Artificial Intelligence Review (2025) 58:370
https://doi.org/10.1007/s10462-025-11375-8

A systematic review of bias detection methods for non-English word embeddings and language models

Alexandre Puttick[1] · Catherine Ikae[1] · Carlotta Rigotti[2] · Eduard Fosch-Villaronga[2] · Mark W. Kharas[3] · Roger A. Søraa[3] · Mascha Kurpicz-Briki[1]

Received: 17 July 2024 / Accepted: 31 August 2025 / Published online: 8 October 2025

Abstract
Biases in applications of machine learning and artificial intelligence are a major limitation of these applications. Stereotypes of the society are reflected in different types of applications, including image generation, machine translation or CV ranking. This is in particular also the case for language models and word embeddings, encoding human language as mathematical vectors. Research addressing the challenging problem of detection (and mitigation) of the bias in these embeddings is often conducted for the English language. However, the stereotypes encoded can be language dependent and impacted by a cultural environment. Thus, dedicated research efforts for languages other than English are required. In this paper, we conduct a systematic literature review to identify and compare existing bias detection methods for non-English word embeddings and language models. In an interdisciplinary team we examine the technical aspects, as well as the definitions of bias used by researchers in the field. Based on our findings, we outline a research plan for making bias detection in the field of NLP more inclusive for languages other than English.

Keywords Natural language processing · Bias · Machine learning · Word embeddings · Language models

1 Introduction

In this paper, we present the results of a systematic literature review to identify and compare existing bias detection methods for non-English word embeddings and language models, focusing on the European region. Our efforts provide researchers with a broad overview of the methods applied in the non-English setting, their limitations, and the various language- and culture-specific issues that arise in the process. Our results are two-fold, addressing the various dimensions and descriptions of bias in the surveyed papers, in addition to describing state-of-the-art technical methods in detail.

Defining bias

Extended author information available on the last page of the article

In the context of AI systems, the term *bias* can have manifold meanings and may not be explicitly defined. Notions of bias on the societal level, already nuanced, collide with bias in the context of statistics and data science. As such, a concentrated effort is made in this survey to define our interpretation of bias, to examine how the term is understood in the collected literature, and to understand how the various types of bias relate to each other (Sect. 4).

The literature offers no clear-cut definition or classification of bias, highlighting the complexity and often contentious nature of the concept. While bias was historically understood in neutral terms—as a mere deviation from a standard (Danks and London 2017)—such deviations frequently involve value-laden, normative judgments that reflect societal structures of privilege and oppression, influencing individual reasoning, actions, and well-being.

To address these nuances, this study adopts as a focal point the term *diversity bias*, anchoring it within the framework of EU anti-discrimination law to ensure terminological clarity. This approach defines bias as the unfair positive or negative treatment of individuals, primarily based on protected grounds, thereby aligning the concept with established legal understandings of discrimination. These grounds include sex, race, color, ethnic or social origin, genetics, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age, or sexual orientation (Article 21 of the Charter of the Fundamental Rights of the European Union).

While the EU socio-legal classification provides a well-established foundation, we assert that the understanding of diversity bias in AI systems must extend beyond this classification in three crucial ways:

First, *AI-driven diversity bias is not confined to traditional power asymmetries and may encompass social groups without current legal protection*, such as single parents, or through proxy characteristics not easily identifiable by humans, such as mouse movement (Wachter et al. 2021).

Second, *we recognize the significance of intersectional discrimination and exclusion*. Intersectional discrimination refers to situations where an individual faces a unique form of cumulative discrimination based on multiple interacting personal characteristics (Crenshaw 1989; Fosch-Villaronga et al. 2022). While already pervasive in the non-digital world, AI applications are anticipated to amplify and introduce new forms of intersectional discrimination through datasets structured along intersecting axes of social inequality (Ovalle et al. 2023; Ireni-Saban and Sherman 2020).

Third, we posit that *a constructive approach to addressing diversity bias necessitates the technical community's openness to viewing values, principles, and rights as endogenous, bridging to the realms of social sciences and humanities (SSH)* (Mulligan et al. 2019). Consequently, this survey endorses SSH knowledge transfer in STEM papers, particularly in terms of methodologies, the consideration of language as a social construct, and a cross-disciplinary understanding of fairness.

Bias in NLP systems

The issue of diversity bias in machine learning and artificial intelligence (AI) has become a major ethical barrier to the application of these powerful technologies. In the field of natural language processing (NLP), the issue of harmful encoded bias and stereotypes has become apparent to the broad public since the advent of ChatGPT and similar services. Hovy and Prabhumoye (2021) highlight five dimensions along the machine learning pipeline in which bias can be introduced: the data, the annotation process, the input represen-

tations, the models, and the research design. In NLP, input representations often come in the form of so-called word embeddings, which encode human language as mathematical vectors. This can be accomplished by constructing a map sending a single word to a single context-independent vector, as in the case of *static word embeddings*, e.g., word2vec (Mikolov et al. 2013), GloVe (Pennington et al. 2014) and fasttext (Bojanowski et al. 2016; Grave et al. 2018). More recent developments lead to *contextual word embeddings*, more nuanced vectorizations that depend on a given word's context. These can be extracted from recent language models based on transformer architecture (Vaswani et al. 2017), for example, Google's BERT (Devlin et al. 2018) or the GPT models (Radford et al. 2018). This new generation of language models has led to significant improvements on many NLP tasks, but they are known to encode social stereotypes, potentially leading to bias and unfairness in downstream applications (e.g., Bolukbasi et al. 2016; Caliskan et al. 2017; Ahn and Oh 2021).

A large body of work has been developed to detect (and partially mitigate) these biases. These efforts have been summarized in various literature reviews (Sun et al. 2019; Meade et al. 2022; Delobelle et al. 2022). An early detection and mitigation method was developed by Bolukbasi et al. (2016) and relied on attempting to delete gender information from word embeddings by contracting the vector subspace corresponding to gender. Another well-known method for bias detection in static word embeddings was proposed by Caliskan et al. (2017). Drawing from the implicit association test (IAT) (Greenwald et al. 1998) used in psychology to detect implicit bias in human subjects, the method was adapted for and applied to static word embeddings. Similar methods exist for contextualized word embeddings, as showcased, for example, in Kurita et al. (2019), Guo and Caliskan (2021) and Ahn and Oh (2021). Various studies have identified significant limitations to existing state-of-the-art methods. Of particular note is the evidence that detecting and mitigating bias in word embeddings and language models does not always correlate strongly with detection and mitigation in downstream tasks (Goldfarb-Tarrant et al. 2020; Delobelle et al. 2022).

Despite this extensive body of work, research addressing the detection of bias in word embeddings and language models is significantly skewed towards the English language. The lack of language diversity in NLP research in general is a source of criticism (Joshi et al. 2020). This poses a real problem because research indicates that (encoded) stereotypes are often language-dependent and impacted by cultural context, as evidenced both in psychological research (e.g., Fiske 2017) and in the context of word embeddings (e.g., Kurpicz-Briki and Leoni 2021). Moreover, structural elements such as grammatical gender complicate the extension of existing detection methods to non-English languages. Dedicated research efforts for languages other than English are thus an absolute necessity.

This literature survey summarizes the methods currently applied to detect bias in word embeddings and language models, with a focus on the challenges that arise from different languages and their corresponding cultural and linguistic properties. In many cases, a simple translation of word lists or sentence templates from existing bias detection methods is not sufficient. The examined work proposes various region- and language-specific adaptations, as well fully novel bias tests. Our findings also summarize the way different studies deal with gendered languages in bias detection.

