual cues in coreference resolution settings in favour of stereotype information.

However, it should also be noted that with growing pressure from the public and academic research pointing at biases in MT systems, there has recently been some positive development such as Google providing several possible translations in the case of words with ambiguous gender [70, 104]. Besides problems in the translations of gender-neutral (pro)nouns, another form of gender bias recorded for MT is stylistic bias. Hovy et al. [60] found that due to the demographic skew in training data, automatic translations made users sound older and “more male.” Moreover, making gender information for first-person narration salient in non-gender marking source languages improved the translation of women’s voices into gender-marking target languages [130]. In addition, there exist several challenge or benchmark datasets to assess gender bias for different MT systems, such as WINoMT [101, 120], the occupations test set [44], the Arabic Parallel Gender Corpus [55] and MUST-SHE [12].

3.1.3 Bias beyond Binary Gender. Gender bias detection and mitigation efforts in NLP until this point have mostly employed a binary conceptualisation of gender, meaning that these works concentrated on equality in representation and quality of service for only male and female gender [38, 39]. Including other genders besides binary male and female was either not mentioned at all [71, 85], mentioned only when discussing future work [142], limitations [8, 41], or mentioned as an issue that the authors were aware of but could not be addressed, because it would complicate experiment setups [29] or the work was building on prior work with a binary conceptualisation of gender [126].

Devinney et al. [39] presented a two-round survey of conceptualisations of gender in NLP research in 2020 and 2021 and found that while awareness for and inclusion of non-binary gender models are increasing, more than half of all research surveyed still subscribed to the binary “folk” model of gender, according to which there are only two immutable categories of gender, *male*

and *female*. They found most works lacking explicit definitions of the conceptualisation of gender that was applied and of how gender was implemented in experiments. In line with this, they also found that social gender and linguistic gender are often conflated. Moreover, there are few works that address intersectional aspects in connection with gender, such as race or socioeconomic status. Devinney et al. [39] recommended future publications to explicitly define gender using appropriate and respectful language, subsequently select a method in line with the chosen definition of gender, and finally base the work on feminist research.

Dev et al. [38] moreover explored harms and challenges related to the exclusion or misrepresentation of gender identities that are non-binary in the context of NLP. Because non-binary gender identities (*non-binary*, *agender*, *genderqueer*, etc.) are not always recognised and not well-understood in large swaths of public discourse, training datasets for language technology reflect this lack of (accurate) representation. This data deficiency leads to language models, which currently function as the basis for most state-of-the-art language technology, creating “meaningless, unstable representations” for words used to express non-binary gender [38]. For example, the neopronouns *xe* and *ze* are treated as out-of-vocabulary tokens by BERT [40]. As a result, downstream applications such as machine translation or coreference resolution systems are likely to fail at resolving neopronouns and other language for expressing non-binary gender identities. The result is either the misgendering and/or erasure of non-binary genders [38]. For future work, the authors mentioned two primary challenges: the need for more real-world data of neopronoun use and a move away from a tripartite view of social gender as *male/female/gender-neutral* but rather a more open conceptualisation of gender to account for its fluidity [38].

3.1.4 Limitations. While measures for the assessment of bias have led to an awareness of how models integrate and emphasise existing biases in the data, a definitive bias measure that works reliably, especially in the context of large-scale language models, does not yet exist.

Aribandi et al. [4] tested the SEAT [85], CROWS-PAIRS [91], and STEREOSET [90] on three BERT models with different random seeds. They found that, while the performance of the models remained stable, predictions of the stereotypical categories in STEREOSET and CROWS-PAIRS, as well as statistical significance of the SEATs, appeared to be erratic (i.e., heavily influenced by the configuration of an individual model). In addition to inconsistencies in their application, Blodgett et al. [18] moreover found that a variety of pitfalls in four bias measuring datasets themselves. They analysed STEREOSET and CROWS-PAIRS, as well as WINOGENDER and WINOBIAS, which all contain contrastive pairs meant to measure a model’s performance on stereotyped versus non- or anti-stereotyped examples. Within these examples, the researchers found a string of inconsistencies in the operationalisation of stereotyping, with some examples being non-meaningful, misaligned, non-relevant, or containing offensive language in place of a stereotype, among others. Blodgett et al. [18] also criticised that all datasets analysed lacked a clear conceptualisation of stereotyping, which is their main focus.

Furthermore, as mentioned earlier, it should not be assumed that when task-agnostic gender bias can be measured for an intermediate representation, such as a word embedding model or language model, that this will definitively translate to task-specific bias in the downstream application [48]. Similarly, different bias measures might not necessarily correlate for the same model [37].

Overall, these works show that while the task of being able to measure problematic behaviour in models is important, it is also equally important to carefully construct measures [18] that remain robust to different configurations of the same model, take model uncertainty into account [4], and illustrate the influences on downstream applications [37].

Table 4. Overview of Model-centric Bias Detection Techniques in CV

Name	Bias Type	Bias Analytics Method	Processing Step	Ref.
Image Embeddings Association Test (IEAT)	Framing bias	Learning Representations	in-processing	[122]
Model Leakage, Bias Amplification	Selection, labelling, and framing bias	Dataset and Model leakage, Bias Amplification	in-processing, intra-processing	[134]
InsideBias	Selection bias	Learning Representations	in-processing	[110]
Bias correlation and amplification	Selection, labelling and framing bias	Learning Representations	in-processing	[140]
Grad CAM	Labelling bias	Learning Representations	post-processing	[69]
WEAT for CV	Selection, labelling, and framing bias	Learning Representations	post-processing	[83]
MCAS	Selection, labelling, and framing bias	Learning Representations	in-processing, intra-processing, post-processing	[82]

3.2 Computer Vision

This section discusses bias detection in CV datasets and models. Bias detection metrics take two approaches: (1) targeting the visual datasets [115, 133] and (2) targeting the models [69, 110]. An overview of model-centric approaches for measuring bias can be found in Table 4.

3.2.1 Data-centric Bias Detection. Gender bias in visual datasets is strongly influenced by the source of the images and the creation of the training labels. Online image hosting platforms, encyclopedias, and social networking sites are popular sources of visual data. Data-centric bias detection involves auditing and measuring gender bias in visual training datasets and is commonly performed via statistical methods, contextual representations, and empirical analysis of the datasets.

Statistical Methods. Statistical measures, such as frequency counts, are often employed to measure and analyse bias in datasets. These range from analysing demographic details, such as age, gender, and race, to using statistical techniques, such as **t-distributed Stochastic Neighbour Embedding (t-SNE)**, to visualise the distribution of images.

Using statistical methods, Singh et al. [116] compared image retrieval results across various image search and hosting platforms such as Bing, Twitter, the New York Times, Wikipedia, and Shutterstock. They used gender-skewed occupations such as *librarian*, *nurse*, *programmer*, and *civil engineer* and found, compared with data from the U.S. Bureau of Labor Statistics on gender participation in that particular occupation, that the New York Times had the most balanced representations while Twitter had the least.

As an example of selection and labelling bias, biased data from these online platforms is then used to curate datasets for training deep learning models. Yang et al. [138] studied image representations in the IMAGENET hierarchy and found gender bias in the very popular dataset. There were instances of labels having gendered and sexist slurs and pejorative keywords. Along with that, most annotations had gender bias such as the term *banker* having mostly male images. Yang et al. [138] used Amazon Mechanical Turk to correct some of these biases by balancing them for race and gender. Other similar debiasing techniques such as *relabelling* and *training data augmentation* have been discussed in Section 4.2.1.

Gender bias in visual datasets comes in many forms. Apart from labels, bias manifests in many implicit ways such as the depiction of gender in visual scenes. Wang et al. [132] analysed various popular vision datasets such as COCO, OPENIMAGES, and YFCC100M. They analysed image scenes and found that outdoor scenes such as transportation, snow and ice, deserts and sky, fields and

parks had more representation of men whereas indoor scenes such as shopping, dining, indoor sports, and leisure and home themes had a higher representation of women. In scenes related to objects, images under the categories *sports* and *vehicle* had more images of men, whereas those under the categories *kitchen*, *appliance*, *indoor*, and *furniture* had more images of women. Such issues can lead to framing bias in vision datasets.

Stereotypical gender representations are often over-amplified in image datasets. Kay et al. [65] studied the effect of stereotype exaggeration, systematic over- and under-representation and people's perception of stereotyping of gender in image search results returned by search engines. They used occupations that have a strong gender skew for their experiments and based their hypotheses on data on occupations collected from the U.S. Bureau of Labor Statistics. Kay et al. [65] found significant stereotype exaggeration with images associated with women for traditionally female dominated occupations and vice versa. Terms such as *sexy* and *attractive* returned a very high percentage of images of women ($\% = 0.8$ and 0.72 , respectively) and *professional* and *trustworthy* returned images of men ($\% = 0.27$ and 0.6 , respectively). This is another example of selection bias.