In a survey of bias in NLP, Blodgett et al. (2020) observe that many studies fail to critically engage with what constitutes bias in the first place, an observation that is corroborated in this work (see Sect. 4). The term *bias* can have many interpretations in the machine learn-

ing context that do not a prior relate to diversity bias, e.g., *measurement bias* or *sampling bias*. However, these forms of bias may still be entangled with diversity bias. Gender, for example, is typically measured as a binary variable, leading to the erasure of non-binary identities from collected data. Similarly, various underprivileged identities may be under-sampled in the training data, leading to inferior model performance on those under-represented groups.

Following the recommendations of Blodgett et al. (2020), which are distilled in Hardmeier et al. (2021), we summarize how bias in NLP systems is generally understood in the surveyed literature, why it can be harmful and who it can be harmful to.

Although the studies in question often fail to explicitly define bias, an implicit definition can be distilled from the technical methods. With only a handful of exceptions (Sentence Completions and Extrinsic Classifiers in Fig. 8), the surveyed works all use methods derived from either Caliskan et al. (2017), Nangia et al. (2020) or Bolukbasi et al. (2016). The methods in Caliskan et al. (2017) and Nangia et al. (2020) aim to quantify the extent to which potentially harmful social stereotypes are intrinsically encoded within word embeddings and language models, comparing the relative distance, for example, between the word vectors for *he* and *she* to the word vector for *business*. In other words, *a model is understood to be biased if it encodes harmful social stereotypes*. On the other hand, while Bolukbasi et al. (2016) also attempt to quantify bias encoded in word embeddings, the DirectBias metric is blind to whether or not the measured biases are stereotypical. For example, in the context of occupations, DirectBias only measures the aggregate gender bias in the embeddings, but is not sensitive to whether the word, e.g., *nurse* is biased in the male or female direction. This implies a weaker notion of bias that is decoupled from harmful stereotypes; *a model is only considered biased if it deviates from complete (gender) neutrality*.

These technical methods pertain to the intrinsic worldview learned by the language model in question, but do not engage directly with potential harmful outcomes in downstream tasks. As an example of how one affects the other, consider an NLP hiring system in which similarity between a job ad and a candidate applications plays a role; the use of a model encoding stereotypical occupational associations could lead to harmful outcomes, such as reinforcing the under-representation of women in tech positions.

A limited number of the studies surveyed in this work attempt to measure bias directly within AI-generated text. These methods are included within the scope of this work because text completion is the primary training task for the models in question and thus intimately linked with intrinsic bias in the model. This line of study has grown in significance, as the societal scope and influence of AI-generated text has been expanding rapidly. Moreover, the profusion of proprietary models pushes researchers from outside institutions to develop methods that are applicable to black-box systems in which only text prompts and generated responses are available.

A further discussion on the position taken with respect to the concept of bias in this work and the surveyed literature is presented in Sect. 4.

Research questions and survey structure

In this systematic literature survey following the PRISMA framework (Page et al. 2021), we collect work on bias detection in non-English word embeddings and language models, and compare them according to the following research questions (RQ):

Organizational aspects and languages (RQ1)

- RQ1.1: Geographical distribution
- RQ1.2: Which year was the paper published?
- RQ1.3: How many different languages does the paper address?
- RQ1.4: Which languages are addressed?**Dimensions of biases (RQ2)**

- RQ2.1: What types of bias have been covered?
- RQ2.2: How is the term bias defined?
- RQ2.3: Are forms of intersectional bias covered?**Technical methods and models (RQ3)**

- RQ3.1: Which bias detection methods have been developed and applied to non-English word embeddings and language models?
- RQ3.2: What are the particular challenges identified when adapting methods developed for English to other languages?
- RQ3.3: How were these challenges overcome?

To address the interdisciplinary research questions, the work is carried out by interdisciplinary team, involving experts from the fields of computer science, social sciences and humanities, and law.

The paper is structured as follows: We first describe the methods used for the systematic literature review including information sources, search and selection strategy as well as the interdisciplinary approach. We first discuss the results concerning organizational aspects and languages used by the selected papers. We then discuss the non-technical results of our literature survey, providing an overview of the collected papers and discussing the different dimensions of bias. In the next section we give an introduction to the technical methods and models, and compare the selected papers in technical terms. Finally, we conclude the paper with a Discussion including an outlook for future work.

2 Methods

The systematic literature survey conducted in this paper followed the PRISMA framework (Page et al. 2021). The following subsections discuss the eligibility criteria, information sources and the collection and selection processes.

2.1 Eligibility criteria

We define the following inclusion criteria for surveyed papers:

- Paper provides new bias detection and/or mitigation methods for word embeddings/language models
- Paper makes an empirical or survey study, or discusses limitations of bias detection and/or mitigation methods for word embeddings/language models
- Paper includes at least 1 language other than English

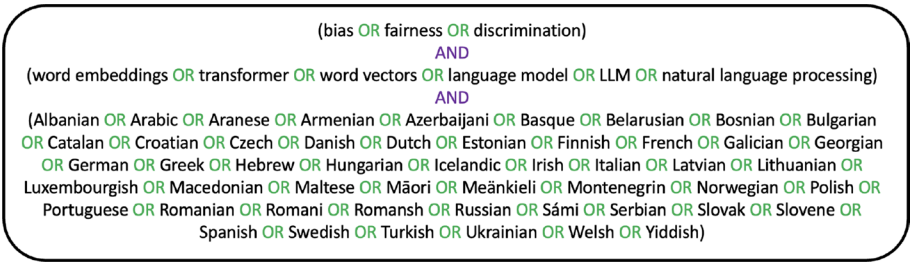

Fig. 1 Search query that was executed on title and abstract on the different databases

2.2 Information sources

We searched IEEE Explore, the ACM Digital Library and the Anthology of the Association for Computational Linguistics (ACL), as we consider these the main venues and conferences relevant to the topics of interest. Additionally, a search on Google Scholar was conducted to complete the results. Once the results of the queries were filtered according to the exclusion criteria, we further collected direct citations from the remaining papers that met the inclusion criteria.

2.3 Search strategy and data collection process

The following query was designed for this systematic literature review:

The query was designed to consider on one hand the relevant aspects of fairness, discrimination or bias, and on the other hand the particular natural language processing technologies of interest in this study. We consider papers published from 2016 on, as this marks the year the first foundational paper in this research field was published (Bolukbasi et al. 2016). Our work aims to select languages relevant for the European context. The languages included in the query are derived from the list of regional members of the United Nations for the groups *Western European and Other States* and *Eastern European States*[1]. Rather than a political construct using EU/EFTA countries as a basis, we view Europe as a geographical region. Based on this list of countries, we compiled a list of official languages using the languages mentioned in the most recent constitutions and/or legal text of the corresponding countries. The resulting languages have an official or co-official status in at least one of the states. Finally, English was excluded from the list. This resulted in the 51 languages displayed in the query in Fig. 1. For ACM and IEEE, the search was conducted by title and abstract, and the query was implemented using the advanced search features of the website. For ACM, two queries were executed to cover the two available database options. Since no such search mask is available for the ACL Anthology, the query was conducted on all fields. Only the first ten pages of results are available with ACL. To complete these results, the query was also executed on Google Scholar, and the results of the first ten pages sorted by relevance were manually examined. Relevant papers were added to the list.

[1] https://www.un.org/dgacm/en/content/regional-groups.

2.4 Selection process

The titles and abstracts of papers collected via the above queries were collected in Excel tables. The titles and abstracts were manually and independently assessed by at least two of the authors with regard to the inclusion criteria. For ACL, links referring to a collection of papers (e.g., proceedings) and papers from demo sessions were excluded (for ACM and IEEE, there were no such cases). Collections were excluded because individual papers of interest within the collections already appeared in the query results, whereas papers from demo sessions did not include sufficient detail or meet the standard of rigor to be considered on par with the other selected studies. Papers which met the inclusion criteria with unanimous agreement were selected, and papers chosen by none of the authors were dropped. The other cases were discussed until a consensus was reached. Finally, any works cited in the initial batch of selected papers that met the selection criteria of the investigation were also included. Based on this, 35 papers were read in full by a least three authors from different disciplinary backgrounds. Three papers were excluded at this step, as they did not include any of the relevant languages, did not contain sufficient detail about bias detection or mitigation, or discussed multimodal models and were therefore not comparable to the other works reviewed in the survey. Figure 2 depicts the paper selection process, resulting in the 32 studies included in this literature review.

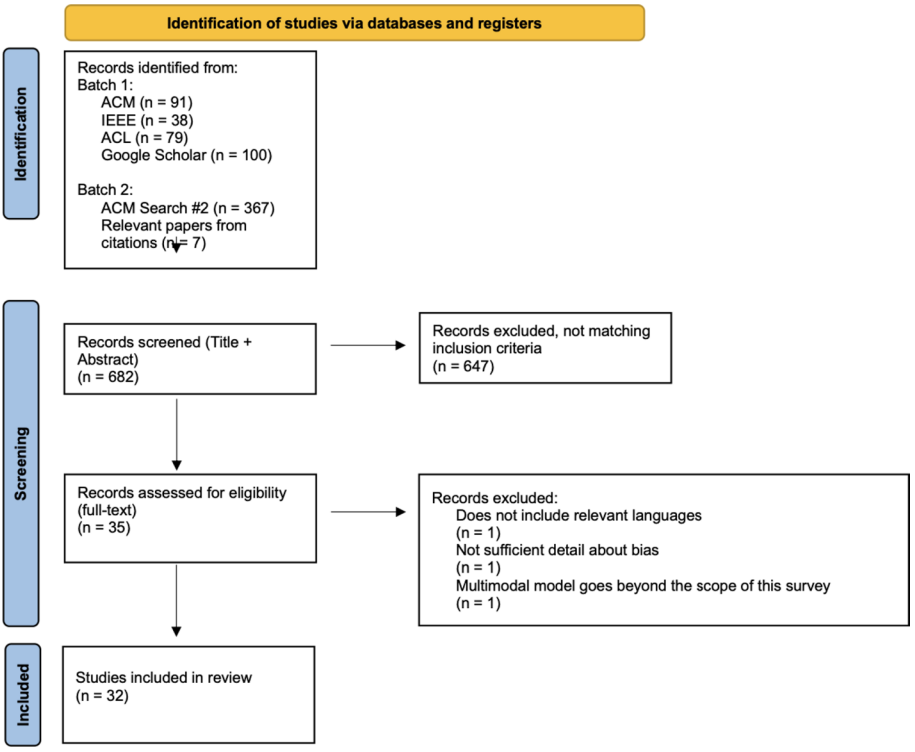

Fig. 2 PRISMA 2020 flow diagram based on the PRISMA framework Page et al. (2021)

2.5 Interdisciplinary collaboration

The 32 papers identified as relevant in the scope of this literature survey were then assessed in detail by the interdisciplinary team of authors. In particular, the three dimensions described in the introduction were examined: organizational aspects and languages, dimensions of bias and technical methods and models. Our interdisciplinary endeavor brought together a team of seven researchers, each contributing unique expertise in areas such as computer science, law, sociology, gender studies, and science and technology studies (STS). We formed three specialized groups—*Group 1 on Demographics*, *Group 2 on SSH*, and *Group 3 on the technical aspects*—strategically aligned with researchers' interests and expertise. The initial phase of our interdisciplinary work involved a constructive dialogue to define and agree upon the scope literature review. This step demanded a delicate balance, considering the diverse perspectives, methodologies, and research interests. Subsequently, each group undertook a comprehensive review of papers, focusing on their respective sub-research questions. To ensure internal consistency and facilitate inter-group understanding, each group created a semi-structured table during the reading and analysis phases summarizing the relevant salient points from each paper. The added value of this interdisciplinary effort lies in the synthesis of diverse knowledge and perspectives, paving the way for a more holistic and nuanced approach to addressing diversity bias in word embeddings and language models. In light of our efforts, it is clear that the understanding and scope of bias addressed by current state-of-the-art technical methods is very limited. Perspectives from other disciplines can thus inform the development of more comprehensive techniques.

3 Organizational aspects and languages

Geographical distribution

The map in Fig. 3 illustrates the geographical distribution of research papers involved in the surveys studies on bias detection. The United States stands out prominently with the highest number of papers, totaling 9 papers, representing approximately 36% of the total. Germany follows closely with 7 research papers, constituting around 28%. The United Kingdom also makes a substantial contribution, with 3 research papers, accounting for

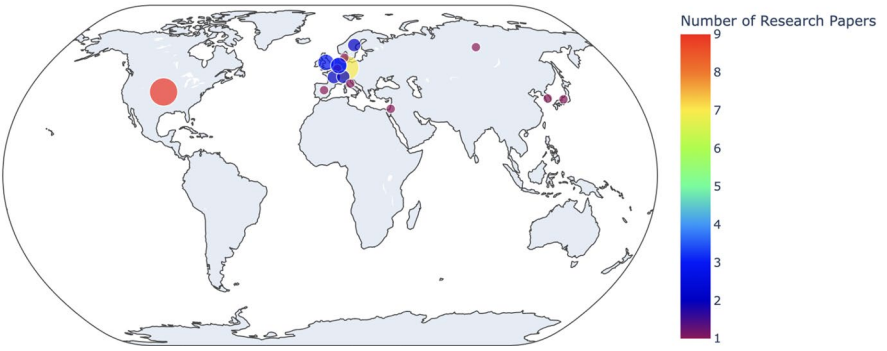

Fig. 3 Author distribution across surveyed papers based on the number of papers per country

about 12%. Noteworthy participation is observed in the Netherlands, Switzerland, Sweden, France, and Spain, each contributing 3, 8, 4, 4, and 4% respectively. Other countries, such as South Korea, Japan, Malta, Belgium, Denmark, Israel, Italy, and Russia, exhibit a more scattered presence, each represented by 1 research paper. This distribution underscores some level of international collaboration in addressing bias in AI systems and language models. Given the focus on the European context, it is not surprising that the majority of research papers are from Europe. However, as is the case with AI research in general, U.S.-based research papers could be considered overrepresented. Moreover, within Europe there is a clear skew towards more wealthy, western European countries (RQ1.1).

Languages

The analyzed papers include a diverse array of languages, as depicted in Fig. 4. German emerges as the most prevalent language, appearing in 16 studies, closely followed by Spanish with 15 instances. Noteworthy representation is also observed for French, Chinese and Arabic, with 10, 8 and 7 occurrences, respectively. The surveyed papers feature further languages such as Italian (5), Portuguese (4), Russian (3), Turkish (4), Japanese (2), and Swedish (3). In addition, single instances are noted for Croatian, Dutch, Romanian, Polish, Farsi, Urdu, Wolof, Danish, Malay, Norwegian, Persian, Finnish, Indonesian, Hebrew, Tagalog, and Hindi, collectively contributing to the rich linguistic variety observed in the analyzed content (RQ1.3)(RQ1.4).