Statistical measures paired with visualisations can provide useful insights into the nature and distribution of bias in data. Karkkainen and Joo [64] used t-SNE to visualise the distribution of images by race to analyse their distribution. Such visualisations are useful to understand data distribution and bias in a high-dimensional space and have the potential to be extended to gender bias.

Contextual Representations. In their analysis of the OPENIMAGES dataset, Wang et al. [132] found that in images of people with musical instruments, men were often depicted as playing or interacting with them whereas women were more likely to be seen as audience. This meant that men were generally closer to the instruments. They found that men were more likely to be engaging with objects such as those related to sports and vehicles, and women with objects related to the kitchen, furniture, and accessories. Such representations can lead to framing and selection bias.

Empirical and Manual Analysis. Implicit gender bias is often difficult to measure using quantitative and statistical tools. Such biases are often hidden in the setting and context of the images and their associated texts and a qualitative analysis approach may be needed. Birhane et al. [15] studied the issue of harmful stereotypes in very large image datasets such as LAION-400M containing 400 million images and text extracted from the alt-text in web pages (i.e., crawled from the Internet). In the dataset they identified both harmful text and sexually explicit images and using text-based image retrieval methods found harmful images returned for terms related to women such as *Maa*, *Aunty* and *Abuela*. Similarly geographic biases were identified. The authors attribute this to the data creation method of crawling the Internet without filtering leading to labelling bias.

Using a similar methodology, Wang et al. [132], analysed the popular image dataset OPENIMAGES and found that in images of people and flowers, women were more likely to be posing with flowers, whereas men would have flowers used for background decoration. They found, using a **convolutional neural network (CNN)** trained on OPENIMAGES, that the model was then more likely to classify people in indoor sports, such as swimming, as women, and in outdoor sports, such as football, as men.

Benchmark Datasets. Another approach to detecting and measuring bias in CV is to create benchmark datasets to measure variance and diversity. Karkkainen and Joo [64] analysed various popular visual datasets such as IMDB-WIKI, LFW+, CELEBA, UTKFACE, among others, and looked into the racial and gender distribution. They found that apart from FOTW and UTKFACE, none of the datasets were balanced. They also created their own dataset called FAIRFACE, which had balanced racial and gender distribution. To compare the performance of their dataset, they trained

a ResNet-34 convolutional neural network on all the datasets individually and tested them on diverse sets of images such as images from geo-tagged tweets, media photographs, and protest datasets, all balanced for race and gender. They measured balanced accuracy on gender by using a variation of equalised odds to measure the difference in true and predicted gender. Karkkainen and Joo [64] found that the model trained on FAIRFACE performed better than the models trained on almost all other datasets.

As discussed in Sections 3.1.1 and 4.2.1, benchmark or challenge datasets as a bias mitigation tool present their own challenges. In an effort to provide a more objective measure of image diversity and to assess the impact of standard image search engines on the creation of image datasets, Mandal et al. [80] used Google Image Search with queries in different languages and using different geolocation settings (via a VPN) to gather images and create a dataset. Neural networks were then used to extract visual features for comparison to judge the resulting variation between searches and the overall diversity of the images independently of any labels.

3.2.2 Model-centric Bias Detection.

Image Embeddings Association Test. A popular methodology for bias detection in CV is to borrow and adapt techniques from other fields (such as NLP). Steed and Caliskan [121] proposed a methodology based on the Word Embedding Association Test (discussed in Section 3.1.1, *Embedding Association Tests*): the **Image Embeddings Association Test (iEAT)** to quantify implicit human biases. They hypothesised that human-like biases are present in the image embeddings used by neural networks. The iEAT measures the correlation between concepts. Using two sets of target concepts and attributes (e.g., *male-career*, *female-family*), the test measures the statistical differential association between them based on the model’s embeddings and produces a standardised measure of the probability that no bias exists.

A similar assessment of model bias was designed by Caliskan et al. [25], who used two CV models: iGPT and SimCLRv2, both pre-trained on IMAGENET. The experiment for gender bias included testing the models on two tests. First, Gender-Career test that measures the relative association of the category *male* with career attributes like *business* and *office*, and the category *female* with family related attributes like *children* and *home*. Second, Gender-Science test that measures associations between *male* with *science* and *engineering*, and *female* with *liberal arts* and *writing*. They found significant gender bias in both models in the Gender-Career test and the Gender-Science test with the standardised probability values being higher for SimCLRv2 than iGPT.

The use of iEAT to measure bias in CV models is relatively recent. Sirotkin et al. [117] used the iEAT to study the effect of **Self-Supervised Learning (SSL)** on bias. They studied three SSL models: *geometric*, *clustering-based*, and *contrastive*. Using the Gender-Career and Gender-Science tests, they found that contrastive models had the highest bias.

Model Leakage. The concept of *model leakage* was defined by Wang et al. [134] in studying **spurious correlations** in vision datasets that lead to bias. For example, in the popular image dataset COCO, there are more images that contain both *plates* and *women* than there are of images that contain both *plates* and *men*. This might lead to gender bias in models that then strongly correlate *plates* with female gender. This ability to infer gender from unrelated predicted image labels (*plate* predicts *female*) is termed “leakage.” The “leakage” in models is measured by the percentage of examples in the dataset that “leak” information about a protected label (e.g., *female*) through the model’s predicted labels (e.g., *plate*), assessed by training a new function that aims to predict the gender from the labels.

Bias Amplification. Models might to not only reflect the bias in the dataset, as in model leakage, but increase or amplify this bias. The term was defined by Wang et al. [134] as the difference

between the evaluated model leakage and the dataset leakage. Alternatively, Zhao et al. [140] measured bias amplification by comparing the effect of bias correlation learnt by a model during training. For example, in **visual semantic role labelling (VSRL)**, labels (*person, spatula, oven*, etc.) are generated for a scene such as a *kitchen* and the resulting activity shown in the image is *cooking*. If the positive correlation between two terms (e.g., women and cooking) is increased by the model over the evaluation dataset, this is termed bias amplification. The total score for the model is estimated as the average magnitude for all pairs that exhibit bias. Both of these metrics aim to quantify the bias influence of a model over a reference dataset, however, there is a risk of oversimplification and of reliance on a well-annotated reference dataset to compare against.

InsideBias. An alternative view of bias measurement is to inspect the internal structures of the model such as the activation of filters in a CNN, commonly used for state of the art CV models. Serna et al. [109] proposed such an approach to measure demographic bias and evaluated it by training two CNN architectures (VGG and ResNet) on DIVEFACE, a diverse dataset with representations from across the world, and on biased data by increasing the representation of a particular group. To assess the impact of biased data, an *Activation Ratio* is calculated. Activation, a measure of the contribution of network layers in generating the feature map, is compared between networks trained differently and the resulting models are considered biased if the ratio is less than a defined threshold. Generally the final layers of the network have the highest contribution and are evaluated in this way. Serna et al. [109] found that the unbiased models had a higher activation ratio for the last layers than the biased ones supporting their claim that this approach can give a good indication of model bias.

Grad CAM. Gradient Weighted Class Activation Mapping developed by Selvaraju et al. [108] provides localised visual explanations for CV models by creating a heatmap over the input image showing the regions contributing to the classification. It is done by computing the gradients flowing into the last (pre-fully connected) layer. This concept was utilised by Reference [83] for bias analytics. They used a visual question-answering machine based on CLIP (Contrastive Language Image Pretraining) similar to Reference [113] to analyse bias in CLIP image encoder models.

Difference Metrics. In contrast to inspecting the internal structures of the model, the difference in the model's output predictions can also be compared statistically. This is distinct from bias amplification methods (see previous section) that train a predictive function to reverse the model's outputs and calculate the leakage or correlation. In their work on debiasing neural networks, Savani et al. [102] used a fairness metric based on the difference in outputs predicted by neural networks for different demographic groups. These include **Statistical Parity Difference (SPD)**, **Equal opportunity difference (EOD)**, and **Average Odds Difference (AOD)**. True- and false-positive rates, standard metrics that measure the accuracy of a models output against the provided evaluation labels, are used to calculate the probability of positive outcomes (predictions or labels) for protected and unprotected groups. SPD measures the difference between the probability outcomes while EOD and AOD look specifically at the differences in true positive rates. Together, these metrics quantify prediction accuracy specifically focusing on protected and unprotected groups. Again, in common with other metrics, this is dependent on the quality of the reference annotations in the evaluation dataset. All these metrics are designed to work on CNNs. However, with the growing popularity of Vision Transformers (ViTs), similar metrics are required to audit bias in them. Mandal et al. [81] proposed *Accuracy Difference*, a metric that can measure bias in both CNNs and ViTs. The metric is a difference in accuracy between two sets of models: one trained on biased data and the other on unbiased data. The accuracy is measured on an unbiased test set. The higher the difference, the more the bias. They found ViTs to be more affected by bias due to a shallower loss

landscape leading to more generalisation and global attention and a larger receptive field allowing ViTs to learn more visual features and capture longer dependencies. These factors enable ViTs to learn more biased features from images.