Timeline of reasearch in non-English bias detection

Over the past five years, the landscape of scholarly inquiry into bias detection in non-English word embeddings has evolved. In 2019, six papers marked the initial foray into this field, laying a foundation for subsequent research. The following year, 2020, saw a notable uptick with nine papers, indicating a growing awareness of the implications and need to address bias in word embeddings. This trend continued in 2021 with eight papers, suggesting sustained attention and a refinement of methodologies. The momentum persisted into 2022, with another nine papers advancing understanding and developing more robust frameworks for bias detection. However, 2023 saw a significant drop to just one paper. This

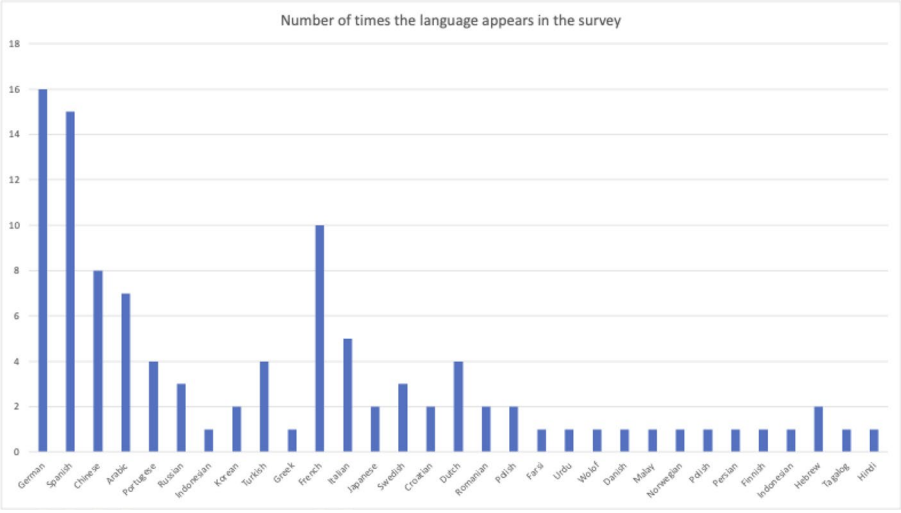

Fig. 4 Languages, aside from English, that are found in the examined papers

does not necessarily signal a shift in research focus or temporary lull. For one, it is possible that more recent work was not yet included in the queried databases at the time studies were collected. Moreover, 2023 marked the release of non-open source LLMs generating enormous amounts of public attention such as GPT-3 and its successors. The proprietary nature of these models greatly obstructs the application of existing bias detection methods and necessitates new directions of research for bias detection in closed-source models beyond the scope of this literature review (RQ1.2).

4 Dimensions of biases

Our analysis advances a common understanding of diversity bias in word embeddings and highlights specific gaps that need addressing, notably the limited attention given to most grounds of discrimination beyond gender. Additionally, it emphasizes the growing trend of interdisciplinarity, where the intersection of SSH with computational frameworks becomes increasingly critical in tackling the complex challenges presented by diversity biases in language representation.

Defining bias: relating the human, AI and legal standpoints

According to the Cambridge Dictionary, the term *bias* encompasses three meanings: allowing personal opinions to influence one's judgment in an unfair way, showing preference towards someone or something, and presenting incorrect information due to flawed collection or presentation methods. In the context of this literature review, all of these definitions hold some relevance. Unfair prejudices or preferences can be encoded into language models, reflected, for example, in a higher probability that a given model completes the prompt *A person from Iran is an...* with the word *enemy* in comparison the prompt *A person from the U.S. is an...*, while sampling bias can lead to under-representation of low-resource languages in training data and, consequently, to poor model performance when applied to such languages.

As stated in the introduction, bias can permeate data collection, annotation, model training, and research methodologies, significantly impacting the development and application of AI systems. While training data is often cited as the main culprit for encoded bias (garbage in, garbage out), it is important to note that the individuals' unique backgrounds, personal characteristics, experiences, and fundamental beliefs, can contribute to model bias along every step of the machine learning pipeline. Bias influences the processing of thoughts and implicit emotions, reinforcing existing attitudes, and perpetuating actions that often stereotype others in prejudicial and inaccurate ways, resulting in subjective preferential treatment towards certain entities to the detriment of others. Without counteractive measures, biased AI systems are poised to reproduce the same unfair actions.

Biases are inherently normative, i.e., they influence our behavior and moral judgments (Blodgett et al. 2020), are deeply rooted in human behavior and social dynamics (Risman 2018), and can harm various social groups for diverse reasons. *None of the reviewed papers offers a normative definition of bias* (RQ2.2), with Pestova (2021) at most defining bias as an *unfair regularity in training data and the model itself* and Sahlgren and Olsson (2019) defining bias as the opposite of fairness. We reference the legal framework of anti-discrimination law in the European Union to ground normative considerations in this literature review: Based on Article 21 of the Charter of Fundamental Rights of the European

Union,"[a]ny discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited." We refer to these characteristics as *protected grounds*, *protected attributes* or *sensitive attributes*, as all three terms appear frequently in the literature.

From a legal standpoint, this would prohibit AI systems from using protected attributes as the grounds for discrimination. EU anti-discrimination law also prohibits *indirect discrimination*—where a seemingly neutral policy has disproportionate effects for individuals from a particular protected group—except in exceptional cases. For example, given that women are typically expected to take on childcare responsibilities, requiring employees to be present at the office from 9am-5pm could discriminate against women. Provisions prohibited indirect discrimination can extend to the phenomenon of *proxy discrimination* in AI systems, wherein a model can infer protected attributes indirectly from data (Xenidis 2020). For example, is it known that language models can be trained to infer the gender of the author of a text sample with high accuracy, even though the author's gender is never explicitly mentioned (Deutsch and Paraboni 2023).

Despite the ambiguity with regard to defining bias in the surveyed studies, it is clear that every paper is concerned with positive or negative stereotypes with respect to particular protected attributes. All of collected works attempt to measure bias according to an array of bias metrics, which are described in detail in Sect. 2. These bias metrics are typically referred to as *intrinsic bias metrics*, because they attempt to measure the degree to which stereotypical associations (e.g., men/women vs. science/arts) are present within word embeddings or language models. This is in contrast to *extrinsic bias metrics*, which attempt to measure bias in downstream tasks. For example, job candidate applications may be converted into numerical data using word embeddings, which are in turn processed in order to predict job suitability. Extrinsic bias metrics could then aim to measure disparities in predictions across sensitive attributes. The bias in word embeddings and language models measured by intrinsic metrics is then analogous to an individual's internalized prejudices, whereas extrinsic metrics measure the analog of how those internal, potentially implicit, prejudices affect that individual's actions.

Just as internal prejudice affects behavior, it would be expected that there is a correlation between the bias intrinsically encoded in word embeddings and language models and the bias demonstrated in downstream tasks. This has been demonstrated in particular cases (Bolukbasi et al. 2016), although the the correlation is not always visible and depends significantly on the combination of intrinsic metric and downstream task (Goldfarb-Tarrant et al. 2020; Delobelle et al. 2022). While a more holistic study of bias would include the intrinsic and extrinsic perspectives, the relation is not well-understood. Since only one of the collected papers considered bias in downstream tasks (Goldfarb-Tarrant et al. 2020), the scope of our survey is limited to intrinsic bias metrics.

Protected grounds included in the collected literature

Table 1 gives an overview of the different sensitive attributes with respect to which bias was analyzed in the collected studies (RQ2.1). We examine the similarities and differences between the formulation of EU protected grounds of discrimination and the dimensions of bias considered in the surveyed literature, including from an intersectional angle (RQ2.3). As a reminder, we use the term *diversity bias* to refer to biases closely linked with personal

Table 1 Main types of bias identified in the analyzed papers

Type of bias	No. of papers	Papers
Gender	31	Alshahrani et al. (2022), Basta et al. (2021), Bartl et al. (2020), Gonen et al. (2019), Kaneko et al. (2022), Katsarou et al. (2022), Kraft et al. (2022), Sahlgren and Olsson (2019), Kurpicz-Briki (2020), Lauscher et al. (2020), McCurdy and Serbetci (2020), Mulsa and Spanakis (2020); Nozza et al. (2021), Zmigrod et al. (2019), Goldfarb-Tarrant et al. (2020), Papakyriakopoulos et al. (2020), Pestova (2021), Sabbaghi and Caliskan (2022), Wagner and Zarrieß (2022), Wambsganss et al. (2022), Zhou et al. (2019), Zhao et al. (2020), Toney-Wails and Caliskan (2020), Lewis and Lupyan (2020), Wevers (2019), Ahn and Oh (2021), Escouflaire et al. (2023), Fort et al. (2022), Matthews et al. (2021), Lauscher and Glavaš (2019), Névéol et al. (2022)
Race	8	Ahn and Oh (2021), Alshahrani et al. (2022), Fort et al. (2022), Sahlgren and Olsson (2019), Lauscher and Glavaš (2019), Lauscher et al. (2020), Névéol et al. (2022), Goldfarb-Tarrant et al. (2020), Wambsganss et al. (2022)
Ethnicity	5	Ahn and Oh (2021), Fort et al. (2022), Sahlgren and Olsson (2019), Kurpicz-Briki (2020), Névéol et al. (2022)
Sexual orientation	4	Fort et al. (2022), Névéol et al. (2022), Nozza et al. (2021), Papakyriakopoulos et al. (2020)
Religion	4	Alshahrani et al. (2022), Fort et al. (2022), Kurpicz-Briki (2020), Névéol et al. (2022)
Socio-economic class	3	Fort et al. (2022), Mulsa and Spanakis (2020), Névéol et al. (2022)
Disability	3	Fort et al. (2022), Mulsa and Spanakis (2020), Névéol et al. (2022)
Age	2	Fort et al. (2022), Névéol et al. (2022)

characteristics that can intersect with axes of social oppression and subordination (Crenshaw 1989; Hill Collins and Bilge 2020).

Our literature review reveals that specific grounds of discrimination received more research attention than others, with gender being the primary focus in 25 papers and mentioned as a potential source of bias in an additional 6 papers. Following gender, racial bias was studied in 9 papers, although the concept of race seemingly overlaps with ethnicity in Ahn and Oh (2021), Fort et al. (2022), Sahlgren and Olsson (2019), Kurpicz-Briki (2020), Névéol et al. (2022). Conversely, some protected grounds, such as property or birth[2], were

[2] Property relates to, e.g., home ownership, whereas birth can relate to, e.g., nationality and citizenship.

not considered. This raises concerns about the creation of an implicit hierarchy of discriminatory grounds within AI bias scholarship, where certain forms of discrimination–in this case, gender and race–are significantly prioritized over other grounds of discrimination. This phenomenon is not unique to AI research; it aligns with broader social sciences and humanities (SSH) literature critiques regarding the establishment of hierarchical discrimination grounds within EU law and the case law of the European Court of Justice. Notably, it is argued that gender and race often receive stronger legal protections compared to other grounds, such as sexual orientation and disability, highlighting systemic disparities in legal frameworks and scholarly attention to discrimination issues (Howard 2006, 2018).

Beyond protected grounds and intersectionality

Some papers discuss diversity biases that go beyond the EU list of protected grounds, addressing issues such as physical appearance–noted in Fort et al. (2022) and Névéol et al. (2022)–and cultural differences–noted in Kurpicz-Briki (2020). In doing so, these analyses further underscore the strong criticism expressed by anti-discrimination law literature about the limitations of the current list of protected grounds in addressing diversity biases, particularly within the context of AI systems (Xenidis 2020; Aloisi 2023; Rigotti and Fosch-Villaronga 2024). On a different note, while physical appearance and cultural differences are treated as specific diversity biases in the collected studies, they could also be seen as proxies for other protected grounds. For example, biased judgments based on physical appearance can be connected to assumptions about someone's age (Kaufmann et al. 2017, 2016). It would be insightful to position physical appearance alongside other diversity biases and protected grounds of discrimination to examine their intersectional interactions. *Intersectionality*, which refers to the interconnected nature of personal characteristics such as gender, age, and sexual orientation within various axes of power, plays a crucial role in shaping unique experiences of discrimination or privilege within society (Crenshaw 1989; Hill Collins and Bilge 2020). Intentionally integrating intersectionality into AI systems could, therefore, make models more reflective and responsive to people's diverse experiences. However, this approach is notably absent from the surveyed literature, except for Sahlgren and Olsson (2019), which explores the intersection of gender and occupation, and Ahn and Oh (2021), which implicitly acknowledges the impact of other personal characteristics.

Ambiguity in the concepts of race and ethnicity

Although the two terms are often used interchangeably, *race* typically refers to physical characteristics that are considered biological, whereas *ethnicity* relates to cultural and ancestral heritage and identity. It is sometimes unclear whether certain terms, such as *population* in Papakyriakopoulos et al. (2020), *cultural differences* in Kurpicz-Briki (2020), and *migration* in Goldfarb-Tarrant et al. (2020) are used with specific meanings and research goals in mind, or if they are employed as synonyms to avoid controversy around terms like *race* and *ethnicity*, which are considered susceptible to social stigmatization and oppression. The concept of race is highly contested—and also differs between languages and regions in terms of the acceptability of the term (Andersson, 2017). This linguistic choice is evident in Névéol et al. (2022) and Fort et al. (2022), where the term *color/ethnicity* is used, which implies the assumption of pigmentocracy (Telles 2014). Reflecting broader political discourse, researchers may express concerns about the use of these terms, fearing that they may inadvertently reinforce racist beliefs and lead to misinterpretations that could adversely affect professional careers (Wilkinson and King 1987; Simon 2015). However, a growing body of guidelines has been published to assist researchers and other stakeholders navigate

these linguistic challenges, recognizing that language evolves and is context-dependent. The concept of race should not be ignored; to effectively mitigate diversity bias, it may sometimes be necessary to gather data on individuals' social positioning and experiences of discrimination based on racial or ethnic origin, and other sensitive categories (High Level Group on, Non-discrimination, Equality and Diversity 2021).

Connecting bias to fairness across disciplines

While we previously highlighted a gap in defining bias, Fort et al. (2022), Kurpicz-Briki (2020), Lauscher and Glavaš (2019), Lauscher et al. (2020), Mulsa and Spanakis (2020), Papakyriakopoulos et al. (2020), Pestova (2021) make additional references to the notion of fairness. These works occasionally present or endorse specific definitions, emphasizing a lack of consensus and acknowledging the elusive nature of defining fairness. Notably, the multidisciplinary construction of fairness evident in these works already indicates knowledge transfer from SSH, reporting similar main arguments, ambitions, and challenges as those within SSH literature.

At the same time, recent interdisciplinary research elucidates the large gap between, e.g., notions of algorithmic bias and fairness and legal notions of discrimination and equality (Weerts et al. 2023). Importantly, algorithmic fairness metrics are primarily concerned with *group fairness* with respect to distributional outcomes with respect to protected attributes. In NLP, this form of fairness relates to both the intrinsic bias of language models as well as extrinsic bias in downstream tasks. In the latter, we may understand bias in terms of observed outcome inequalities, whereas fairness pertains to which outcome distribution is considered ethical. For example, an AI recruitment algorithm may demonstrate a bias towards recommending male over female job candidates, whereas the fairness goal may be complete gender parity.

Multidisciplinary collaboration highlights the connection between diversity biases in AI applications, various social constructs (e.g., language, gender) and social asymmetries of power. Such collaboration is promising and should be further endorsed to advance a nuanced understanding of diversity biases in AI systems and avoid falling into the trap of technical fixes that operate, e.g., on the level of optimizing fairness metrics but not necessarily in alignment with those metrics' ethical underpinnings. A more nuanced and collaborative understanding enables the design of AI systems and mitigation solutions that are inclusive and responsive to social needs.