3.2.3 Bias beyond Binary Gender. Similar to NLP, bias analytics in CV has focussed mainly on binary gender bias. However, unlike NLP, research on non-binary gender bias in CV is extremely limited. Luccioni et al. [78] studied the presence of intersectional gender bias in TTI models by evaluating their output using image captioning models and creating clusters based on visual features. Their tool **StableBias** also allowed for visual analysis of the outputs. They used prompts that included multiple identities such as occupation, ethnicity, and gender. Their tool allows for exploratory analysis of the output of TTI models but does not allow for quantitative measurement of bias, especially in the representation space. To generate a diverse dataset based on social characteristics, they used a pattern *Photo portrait of [X][Y]* with X and Y being characteristics related to ethnicity/gender and professional attributes. This dataset was used to evaluate three TTI models: DALL-E2 and Stable Diffusion v1.4 and v2. Then three types of analysis were done. In the first set, they performed an analysis of the text features of the generated images using two Image-to-Text models: a ViT-GPT-2 model trained on MS COCO and a VQA system based on BLIP. They analysed the text features for gender and ethnic markers for professional attributes and compared them with data sourced from labour bureaus. The analysis revealed DALL-E 2 has the highest deviation from the real-world data followed by Stable Diffusion v2 and v1.4. The second analysis involved clustering visual features extracted from the images using the same BLIP VQA used before. The results indicated that men made up most of the professional clusters. The third analysis involved creating an interactive tool to study these biases on a case-by-case basis.

3.3 Multimodal Models

Apart from considering data as strictly “text” or “visual,” there are emerging applications using multimodal or visual-linguistic models. Work on measuring gender bias in VL-BERT [123], a visual-linguistic model, was conducted by Srinivasan and Bisk [118]. The researchers measured associations between the gender of an agent (*man*, *woman*, *person*) and objects that have a stereotypical association with either male or female gender, such as *briefcase* versus *purse*. For this purpose, they adapted Kurita et al.’s [71] method for measuring associations in LLMs to the multi-modal setting, analysing the influence of visual-linguistic pre-training, as well as both single-modality contexts. Srinivasan and Bisk [118] found that visual-linguistic pre-training of VL-BERT shifts associations of the queried objects toward men. Moreover, the presence of a gendered agent in an image made the model more confident in predicting the object to be one that has a stereotypical association with the agent’s gender, even in the presence of contrary visual evidence. Similarly, Hendricks et al. [57] found that stereotypical associations with objects, such as between men and computers, influences caption generation even if there is contrary visual evidence (i.e., a woman sitting at a computer).

Generative vision models, especially text-to-image diffusion models, present a foundational shift in combining language and images. Cho et al. [30] studied gender bias in DALL-E, a text-to-image diffusion model by OpenAI. They generated images of humans using various attributes such as profession and politics and annotated them based on gender and skin tone using automated and human annotators. They then analysed the distributions of the annotations using *standard deviation* and *mean absolute deviation*. Their experiments revealed that Stable Diffusion suffered from more gender bias as compared to DALL-E. Similar observations were made by Mandal et al. [82] who developed **Multimodal Composite Association Score (MCAS)** to effectively and comprehensively detect and measure multimodal bias in text-to-image diffusion models. MCAS consists

of four individual association scores based on the WEAT association scores developed by Caliskan et al. [26]. They each measure stereotypical associations between binary gender attributes and real-world target concepts such as professions, scenes, sports, and objects in multiple modalities: image-text, image-image, and text-text. They found that Stable Diffusion generates more biased results than DALL-E, i.e., it is more likely to generate an image of a man for the word *engineer* and an image of a woman for the word *nurse*.

Most diffusion models make use of a visual-linguistic model like CLIP [97], which generates embeddings for the image generating diffusion process. Such models, such as CLIP and ALBEF [73], were analysed for gender bias by Zhou et al. [143], who developed *vision-language bias score (vlbs)* and *idealized vision-language ability score (ivlas)* to measure stereotypical associations in pre-trained vision-language models. *Vlbs* refers to the percentage of stereotypical predictions by a model for anti-stereotypical images. *Ivlas* is a combination of *vlrs*, which refers to the percentage of times a model ranks a stereotypical or anti-stereotypical caption higher than an irrelevant caption and *vlbs*. The authors used many vision language models in their study, such as CLIP, ALBEF, ViLT, and VisualBERT. Their experiments revealed that ALBEF has the least amount of bias and CLIP the highest. As these multimodal models combine language and vision, methods used in NLP for bias detection can be used for vision as well. One such cross-domain adaptation was the use of *Word Embeddings Association Test* (discussed in Section 3.1.1 by Mandal et al. [83]) to measure gender bias in CLIP. The authors used CLIP to predict labels for images of men and women from across the world and used the WEAT Association Score to measure the relative associations of real-world concepts such as those related to occupations (e.g., programmer and nurse) and adjectives (e.g., knowledgeable and feminine) to that of words representing men (e.g., man and he) and women (e.g., woman and she). They found traditionally male-dominated occupations and adjectives to be more closely associated with men and traditionally female-dominated ones with women. This approach showed a successful cross-domain adaptation of an NLP technique to audit bias in vision and multimodal models. The *Multimodal Composite Association Score* discussed earlier is based on this.

3.4 Comparative Analysis of Gender Bias Detection in NLP and CV

Perhaps the most straightforward connection between NLP and CV is the reliance on text for analysing bias in vision models. For example, Yang et al. [138] used the labels of images for part of their analysis. Another way of detecting bias in CV models is comparing generated captions for images that show gendered agents in specific contexts [140].

Some of the methods for the detection of gender bias in NLP and CV systems follow similar reasoning. One way to measure bias, which is employed in both fields, is to measure **performance differences**, which can create allocational harms [7]. For example, smile detection and facial detection have been shown to work better on White men's faces than on the faces of White women and women of colour [23, 102], and occupation classification from short biographies worked better for men's than for women's biographies [36].

Another area of overlap are methods derived from psychology's **Implicit Associations Test (IAT)** [52], which are the WEAT [26] and SEAT [85] in NLP for static and contextualised word embeddings, respectively, and the iEAT [121] for CV. The IAT was first adapted for word embeddings and subsequently for image embeddings, and measures associations between explicitly gendered words (*he*, *she*, etc.) and concepts that carry stereotypical associations with either female or male gender, such as the arts versus science. Another line of gender bias research in CV, which is similar to measuring associations, is to evaluate contextual cues in an image. These include the presence and framing of certain objects, such as flowers or musical instruments, in the presence of gendered agents [132].

In a similar fashion to measuring associations, there are methods for discovering **spurious correlations**, also referred to as model leakage, for both NLP and CV models [137, 140]. Spurious correlations describe correlations that are leveraged by the model to infer the gender of an agent, but which are based on stereotypes and thus undesirable. For example, a visual model might infer from the presence of a kitchen that an agent standing in a kitchen is a woman, because women are more often seen in kitchens in the training data, even if the agent is in fact a man [140].

When it comes to analysing the **training data** for gender bias, it is more common to gather statistics for the depiction or mention of different genders in CV than for NLP datasets [116, 138]. In NLP, for example, counting the presence of masculine and feminine pronouns is only mentioned as initial probes, if at all. However, both the latest NLP and CV models suffer from insufficient documentation when it comes to issues around gender bias or gender skew in their training data due to the fact that these data are obtained by web-crawling and thus reach a very large size while also containing high levels of noise [11].

While there are some **challenge datasets** specifically designed to show gender bias in models, such as CROWS-PAIRS [91] and STEREOSET [90], widely accepted versions of these kinds of specific benchmarks are still missing from the CV literature. There are some datasets, such as UTKFACE [139] and the Pilot Parliament Benchmark [23], but these mainly target performance differences instead of providing opportunities to measure stereotyping.

Another element missing from the CV literature are discussions around the handling of queer and non-binary gender identities, though some initial work is started to be done in this regard as discussed in Section 3.2.3. As noted in Section 3.1.3, most research on gender bias in NLP still uses the binary distinction of male and female to detect stereotyping or performance differences in systems [39]. However, many of these works mention the inclusion of other gender identities as aims for future work. A few papers have started to close that gap by, e.g., including neutral gender pronouns in co-reference resolution challenge datasets [27, 100], or the handling of non-binary gender expressions within large language models [38, 58].

This development of moving toward a more open conception of gender, which is subsequently integrated into the ways bias in models is detected, has not yet been as widespread in bias research in CV. This is a significant limitation. In CV, most models and datasets are assessed for binary gender bias with often no mention of including non-binary gender in future research efforts. However, as visual-linguistic models bring the two fields together, some works have started to assess non-binary gender representations in these models. One example is Ungless et al.'s [128] work on non-cisgendered representation in image-generation models, who found that generated images of transgender people appear to be less human and more sexualized. A successful example of cross-domain adaptation of NLP techniques is the use of **WEAT Scores for Vision**. This allows for more comprehensive bias analytics than previous methods based on metrics such as accuracy. Such methodologies can also be used for multimodal models. The **Multimodal Composite Association Score** is an extension of this method and is a promising direction for bias analytics for both vision and multimodal models.

4 Bias Mitigation

Model Pipelines for Bias Analytics. Modern deep learning models are extremely large, often having billions of parameters, and are trained and deployed using complex machine learning pipelines. Therefore, specific techniques targeting parts of these pipelines are required when mitigating bias [10]. Bellamy et al. [10] proposed AI Fairness 360, a toolkit and framework to detect and mitigate bias in these pipelines. They divided the techniques into three broad categorisations: **(1) Pre-processing** algorithms optimise the training data to make it fairer. These include reweighing (increasing the “weight” of features of minority groups), demographic parity (increasing the

Table 5. Data- and Model-centric Mitigation Techniques for NLP

Appl.	Name	Method	Process. Step	Ref.
training/ fine-tuning/ data	Counterfactual Data Augmentation (CDA)	gendered words (pronouns, nouns, names) are swapped (m/f)	pre-processing	[77], [84]
	Gender Neutralisation	gendered words (pronouns, nouns) are replaced with gender-neutral equivalents	pre-processing	[125], [129]
WE	Direct debiasing	remove projection onto gender subspace from original vectors	post-processing	[19]
	Gender-Neutral Global Vectors (GN-GloVe)	gender information concentrated in specific dimensions of vector which are subsequently removed	in-processing	[142]
LLMs	SentDebias	adaptation of <i>direct debiasing</i> [19]; contextualization of gendered words within randomly extracted sentences	post-processing	[74]
	Iterative Nullspace Projection	linear classifier to learn gender direction, use the nullspace of the classifier's weight matrix to debias sentence representations	post-processing	[99]
	dropout regularisation	reduce bias introduced by spurious correlations through dropout regularisation	in-processing	[137]
	Self-Debias	use toxic text generation to scale down the probabilities of the toxic generation in a second generation	post-processing	[105]
	Auto-Debias	automatically generate prompts that show a large difference in probability distribution for masked tokens in presence of gendered words, then use prompts to minimize difference during finetuning	post-processing	[54]
	Bias removal without losing factual gender information	use orthogonal probe to separate factual and stereotypical gender information and filter out the gender bias subspace from embedding space	post-processing	[75]
Equalize and Decluster	using bias mitigation losses (equalizing loss, declustering loss) during further pre-training	in-processing	[47]	

m = male, f = female, WE = Word Embeddings.

representation of underrepresented groups), optimised pre-processing (data transformation subject to fairness constraints) and learning fair representations (obfuscating protected attributes). Some of these algorithms have been adapted for CV. Some of these algorithms can also be used for creating fairer datasets as well. (2) **In-processing** involves modifying the deep learning models, either by changing the network or its training process. Examples include adversarial debiasing and layer-wise optimisation. These algorithms can target classification layers and data representations learnt by the network. This can also be used for detecting bias such as in case of the iEAT. (3) **Post-processing** algorithms modify the outputs of the deep learning algorithms to make them more fair.

We will use this categorisation in this section when presenting several current techniques for gender bias mitigation in NLP and CV in Tables 5 and 6. Overall, we divided each subsection on NLP and CV, respectively, into *data-centric* and *model-centric* approaches. Data-centric approaches target a model's training data through augmentation or alteration thus covering pre-processing approaches, while model-centric approaches make changes to the model's parameters, either during training (in-processing) or as a post-hoc method (post-processing). Similar to Section 3, we further discuss research on bias mitigation in visual-linguistic models and conclude this section with a comparative examination of gender bias mitigation in NLP and CV.

4.1 Natural Language Processing

This section presents research on the mitigation of gender bias in NLP. We will first discuss methodologies aimed at increasing gender parity or gender neutrality in the training data in Section 4.1.1. Second, Section 4.1.2 presents methods that mitigate bias in word embeddings and LLMs.

Table 6. Overview of Model-centric Bias Mitigation Techniques in CV

Technique/Metric	Bias Type	Bias Mitigation Method	Processing Step	Ref.
Adversarial Loss	Selection, labelling, and framing bias	Adversarial Debiasing	in-processing, intra-processing	[134]
Random perturbation, Layer-wise Optimisation, Adversarial Fine-tuning	Selection bias	Learning Representations	intra-processing	[103]
Structured output Prediction, Corpus level Constraints, Lagrangian Relaxation	Selection, labelling, and framing bias	Model fine-tuning	intra-processing, post-processing	[140]
Strategic Sampling, Domain Discriminative Training, Prior shift inference, Domain independent training	Selection bias	Sampling and Adversarial debiasing	pre-processing, in-processing	[135]

4.1.1 Data-centric Bias Mitigation. As we described in Section 2.3, the training data are one of the primary entry point for gender bias into NLP models. Therefore, changing the data in a way that counters prevalent gender imbalances and stereotypes is an obvious starting point for bias reduction.

Counterfactual Data Augmentation (CDA). One of the most common methods is **Counterfactual Data Augmentation (CDA)**. In CDA, pronouns and nouns referring to female (male) gender are swapped for those referring to male (female) gender. For English text, Lu et al. [77] appended sentences edited in such a way to the original training data, thereby augmenting the corpus. Maudslay et al. [84], however, substituted the original sentence for the augmented one, which they called **Counterfactual Data Substitution (CDS)**. They additionally developed a method of also swapping first names in such a way that name frequency (*James* vs. *Bart*) and gender-specificity (*Anna* vs. *Taylor*) were preserved. Bartl et al. [8] applied CDS to fine-tuning data for English BERT and demonstrated bias reduction using the log probability bias score by Kurita et al. [71]. Dinan et al. [43] and Webster et al. [137] moreover demonstrated the usefulness of CDA for bias mitigation on dialogue generation tasks. While the previous works have focused on English, Zmigrod et al. [145] showed that CDA is also useful for reducing gender stereotyping in gender-marking languages (Hebrew, Spanish, French, Italian) and moreover provided a method for adjusting the gender of dependants of a swapped instance according to the “new” gender.

Gender Neutralisation. Instead of swapping gendered words, another strand of research targeting the training data concerns the creation of gender-neutral text. Vanmassenhove et al. [129] and Sun et al. [125] both developed applications designed to turn gender-specific into gender-neutral sentences. For instance, the sentence “Every stuntman accepts a considerable risk of injury in his job.” would be turned into “Every stunt performer accepts a considerable risk of injury in their job.” The researchers mentioned machine translation as one possible area for the application of post-hoc gender-neutralisation. Vanmassenhove et al. [129] additionally mentioned that gender-neutral text created by their system could also be used to mitigate bias in training data, but left this to future research.

4.1.2 Model-centric Bias Mitigation. Model-centric debiasing techniques have traditionally focused more closely on mitigating semantic bias, meaning bias in intermediate textual representations, than bias in the models themselves.

Word Embeddings. As one of the earlier works on illustrating gender bias in word embeddings, Bolukbasi et al. [19] presented a method called *direct debiasing*, with the rationale of removing associations with gender from the embeddings. They obtained a gender subspace through combining the vectors for a variety of words containing masculine and feminine gender, such as pronouns,

then projected the word vectors onto this subspace, and subsequently removed the projections from the original vectors.

Zhao et al. [142] critiqued Bolukbasi et al.'s [19] method for completely removing gender information from word embeddings, which might not always be desirable. To overcome this flaw, they presented a method for learning gender-neutral GLoVe word embeddings [94]. Their approach concentrated gender information in specific dimensions of the vectors, which could then be removed to reduce biased gender associations but preserve factual gender information. Using Bolukbasi et al.'s [19] gender direction, they illustrated a reduction in gender stereotyping and moreover showed a reduction in bias on a coreference resolution task [142].