5 Technical methods and models

5.1 Static and contextual word embeddings

5.1.1 Static word embeddings (SWEs)

It is very non-trivial task to translate the meaning of a word into numbers that can then be processed by a computer. The invention of *word embeddings* streamlined automated text processing methods and thus the field of natural language processing.

Each word from the language in question is encoded as a mathematical vector of fixed size, on the order of several hundred dimensions. These dimensions aim to capture the semantics of the word in question. Word embeddings (also referred to as *word vectors*)

are constructed so that the embeddings of words with similar meaning are close together. For example, the word vectors for *cat* and *dog* would be relatively close to each other, but further away from the word vector for *thunderstorm*. A given word is always mapped to the same vector. Word embeddings with this property are referred to as *static word embeddings*, in contrast to contextual word embeddings, which are introduced below.

For mathematical reasons, the distance between word vectors is not measured in terms of the usual Euclidean distance. Instead, the similarity between two words is measured using *cosine similarity*, which measures the extent to which the corresponding vectors point in the same direction. Given two words w and v and their word vectors $\vec{w}$ and $\vec{v}$, the similarity between w and v in given by $\cos(\vec{w}, \vec{v})$. Figure 5 illustrates simplified two-dimensional word embeddings, where semantic similarity between words corresponds to cosine similarity between the corresponding word embeddings.

Using the relations between word vectors and computing the similarity metrics, powerful applications can be developed: one can compute the best matching answer to a chatbot question or identify relevant documents for a given search query. For a more detailed introduction to this topic for a non-technical audience, the authors refer to Kurpicz-Briki (2023). Word embeddings are usually available out-of-the-box and are freely available. Common word embeddings are word2vec (Mikolov et al. 2013), GloVe (Pennington et al. 2014) and, particularly in the context of Non-English languages, fasttext (Bojanowski et al. 2016; Grave et al. 2018) which is available in several languages.

5.1.2 Contextual word embeddings (CWEs)

The static word embeddings described in the previous section have the disadvantage that they do not provide any contextual information for the corresponding words; the static word embedding of the word *orange* does not distinguish between the contexts *I eat an orange,* and *My shirt is orange.* This is not the case for *contextual word embeddings (CWEs)*, which would assign a different vector to the word *orange* given different contexts.

The main technological breakthrough that allowed words to be encoded into high quality CWEs was the invention of *transformers* (Vaswani et al. 2017). A transformer is a particu-

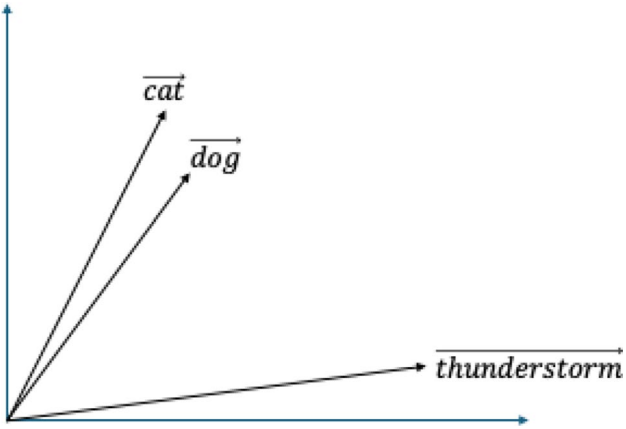

Fig. 5 A simplified depiction demonstrating the notion of word vectors. The cosine similarity $\cos(\overrightarrow{cat}, \overrightarrow{dog})$ is larger than $\cos(\overrightarrow{cat}, \overrightarrow{thuderstorm})$ and $\cos(\overrightarrow{dog}, \overrightarrow{thunderstorm})$

lar type of neural network architecture that accepts sequential input (such as a sentence) and applies an *attention mechanism*, which allows the model to encode relevant contextual information into word vectors. For instance, given the input text *She is a skilled programmer, he is a farmer*, the model outputs an embedding vector for each word in the sentence. The attention mechanism has the effect that the embedding for the word *programmer* places more weight on relevant context words like *she* and *skilled*, while de-emphasizing irrelevant context.

Masked language models (MLMs)

Some transformer models, such as BERT (Devlin et al. 2018), are trained on the task of *masked token prediction*, i.e., given a sentence such as [MASK] *is a programmer,* the model should predict likely words that would appear in place of the [MASK] token. Language models trained on such tasks are called *Masked Language Models (MLMs).* Given a pre-trained MLM, there are several ways to obtain a CWE associated to a given word. For example, given any input sequence, e.g., *John likes art,* the last hidden layer of the BERT model outputs a vector for every word in the sentence. To obtain a CWE for the word *art*, one could simply take the corresponding vector. However, this is not the only option. BERT was trained on an additional task, *next sentence prediction*, i.e., given two sentences, predict the likelihood that the second directly follows the first. For this, an additional tokens were added to the model input. The [CLS] token is added to the beginning of every input sequence, and the [SEP] token is inserted at the end of sentences. After training, the last hidden layer output of the [CLS] token captures the semantics of the entire input sequence, and is therefore often used when we are interested in the sentence-level meaning of the input. This is done, for example, in the Sentence-Embedding Association Test (SEAT) (May et al. (2019), see Sect. 5.2.1 below), where a variety of CWEs for a word, such as *art*, are obtained by taking the vector associated to the [CLS] token for input sentences such as *This is art*, *That is art* etc. Figure 6 depicts CWEs and predictions associated to an MLM.

Causal language models (CLMs)

Causal language models (CLMs) such as those in the GPT series (Radford et al. 2018) are another variant of transformer-based language model. Instead of being trained on the task of masked token prediction, CLMs are trained on *next word prediction*. Consider an input sequence s of text, e.g., $s =$ *She is a...* For every word w in the model's vocabulary, the CLM estimates the probability $p(w|s)$ that w is the next word in the sequence. This training objective makes CLMs ideal for text generation. By sampling among the most probable next words, one can extend the input sequence, e.g., *She is a programmer...* Iterating this process

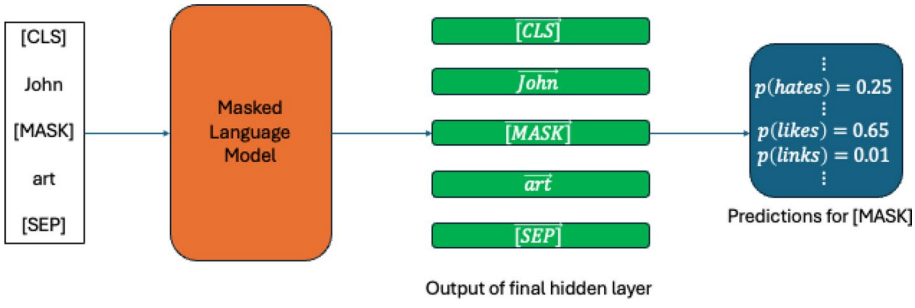

Fig. 6 Illustration of the input and output of an MLM. A sequence of text is converted into a sequence of vectors, which are then used to make predictions for the [MASK] token

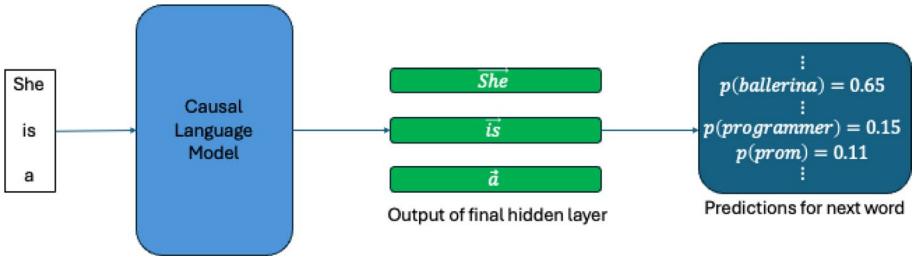

Fig. 7 Illustration of the input and output of a CLM. A sequence of text is converted into a sequence of vectors, which are then used to make predictions for the next word in the sequence