In a seminal work, Gonen and Goldberg [49] then assessed the effectiveness of the two previously described debiasing techniques by Bolukbasi et al. [19] and Zhao et al. [141]. They stated that the way bias removal is conceptualised in these works, that bias is removed if definitionally gender-neutral words all have an equal distance to all pairs of explicitly gendered words, ignores more implicit associations relating to gender stereotypes. In their experiments, Gonen and Goldberg [49] used clustering and gender prediction to show that stereotypical gender information can still be easily recovered from de-biased embeddings.

A method using projections to remove bias, *Iterative Nullspace Projection*, was proposed by Ravfogel et al. [99]. They trained a linear classifier to learn the direction corresponding to attributes of a protected group. Then, to debias, they projected the sentence representations into the nullspace of the linear classifier's weight matrix, thereby removing information about the protected group.

Large Language Models. Since the emergence of transfer learning from pre-trained, transformer-based LLMs and their widespread adoption within the field of NLP, the most recent efforts at debiasing NLP models have focused on language models.

One approach taken by Liang et al. [74] was the adaptation of Bolukbasi et al.'s [19] direct debiasing technique for LLMs, which is called SENTDEBIAS. Liang et al. [74] contextualised a set of identity terms and terms related to a protected group, gender in this case, in randomly extracted sentences, estimated a bias subspace from these sentence representations, and subtracted the projection onto the subspace from the LLM's representations.

Webster et al. [137] targeted model bias that is created through the exploitation of spurious correlations with gendered identity terms. To reduce those correlations, they increased the dropout parameter during additional pre-training for BERT [40] and ALBERT [72]. Dropout regularisation is normally used to avoid overfitting, but Webster et al. [137] showed that the effect of reducing superfluous correlations also reduces correlations that express stereotyping in masked LLMs, while keeping performance consistent.

Opposed to the previous methods, which essentially change the model's internal representations, Schick et al. [105] proposed a post-hoc method called SELFDEBIAS. They base their method on the observation that LLMs are able to detect when their own output contains toxic or biased text, which they call *self-diagnosis*. Based on this observation, they then first prompted the model to create a text containing a form of bias, and subsequently de-biased by scaling down the probabilities for the generated biased text for a secondary generation of text.

Guo et al. [54] presented their approach called AUTODEBIAS. It is similar to Schick et al.'s [105] SELFDEBIAS, however, instead of crafting prompts to elicit biased text, Guo et al. [54] automatically found prompts that can be used for de-biasing a **masked language model (MLM)**, thereby reducing the reliance on external corpora. They choose prompts "such that the cloze-style completions have the highest disagreement in generating stereotype words (e.g., manager/receptionist) with respect to demographic words (e.g., man/woman)" [54]. These prompts were then used to fine-tune the MLM in such a way that the disagreement between the two generations for binary

gender words are minimised. Guo et al. [54] showed that this kind of de-biasing does not hurt model performance on the GLUE benchmark.

Limisiewicz and Mareček [75] aimed to preserve factual gender information while removing gender bias from the top layer of pre-trained, transformer-based language models. They used an orthogonal probe to distinguish between gender associations related to factual gender versus gender bias, and then filtered out the bias subspace from the embedding space. They showed that, while not all of the stereotypical bias is removed, their method succeeded in mitigating bias while preserving language modelling ability [75].

Garimella et al. [47] presented another approach to bias mitigation by introducing two additional loss functions during additional pre-training, an *equalising loss* and a *declustering loss*, aimed to “equalize the associations of words with different groups of a given demographic” and “decluster the various clusters of words that may be indicative of certain kind of implicit bias with respect to the demographic” [47]. They evaluated this method on a BERT model using SEAT [85] as well as human evaluations and found sentence completions to be less biased. In addition to bias reduction during pre-training, the researchers also presented a bias mitigation objective during decoding for a specific language generation task, text summarisation in this case.

Meade et al. [87] compared several of the previously presented gender bias mitigation techniques for LMs: Dropout regularisation [137], SENTENCEDEBIAS [74], SELFDEBIAS [105], Iterative Nullspace Projection [99], and CDA [77, 145]. Measuring bias with the SEAT, STEREOSET, and CROWS-PAIRS, they found SELFDEBIAS [105] to be the most effective technique, that also consistently preserved language modelling ability.

4.2 Computer Vision

Bias mitigation techniques in CV can be generally categorised into two categories: debiasing the training data (Section 4.2.1), and modifying the learning representations (Section 4.2.2).

4.2.1 Data-centric Bias Mitigation Techniques. Data-centric bias mitigation techniques involve modifying the training data to either have unbiased training datasets or use of specific datasets to de-bias existing models.

Relabelling. The most straightforward method to reduce data-centric bias is to relabel or refine the existing annotations and classifications. This can potentially mitigate framing, labelling, and selection bias. Relabelling can be expensive, time-consuming and require significant domain expertise but allows the utilisation of existing large image collections. OPENIMAGES contains about nine million images that contain five person-level annotations: *person*, *man*, *woman*, *boy*, and *girl*. Schumann et al. [106] studied the gender bias in these annotations and proposed a new framework called **More Inclusive Annotations for People (MIAP)**. For example, they found that in images containing both men and women in settings such as weddings, the bounding box focused on women whereas it was reversed in case of images depicting military personnel. They introduced MIAP to replace the five person-related keywords with three terms: *predominantly feminine*, *predominantly masculine* and *unknown* in an effort to mitigate these effects.

Training Data Augmentation. One of the more fundamental approaches to bias mitigation involves modifying the training data with respect to model behaviour. Zietlow et al. [144] used an *Adaptive Sampling* method to improve fairness in vision classifiers. They started with two sets of training data: an original set and an extended set. They trained a classifier on the original set and determined the worst performing group using a hold-out dataset. Then they added the group to the extended dataset and measured the resulting changes in the classifier’s performance using a sampling approach (g-SMOTE) that promotes oversampling of minority classes from the data. The

results of their experiments showed a considerable improvement of the model when retrained with the augmented data with increasing representation of the worst performing group. This method also outperformed other fairness techniques such as weighing and fairmix [144], and help mitigate selection bias.

Benchmark Datasets. Rather than trying to improve an existing labelled dataset, benchmark datasets, such as the FAIRFACE dataset [64], are specifically created to serve as a standard against which training data can be checked. They aim to provide a reference for gender and racial diversity. Similar datasets such as UTKFACE [139] and the Pilot Parliament Benchmark [23] have also been proposed. However, these benchmark datasets risk incorporating their own selection bias, most prominently a Western-centric bias on how race and gender are conceptualised, and the authors proposed various methods to avoid compounding such bias. A dataset creation process where a variety of terms in different languages and different geolocations for Google Image Search was used in Mandal et al. [80] to assess the impact of the dataset creation choices on bias. Such datasets can help mitigate selection bias.

4.2.2 Model-centric Bias Mitigation. Model-centric bias mitigation techniques generally involve targeting the internal representation learnt by the model in its embedding space. These include modifying the training objective function to focus on debiasing and adversarial debiasing. An overview of model-centric bias mitigation techniques in CV is provided in Table 6.

Learning Representations. Sampling techniques, such as those used in the previous approach, come with their own issues including potential for oversampling and overfitting. Wang et al. [135] proposed two techniques to overcome these issues: *Domain Discriminative Training* and *Domain Independent Training*, which can help mitigate sampling bias. **Domain discriminative training (DDT)** works on the opposite principle of the “fairness through blindness” concept. In DDT, information is first encoded and then mitigated. The model first learns correlations between the target class and the domain that leads to bias (such as *man-programming* and *woman-cooking*). Then the model is trained to minimise these correlations to reduce bias. In **Domain Independent Training (DIT)**, the model learns these correlations but is trained to ignore these class boundaries. The authors tested both the methods on the CELEB-A dataset. Their aim was to remove gender bias by using a weighted **mean average precision (mAP)** metric to simulate equally distributed samples between the genders. They found that the DDT model performed worse than the base model (73.8% vs. 74.7% mAP), while the DIT performed better with 76.3% mAP.

Model Fine-tuning. Large pre-trained models work well on general CV problems. They are however difficult and expensive to train. Therefore, any significant retraining is expensive and time and resource consuming. This has led to the development of fine-tuning algorithms where models are retrained to achieve specific goals including for debiasing. One such method is *Adversarial Debiasing* proposed by Wang et al. [134] to reduce model leakage (see Section 3.2.2 *model leakage*) by discouraging the model from building representations from protected attributes such as gender. They construct a *critic* model that tries to predict protected attributes from an intermediate representation for an image from a competing predictor model. In its simplest form, the predictor tries to improve classification performance at the expense of the critic (meaning that the critic’s ability to predict protected attributes decreases) to result in a more balanced and less biased system. The authors also experimented with optimisation of the adversarial loss on the input feature space by using an encoder-decoder model and auto-encoding input image.

Wang et al. [134] used three types of adversaries to remove leakage at different stages in a ResNet-50 classification model. The first targets the image directly, trying to remove gender information by using a U-Net as the encoder-decoder network to predict a mask. The second removes

gender information from an intermediate representation of ResNet-50 (the fourth convolutional block) using an adversary having three convolutional layers and four linear layers. The third method removes gender information from the last convolution layer of ResNet-50 using a linear adversary taking a vectorised form of the output feature map as input and a four-layer MLP as classifier. These approaches try varying methods of crafting suitable adversaries (the critic and the predictor) based on the image or specifically targeting a layer of the model. Various models are used as the base model, trained on original data, and models trained on different augmented data such as with Gaussian blur, face blackout and blur. From their experiments, Wang et al. [134] found that the three adversarial models resulted in less bias amplification than the baselines with the second approach (targeting an intermediate representation layer) performing the best.

A second approach to model fine-tuning are *Intra-processing Methods*. As demonstrated by the success of the adversarial method targeting an intermediate layer, it is possible to debias pre-trained models by focusing on CNN layers. Savani et al. [102] proposed intra-processing algorithms for debiasing vision models trained on large generic datasets, as a complement to in-processing methods. They propose three intra-processing algorithms: *random perturbation*, *layer-wise optimisation*, and *adversarial fine-tuning*, which we discussed above as an example of adversarial debiasing.

The intra-processing algorithm takes the validation dataset and a trained model with a set of weights and outputs a fine-tuned weights set that optimises the desired outcome. The authors proposed the following intra-processing debiasing algorithms that optimise metrics similar to the difference metrics explained in Section 3.2.2, which are based on true and false positive rates. *Random perturbation* is an iterative algorithm in which every weight in the network is replaced by a Gaussian random variable (mean of 1 and standard deviation of 0.1) in every iteration. This aims to disrupt the training and force the layers to avoid over-generalisation that can lead to bias. *Layer-wise optimisation* debias the model by debiasing individual layers using a more reliable means of finding an optimum network point, and that can only operate on a feed-forward neural network. In their experiments, Savani et al. [102] used Gradient Boosted Regression Trees as the optimiser. The authors found significant bias reduction using a ResNet model tested on the CELEBA dataset with the Layer-wise optimised model outperforming the random perturbations, again showing the advantages of a more targeted approach.

A third technique for model fine tuning is **Reducing Bias Amplification (RBA)**. Deep neural networks learn representations in the data by creating correlations between the features in the input. This can lead to the network amplifying certain correlations that may then amplify any bias present in the training data. Zhao et al. [140] proposed RBA to reduce bias arising out of spurious correlations in visual datasets. Details in images can often contain features introducing bias in models trained on them (as explained in Section 3.2.2: *Bias Amplification*). The authors here aimed to mitigate such biases by injecting constraints to make sure that the model follows the distribution present in the training data. The proposed algorithm is a meta-algorithm based on Lagrangian relaxation consisting of three main parts. The first part involves structured output predictions, where a scoring function is created based on the model and decomposed to extract the part concerned with semantic labelling. In the second part, corpus level constraints are introduced to ensure that the output labels follow a desired distribution. For example, the gender ratio for each activity can be constrained. The third part involves solving this constrained problem, expressed as an integer linear program – a set of linear constraints over integer variables, by using a solver (the authors used Gurobi Optimisation in their experiments).

This algorithm was evaluated on two tasks: **visual semantic role labelling (vSRL)**, and **multi-label classification (MLC)**. They focused on gender-specific terms (*man* and *woman*) and the agent in vSRL and text association with images in MLC. For vSRL, they used the IMSITU dataset

containing about 125,000 images with activity classes drawn from FRAMENET and noun categories drawn from WORDNET. Non-human activities were filtered out. They build a **Conditional Random Field (CRF)** model for testing. For MLC, they used MS-COCO, an object detection benchmark containing 80 different object types and no gender related captioning. They used a CRF based on ResNet-50 as the model. Both the datasets are biased toward men with 64.6% and 86.6% for IMSITU and MS-COCO, respectively. The results showed that the debiased models had bias reduction as compared to the baseline models.

4.3 Multimodal Models

Methods have been developed to mitigate social biases in multimodal models. Berg et al. [13] proposed a method to debias multimodal models like CLIP [97] by using an objective function to reduce bias and hyperparameter optimisation for bias reduction. They combined their approach with adversarial debiasing and found a significant reduction in bias, especially in CLIP. When the methods were used individually, the bias reduction was limited. They also qualitatively demonstrated the effectiveness of their method.

Tang et al. [127] developed a method to debias visual captioning models using a self-guidance mechanism on visual attention to learn from the correct gender features. They used two parallel streams to simultaneously generate captions and focus the model's attention on the correct regions of an image allowing the model to focus less on stereotypical features of an image. The authors found significant bias reduction in the trained models on metrics such as gender accuracy and attention correctness.

4.4 Comparative Analysis of Gender Bias Mitigation in NLP and CV

Natural Language Processing and Computer Vision have many similarities when it comes to bias mitigation. The methods in both fields are drawn from a diverse set of research areas including machine learning, social sciences and statistics. Many of the methods used are similar at a conceptual level and some at implementation level. In the following, we will discuss some of the similarities and differences in bias mitigation in the two fields.

First, it has become a convention to release large-scale models in both NLP and CV together with **model cards**. Model cards were introduced by Mitchell et al. [89] to increase documentation and inform intended users about the risks of using a model. These cards provide information such as general information (model type, version, developer, and fairness constraints), factors (demographic groups, environmental conditions, and technical attributes), and ethical considerations. Through more comprehensive documentation, especially regarding ethical implications of their models, model engineers are encouraged to mitigate biases in their models. Additionally, if models are released without addressing ethical considerations, engineers of integrated systems might be reluctant to use the respective model.

A second area of similarity is related to **gendered associations** that models learn from data, which may introduce or amplify gender bias. Research on decreasing these associations to reduce gender stereotyping is conducted in both NLP and CV. In NLP, Garimella et al. [47] introduced loss functions to equalise association of words belonging to different demographics. In CV, Wang et al. [135] advocated for using techniques to train models, which separates domain information and either avoids correlation entirely or minimise it by actively identifying it. Webster et al. [137] showed that using dropout can reduce gendered correlations in language models and Savani et al. [102] debiased convolutional neural networks by fine-tuning parameters of individual layers.

Another method for debiasing used in both the fields is by harnessing the **learning techniques** of the models. Schick et al. [105] created SELFDEBIAS to determine if their own output contains bias and Wang et al. [134] used adversarial debiasing to explicitly reduce bias. Both fields have used

fine-tuning to debias large pre-trained models. Zhao et al. [140] used corpus-level constraints and Lagrangian relaxation to enforce distributions learnt from a debiasing dataset on the model outputs. Bartl et al. [8] used Counterfactual Data Augmentation to fine-tune BERT and Zietlow et al. [144] used Adaptive Sampling to augment training data by iteratively analysing model performance on target demographics.

Along with the similarities, there are differences between between the fields as well. Bias mitigation techniques in CV are more quantitative and almost always use metrics such as bias amplification and model leakage to mitigate bias based on a human-labelled reference dataset. CV aims to attach meaning to visual data that is ambiguous, poly-semantic and multi-layered by nature. At the current stage of research this precludes general application of the more structured understanding used in NLP. What's more, bias mitigation in NLP can draw upon lexical properties of gender to perform model debiasing. Zhao et al. [142] and Bolukbasi et al. [19] proposed making gender neutral words equidistant from masculine and feminine words—something that is not possible in CV.

The process of comparing gender bias mitigation in both NLP and CV has identified the following common limitations. First, there is a focus on binary gender without proper consideration of the nuances and changes in societal views. This is commonly done to simplify the training of models but often results in challenges for detecting or mitigating bias [39, 133, 140]. Second, datasets are often composed with a focus on a small pool of gendered words/stereotypical occupations (NLP) or, in vision, with small numbers of examples relating to gender [54, 69, 102]. This leads to a narrowing of the understanding of gender bias and restrictions in discussing or applying mitigation practises. Third, the increasing automation of dataset creation at significant scale. To achieve this, generalisation often occurs, which tends to increase the probability of bias or imbalance in datasets [11, 115].

5 Conclusion

In this survey, we presented research on the detection and mitigation of gender bias in the fields of Natural Language Processing and Computer Vision. We first introduced theory on the conceptualisation of gender, terminology related to bias and fairness, possible sources of bias in the machine learning pipeline, as well as legal dimensions of trustworthy AI in the European Union. The main part of the survey presented strategies for gender bias detection and mitigation for both NLP and CV, respectively, as well as for combined visual-linguistic models. The sections on bias detection and mitigation were each closed with comparative analyses of methods in the two fields.