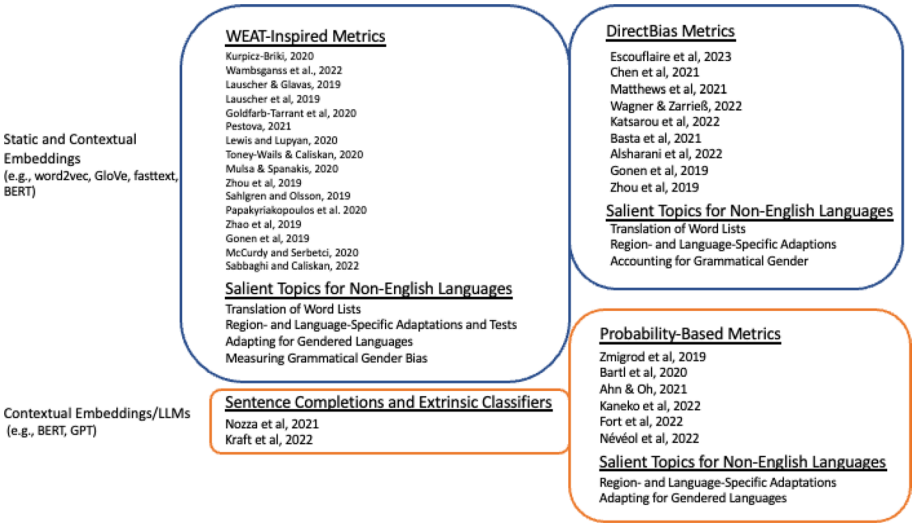

Fig. 8 Grouping of papers by technical methods

results in a completion for a given prompt. CWEs may be extracted from CLMs in a process similar to the one described above for MLMs (Fig. 7).

5.2 Bias detection

In this section, we present an in-depth overview of the technical methods employed in the collected studies. We elect to divide these methods into four categories, which are depicted in Fig. 8.

First, we consider papers that employ methods applicable to either static or contextual word embeddings. These can be subdivided according to which foundational work they take inspiration from. The largest category is descended from the Word-Embedding Association Test (WEAT), first presented by Caliskan et al. (2017). The second sub-category is derived from the methods of the well-known work of Bolukbasi et al. (2016).

Next, we turn to techniques that apply exclusively to large language models such as BERT and GPT. The majority of these techniques apply to masked language models and

are derived from Kurita et al. (2019) and Nangia et al. (2020), with the remaining studies examining causal language models, inspired by Sheng et al. (2019). All of these methods do not work directly on the level of word embeddings, but rather the output of the model in question.

The methods for bias-detection we describe share certain characteristics. Given a language model, i.e., a choice of static word embeddings or large language model, one aims to construct *bias metrics*. Each such metric applies to one or several types of social bias. There are *word-level bias metrics*, which aim to measure the extent to which the model attributes a particular bias to a given word. For example, one might be interested in measuring the gender bias of the word *programmer.* Then there are aggregated *model-level bias metrics*. Roughly speaking, these aim to measure the overall bias of the language model in question. These can be context specific, e.g., measuring gender bias in the context of occupation [e.g., WEAT and DirectBias, or more general, measuring a model's preference for stereotypes or particular demographic groups (e.g., methods derived from Nangia et al. (2020) and Sheng et al. (2019)].

5.2.1 The word embedding association test (WEAT) and related bias metrics

A number of intrinsic bias metrics are based on measuring the level of association in the model between sensitive traits (e.g., gender, ethnicity) and concepts related to pervasive social biases and stereotypes (e.g., occupation). Such techniques typically use lists of words to represent the various categories of interest. For example the concept of *male* could be represented by the set of words $\{$*he, brother, father,...*$\}$. The level of association between two different concepts is measured in terms the average cosine similarity between the word embeddings representing those concepts.

One of the most widely used bias metrics of this type is the Word-Embedding Association Test (WEAT) (Caliskan et al. 2017). WEAT is inspired by the Implicit Association Test (IAT) (Greenwald et al. 1998), which is used is Psychology research to measure implicit bias in human subjects, where human reaction time serves as a proxy for implicit bias. In WEAT, the notion of reaction time is replaced with the cosine similarity between the word vectors. Several common stereotypes measured in humans using IAT were confirmed in to also be present in word embeddings using WEAT.

Each WEAT test requires two categories of wordlists: *attributes* and *targets.* The attributes consist of wordlists A and B representing opposing concepts relating to an aspect of social bias. For example, $A = \{$*executive, management*,...$\}$ and $B = \{$*home, parents,*...$\}$ are attribute lists representing the concepts of *career* and *family* respectively. The targets are also wordlists X and Y. In the context of social biases, these typically represent different values of a particular sensitive trait, e.g., $X = \{$*male, man,*...$\}$ and $Y = \{$*female, woman,*...$\}$ in the case of gender. Using these wordlists, the WEAT test provides a quantitative measure of the degree of bias present in the word embeddings being studied. For example, given the *career vs. family* and *male vs. female terms*, the WEAT test provides a single number measuring the extent to which men are more closely associated to career, whereas women are more closely associated to family. The original WEAT paper defines eight tests related to social bias, which are detailed in Table 2.

We now describe in detail how the WEAT-metric is computed. Let w be a word with corresponding word-embedding $\vec{w}$. The expression

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b}) \quad (1)$$

measures to what extent *w* is more closely associated to *A* or *B*. The sign of *s*(*w*, *A*, *B*) indicates the direction of the bias, while the magnitude indicates the level of bias. For example, if $w = man$ and *A* and *B* correspond to *career* and *family* respectively, and the embedding space indeed encodes stereotypical bias, we would expect *s*(*w*, *A*, *B*) to be a large positive number. The relative association between the target words *X*, *Y* and the attribute words *A*, *B* is then given by

$$s(X, Y, A, B) = \text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B). \quad (2)$$

The overall WEAT bias metric, called the *effect size*, is computed by normalizing *s*(*X*, *Y*, *A*, *B*):

$$es(X, Y, A, B) = \frac{s(X, Y, A, B)}{\text{stddev}_{w \in X \cup Y} s(w, A, B)}. \quad (3)$$

Typically *X*, *Y* and *A*, *B* are chosen so that positive effect sizes reflect stereotypical bias and negative values reflect anti-stereotypical bias, as in the above examples with targets *male vs. female terms* and attributes *career vs. family*. The role of targets and attributes can be switched, and we observed several cases in the literature where wordlists originally designated as attribute sets were used as targets, particularly in the case of *male vs. female terms*. However, switching the role of targets and attributes does affect the normalization factor in the denominator of *es*(*X*, *Y*, *A*, *B*), which should be taken into account when comparing results.

Figure 9 illustrates an example in which the word vectors for the concept of *male* are more closely associated to *career* than *family*, while the opposite is true for the concept of *female.* In this case, the effect size would be positive, indicating stereotypical bias.