Comparing the state of gender bias detection and mitigation in NLP and CV, we found much conceptual overlap, even if the actual operationalisation was necessarily constructed to work with the respective model architectures. We found conceptual overlap for example in the observation that gender bias was often measured through the associations or correlations between words that contain gender (*she*, *man*, etc.) or gendered agents, and concepts that are related to gender stereotypes, such as specific occupations. Gender bias mitigation then aimed at reducing these associations.

A more concrete example that not only illustrates conceptual similarities but active interdisciplinarity is the adaption of the WEAT [26], which detects stereotypical associations, for CV models into the iEAT [121]. This shows potential for transferring further bias-related methodologies from NLP to CV, especially seeing the inherent connection of text and images through labels and captions.

Another area in which work on gender bias in CV could benefit from previous approaches in NLP is the adoption of theoretical frameworks outside of the field, such as from the social sciences, psychology, and gender theory. This will allow for a better and more comprehensive conceptualising of gender leading to a better understanding of gender bias. In NLP, work on gender bias has

previously been criticised for not being sufficiently grounded in theories outside the field, which resulted in vague definitions of bias [17], unclear conceptualisations of what was meant by *gender* and how gender was operationalised [39]. While these criticisms have inspired more recent works to engage with and discuss the concepts of *gender* and *bias*, and used these considerations to inform their research [28, 104], clear conceptualisations of either *gender* or *bias* are still missing from CV works. Moreover, research on gender bias in NLP builds on theories related to the construction and performance of gender through language as well as linguistic categories such as referential and lexical gender (see Section 2.1), but in research on gender bias in CV models it is unclear what visual attributes of the agent themselves contribute to identifying their gender.

It is not only the case that the conceptualisation and operationalisation of gender is often not made explicit, but at present most works on gender bias in NLP and nearly all in CV treat gender as binary. This focus on binary gender is also contained and reproduced through datasets, which include few mentions or instances of non-binary genders. In CV, for example, labelling images of people as only *men/boys* or *women/girls* will further solidify this distinction and lead classifiers to identify only those two categories and thus possibly misgender people in images. Generally, there is an argument to be made regarding the necessity for gender classifiers in CV that make a decision based on visual features as these remove the option for self-identification. Therefore, implementing a more open view of gender that allows for more than two, and ideally more than three non-discrete categories, presents not only interesting, but vital avenues for future research in both fields [38, 39, 118].

Overall, we have illustrated parallels and potential for inter-disciplinary cooperation between the fields of Natural Language Processing and Computer Vision with regards to detecting and mitigating gender bias. Both NLP and CV models are contained in a variety of applications that have become part of everyday life, such as social media, search engines, and news aggregators with high potential for life-changing impact and harm. It is therefore important to be able to assess biases in a joint fashion, especially as multimodal, visual-linguistic models gain more popularity and widespread use. We therefore encourage future collaboration between the fields of CV and NLP to create trustworthy AI systems.

Acknowledgments

For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

- [1] Ahmed Abbasi, Ammar Hassan, and Milan Dhar. 2014. Benchmarking Twitter sentiment analysis tools. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), 823–829.
- [2] Lauren Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa* 4, 1 (2019), 1–27.
- [3] Afra Feyza Akyürek, Muhammed Yusuf Kocyigit, Sejin Paik, and Derry Tanti Wijaya. 2022. Challenges in measuring bias via open-ended language generation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP'22)*. Association for Computational Linguistics, 76–76.
- [4] Vamsi Aribandi, Yi Tay, and Donald Metzler. 2021. How reliable are model diagnostics? In *Proceedings of the Association for Computational Linguistics (ACL-IJCNLP'21)*. Association for Computational Linguistics, 1778–1785.
- [5] Muhammad Hilmi Asyrofi, Zhou Yang, Imam Nur Bani Yusuf, Hong Jin Kang, Ferdian Thung, and David Lo. 2021. Bi-asFinder: Metamorphic test generation to uncover bias for sentiment analysis systems. *IEEE Transactions on Software Engineering* 48, 12 (2021), 5087–5101.
- [6] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *Proceedings of the 9th Annual Conference of the Special Interest Group for Computing, Information and Society*.

- [7] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org.
- [8] Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT’s gender bias. In *Proceedings of the 2nd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 1–16.
- [9] Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurltlex: A multilingual lexicon of words to hurt. In *Proceedings of the CEUR Workshop*, Vol. 2253. Accademia University Press.
- [10] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. Retrieved from <https://arXiv:1810.01943>.
- [11] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT’21)*, 610–623.
- [12] Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? Evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 6923–6933.
- [13] Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. 2022. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 806–822.
- [14] Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good secretaries, bad truck drivers? Occupational gender stereotypes in sentiment analysis. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 62–68.
- [15] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. Retrieved from <https://arXiv:2110.01963>.
- [16] Abeba Birhane, Vinay Uday Prabhu, and John Whaley. 2022. Auditing saliency cropping algorithms. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4051–4059.
- [17] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “Bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.
- [18] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 1004–1015.
- [19] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc.
- [20] Samuel Bowman. 2022. The dangers of underclaiming: Reasons for caution when reporting how NLP systems fail. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 7484–7499.
- [21] Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. How Conservative are Language Models? Adapting to the Introduction of Gender-Neutral Pronouns. Retrieved from <https://arXiv:2204.10281>.
- [22] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. Retrieved from <https://arXiv:2005.14165>.
- [23] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. PMLR, 77–91.
- [24] Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, London/New York.
- [25] Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. 2022. Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics. Retrieved from <https://arXiv:2206.03390>.

- [26] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.
- [27] Yang Trista Cao and Hal Daumé, III. 2021. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle. *Comput. Ling.* 47, 3 (Nov. 2021), 615–661.
- [28] Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-Grounded Measurement of U.S. Social Stereotypes in English Language Models. Retrieved from <https://arXiv:2206.11684>.
- [29] Serina Chang and Kathy McKeown. 2019. Automatically inferring gender associations from language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 5746–5752.
- [30] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Models. Retrieved from <https://arXiv:2202.04053>.
- [31] Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 173–181.
- [32] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (Apr. 2020), 82–89.
- [33] Greville G. Corbett. 2013. *The Expression of Gender*. De Gruyter, Inc., Berlin/Boston.
- [34] Kate Crawford. 2017. The Trouble with Bias - NIPS 2017 Keynote - Kate Crawford #NIPS2017. https://www.youtube.com/watch?v=fMym_BKWQzk
- [35] Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Trans. Assoc. Comput. Ling.* 9 (Nov. 2021), 1249–1267.
- [36] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*’19)*. Association for Computing Machinery, 120–128.
- [37] Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2021. Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. Retrieved from <https://arXiv:2112.07447>.
- [38] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1968–1994.
- [39] Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of “Gender” in NLP Bias research. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT’22)*.
- [40] Jacob Devlin, Ming-Wei Chang, Lee Kenton, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT’19)*. 4171–4186.
- [41] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT’21)*. Association for Computing Machinery, New York, NY, 862–872.
- [42] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: A multilingual speech translation corpus. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2012–2017.
- [43] Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP’20)*. Association for Computational Linguistics, Online, 314–331.
- [44] Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 147–154.
- [45] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1696–1705.
- [46] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. 2022. A survey on bias in visual datasets. *Comput. Vision Image Understand.* 223 (Oct. 2022), 103552.
- [47] Aparna Garimella, Akhsh Amarnath, Kiran Kumar, Akash Pramod Yalla, Anandhavelu N, Niyati Chhaya, and Balaji Vasani Srinivasan. 2021. He is very intelligent, she is very beautiful? On mitigating social biases in language

modelling and generation. In *Proceedings of the Association for Computational Linguistics (ACL-IJCNLP'21)*. Association for Computational Linguistics, 4534–4545.

- [48] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 1926–1940.
- [49] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 609–614.
- [50] Maria J. Grant and Andrew Booth. 2009. A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Inf. Libraries J.* 26, 2 (2009), 91–108.
- [51] Ben Green. 2019. Good” isn’t good enough. In *Proceedings of the AI for Social Good Workshop at NeurIPS*, Vol. 17.
- [52] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *J. Personal. Soc. Psychol.* 74, 6 (1998), 1464. Publisher: American Psychological Association.
- [53] Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 122–133.
- [54] Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-Debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1012–1023.
- [55] Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reflection in Arabic. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 155–165.
- [56] Marlis Hellinger and Hadumod Bussmann. 2003. *Gender Across Languages: The Linguistic Representation of Women and Men*. Vol. 11. J. Benjamins, Amsterdam/Philadelphia.
- [57] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 771–787.
- [58] Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of large language models in understanding pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 5352–5367.
- [59] Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 752–762.
- [60] Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. “You Sound Just Like Your Father” commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1686–1690.
- [61] Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Lang. Linguist. Compass* 15 (Aug. 2021).
- [62] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (Un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*’19)*. Association for Computing Machinery, New York, NY, 49–58.
- [63] Abigail Jacobs, Su Blodgett, Solon Barocas, III Daumé, Hal, and Hanna Wallach. 2020. The meaning and measurement of bias: Lessons from natural language processing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 706–706.
- [64] Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1548–1558.
- [65] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3819–3828.
- [66] Os Keyes, Chandler May, and Annabelle Carrell. 2021. You keep using that word: Ways of thinking about gender in computing research. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (Apr. 2021), 39:1–39:23.
- [67] Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, 43–53.

- [68] Alex Koch, Roland Imhoff, Ron Dotsch, Christian Unkelbach, and Hans Alves. 2016. The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *J. Personal. Soc. Psychol.* 110, 5 (May 2016), 675–709. Publisher: American Psychological Association.
- [69] Arvindkumar Krishnakumar, Viraj Prabhu, Sruthi Sudhakar, and Judy Hoffman. 2021. Udis: Unsupervised discovery of bias in deep visual recognition models. In *Proceedings of the British Machine Vision Conference (BMVC'21)*, Vol. 1. 3.
- [70] James Kuczumski. 2018. Reducing gender bias in Google Translate. <https://blog.google/products/translate/reducing-gender-bias-google-translate/>
- [71] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*. 166–172.
- [72] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. Retrieved from <https://arXiv:1909.11942>.
- [73] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before Fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., 9694–9705.
- [74] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5502–5515.
- [75] Tomasz Limisiewicz and David Mareček. 2022. Don't forget about pronouns: Removing gender bias in language models without losing factual gender information. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP'22)*. Association for Computational Linguistics, 17–29.
- [76] Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. 923–929.
- [77] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*. Springer, Berlin, 189–202.
- [78] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2024. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems* 36 (2024).
- [79] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 142–150.
- [80] Abhishek Mandal, Susan Leavy, and Suzanne Little. 2021. Dataset diversity: Measuring and mitigating geographical bias in image search and retrieval. In *Proceedings of the 1st International Workshop on Trustworthy AI for Multimedia Computing (Trustworthy AI'21)*. Association for Computing Machinery, 19–25. <https://doi.org/10.1145/3475731.3484956>
- [81] Abhishek Mandal, Susan Leavy, and Suzanne Little. 2023. Biased attention: Do vision transformers amplify gender bias more than convolutional neural networks? In *Proceedings of the 34th British Machine Vision Conference (BMVC'23)*. BMVA. Retrieved from <https://papers.bmvc2023.org/0629.pdf>.
- [82] Abhishek Mandal, Susan Leavy, and Suzanne Little. 2023. Measuring bias in multimodal models: Multimodal composite association score. In *Proceedings of the International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 17–30.
- [83] Abhishek Mandal, Suzanne Little, and Susan Leavy. 2023. Multimodal bias: Assessing gender bias in computer vision models with nlp techniques. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 416–424.
- [84] Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. 5267–5275.
- [85] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 622–628.
- [86] Sally McConnell-Ginet. 2013. Gender and its relation to sex: The myth of “natural” gender. In *The Expression of Gender*, Greville G. Corbett (Ed.). De Gruyter Mouton, 3–38.
- [87] Nicholas Meade, Elinor Poole-Dayana, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1878–1898.

- [88] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *Comput. Surveys* 54, 6 (July 2021), 115:1–115:35.
- [89] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–229.
- [90] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 5356–5371.
- [91] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*. Association for Computational Linguistics, 1953–1967.
- [92] Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2398–2406.
- [93] Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring harmful sentence completion in language models for LGBTQIA+ Individuals. In *Proceedings of the 2nd Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics, 26–34.
- [94] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. Association for Computational Linguistics, 1532–1543.
- [95] Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 507–511.
- [96] Marcelo O. R. Prates, Pedro H. Avelar, and Luís C. Lamb. 2020. Assessing gender bias in machine translation: A case study with Google Translate. *Neural Comput. Appl.* 32, 10 (May 2020), 6363–6381.
- [97] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*. PMLR, 8748–8763.
- [98] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. Retrieved from <https://arxiv.org/abs/2204.06125>.
- [99] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 7237–7256.
- [100] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 8–14.
- [101] Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the 2nd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 35–43.
- [102] Yash Savani, Colin White, and Naveen Sundar Govindarajulu. 2020. Intra-processing methods for debiasing neural networks. *Adv. Neural Inf. Process. Syst.* 33 (2020), 2798–2810.
- [103] Yash Savani, Colin White, and Naveen Sundar Govindarajulu. 2020. Intra-processing methods for debiasing neural networks. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 2798–2810.
- [104] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Trans. Assoc. Comput. Ling.* 9 (2021), 845–874.
- [105] Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Trans. Assoc. Comput. Ling.* 9 (2021), 1408–1424.
- [106] Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. 2021. A step toward more inclusive people annotations for fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 916–925.
- [107] Carsten Schwemmer, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart. 2020. Diagnosing gender bias in image recognition systems. *Socius* 6 (2020), 2378023120967171.
- [108] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 618–626.

- [109] Ignacio Serna, Alejandro Pena, Aythami Morales, and Julian Fierrez. 2021. InsideBias: Measuring bias in deep networks and application to face gender biometrics. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR'21)*. IEEE, 3720–3727.
- [110] Ignacio Serna, Alejandro Peña, Aythami Morales, and Julian Fierrez. 2021. InsideBias: Measuring bias in deep networks and application to face gender biometrics. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR'21)*. 3720–3727.
- [111] Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5248–5264.
- [112] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *Proceedings of the Workshop on Machine Learning for the Developing World (NIPS'17)*.
- [113] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How much can clip benefit vision-and-language tasks? Retrieved from <https://arXiv:2107.06383>.
- [114] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 3407–3412.
- [115] Vivek K. Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. 2020. Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms. *J. Assoc. Inf. Sci. Technol.* 71, 11 (2020), 1281–1294.
- [116] Vivek K. Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. 2020. Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms. *J. Assoc. Inf. Sci. Technol.* 71, 11 (2020), 1281–1294.
- [117] Kirill Sirotkin, Pablo Carballeira, and Marcos Escudero-Viñolo. 2022. A study on the distribution of social biases in self-supervised learning visual models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10442–10451.
- [118] Tejas Srinivasan and Yonatan Bisk. 2022. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP'22)*. Association for Computational Linguistics, 77–85.
- [119] Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. Retrieved from <https://arXiv:2112.14168>.
- [120] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1679–1684.
- [121] Ryan Steed and Aylin Caliskan. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 701–713.
- [122] Ryan Steed and Aylin Caliskan. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT'21)*. Association for Computing Machinery, New York, NY, 701–713.
- [123] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2021. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. <https://openreview.net/forum?id=SygXPaEYvH>
- [124] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1630–1640.
- [125] Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral English. Retrieved from <https://arXiv:2102.06788>.
- [126] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark D. M. Leiserson, and Adam Tauman Kalai. 2019. What are the Biases in my word embedding?. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. Association for Computing Machinery, New York, NY, 305–311.
- [127] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. 2021. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference (WWW'21)*. Association for Computing Machinery, New York, NY, 633–645.
- [128] Eddie Ungless, Bjorn Ross, and Anne Lauscher. 2023. Stereotypes and smut: The (Mis)representation of non-cisgender identities by text-to-image models. In *Proceedings of the Association for Computational Linguistics (ACL'23)*. Association for Computational Linguistics, 7919–7942.

- [129] Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 8940–8948.
- [130] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 3003–3008.
- [131] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It’s a man’s Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*.
- [132] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. 2022. REVISE: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision* 130, 7 (2022), 1790–1810.
- [133] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. 2022. REVISE: A tool for measuring and mitigating bias in visual datasets. *Int. J. Comput. Vision* 130, 7 (July 2022), 1790–1810.
- [134] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5310–5319.
- [135] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8919–8928.
- [136] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Trans. Assoc. Comput. Ling.* 6 (Dec. 2018), 605–617.
- [137] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. Measuring and Reducing Gendered Correlations in Pre-trained Models. Retrieved from <https://arXiv:2010.06032>.
- [138] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 547–558.
- [139] Zhifei Zhang, Yang Song, and Hairong Qi. 2017. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5810–5818.
- [140] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2979–2989.
- [141] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 15–20.
- [142] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 4847–4853.
- [143] Kankan Zhou, Eason Lai, and Jing Jiang. 2022. VLStereoSet: A study of stereotypical bias in pre-trained vision-language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 527–538.
- [144] Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell. 2022. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10410–10421.
- [145] Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1651–1661.

Received 15 September 2022; revised 30 April 2024; accepted 22 August 2024