Although the effect size may indicate of the presence and magnitude of bias, it cannot be ruled out that the measured effect size is an artifact of the way word vectors are distributed in the embedding space. Caliskan et al. (2017) propose a significance test, the *one-sided permutation test*, in order to ensure that random partitions of the target words $X \cup Y$ do not

Table 2 A list of the WEAT tests that directly pertain to social biases

Test	Bias type	Targets	Attributes
WEAT 3–5*	Ethnic/racial	European American vs. African American first names	Pleasant vs. unpleasant
WEAT 6	Gender	Male vs. female first names	Career vs. family
WEAT 7	Gender	Math vs. arts	Male vs. female terms
WEAT 8	Gender	Science vs. arts	Male vs. female terms
WEAT 9	Health	Physical vs. mental illness	Temporary vs. permanent
WEAT 10	Age	Young vs. old people's names	Pleasant vs. unpleasant

*WEAT 4 uses a subset of names from WEAT 3, and WEAT 5 uses a subset of attributes from WEAT 4

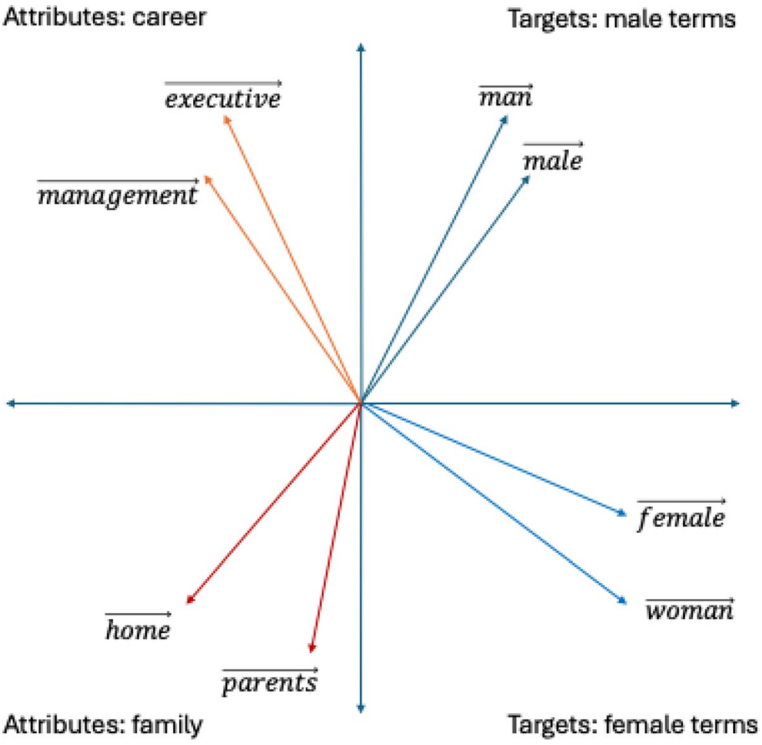

Fig. 9 Example of word embeddings encoding sterotypical bias. The measured effect size would be positive in this case

yield large spurious effect sizes. Let $\{X_i, Y_i\}_i$ denote the set of partitions of $X \cup Y$ into two sets of equal size.[3] The *p*-value for the permutation test is given by

$$p := \Pr_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)], \tag{4}$$

i.e., the fraction of partitions for which $s(X_i, Y_i, A, B) > s(X, Y, A, B)$. A common threshold for statistical significance is $p < 0.05$, meaning that the null hypothesis (that there is no significant bias present) can be rejected at a 5% level of significance.

Sentence embedding association test (SEAT)

WEAT was originally conceived for static word embeddings (SWEs), but one can also carry out the same measurements in the for contextual word embeddings (CWEs). However, unlike the SWE setting, there are several options for how one obtains the CWE vector $\vec{w}$ associated to a target or attribute word w. The simplest method is to simply input a sequence consisting solely of the word w into the CWE-model in question and take, e.g., the output of the final hidden layer(s) as $\vec{w}$. However, such an approach makes no use of the sort of contextual information that makes CWEs so useful. Alternative approaches insert target or attribute words w into sentence templates and use embeddings associated to the entire sentence as $\vec{w}$. May et al. (2019) introduce the Sentence Embedding Association Test

[3] If $|X| = |Y| = n$, then there are exactly $\binom{2n}{n}$ such partitions.

Table 3 A list of the additional WEAT/SEAT tests defined in May et al. (2019)

Test	Bias type	Targets	Attributes
AWBS	Intersectional (race/gender)	European vs. African American female first names	Antonymic traits vs. ABWS traits
DB: competence	Gender	Male vs. female first names	Competent vs. incompetent
DB: likability	Gender	Male vs. female first names	Likable vs. unlikable

(SEAT) as simple method for extending WEAT to the sentence context. Target and attribute words are inserted into semantically bleached templates such as *This is WORD* or *WORD is here.* Apart from replacing word embeddings with sentence embeddings, SEAT scores are computed in the exact same manner as WEAT scores. For each of the static and contextual embeddings considered, May et al. (2019) perform WEAT and SEAT tests corresponding to WEAT 1, 3 and 6.

May et al. (2019) also create additional bias tests. The first is meant to study intersectional bias in the specific case of the *angry black woman stereotype* (ABWS). This is a standard WEAT/SEAT test with targets *European American vs. African American female first names* and attributes *antonymic traits vs. ABWS traits*, where *antonymic traits* refer to the antonyms of the traits associated with the angry black woman stereotype. The second set of tests is geared toward testing the *double binds* faced by many women: If women succeed at a stereotypically male job, then they are often viewed as less likable, whereas, if success is ambiguous, they are viewed as less competent. These double binds are divided into two separate tests. In the first, the targets are *male vs. female first names* and the attributes are *competent vs. incompetent.* The second test uses the same targets, but has attributes *likeable vs unlikeable*. These tests are summarized in Table 3.

The authors perform the double bind test with both semantically bleached and unbleached templates.[4] Both tests use the unbleached attribute template *The engineer is ATTRIBUTE*. For the competence test, the unbleached target template is *TARGET is an engineer*, whereas for the the likability test, the target template is *TARGET is an engineer with superior skills*. Interestingly, for the transformer-based models GPT and BERT, higher magnitude effect sizes are recorded for the unbleached sentence templates compared to the bleached templates.

SEAT is intended to work with both static and contextual word embeddings, but the manner in which the sentence embeddings are obtained depends on the model being used. For example, for GloVe SWEs, sentence embeddings are simply the average of the individual word embeddings corresponding to the words in the sentence. For the transformer-based models GPT and BERT, the authors follow the same procedure for obtaining sentence embeddings as in the original work GPT and BERT. For the causal language model, GPT, this corresponds to take the final hidden state of the last word in the sentence (Radford et al. 2018). For BERT, the final hidden state of the [CLS] token is used (Devlin et al. 2018).

[4] A template is called semantically bleached if all of the non-target/attribute words in the template are semantically neutral. For example, the template *This is WORD* is semantically bleached, while template *The engineer is WORD* is unbleached.

5.2.2 Non-English bias detection using WEAT-inspired methods

Of the studies considered in this survey, fifteen use variants of the WEAT bias metric, summarized in Table 4. Several works carried out straightforward translations of the original English WEAT tests (Toney-Wails and Caliskan 2020; Mulsa and Spanakis 2020; Lauscher et al. 2020). However, the majority of the studies required specific adaptations according to the language(s) in question.

In particular, WEAT 5–6 use first names particular to U.S. English. These tests were adapted in one of two ways: replacement with language/region specific names (Sahlgren and Olsson 2019; Goldfarb-Tarrant et al. 2020; Kurpicz-Briki 2020; Sabbaghi and Caliskan 2022) or, in the case of WEAT 6, using *male vs. female terms* in place of names (Pestova 2021; Lewis and Lupyan 2020; McCurdy and Serbetci 2020).

Another significant hurdle involves adaptation to languages with grammatical gender. Strategies for taking grammatical gender into account involve including all gender forms of translated words in wordlists (Lauscher and Glavaš 2019; Zhou et al. 2019; Zhao et al. 2020), replacing nouns with similar masculine and feminine adjectives (Goldfarb-Tarrant et al. 2020), or comparing the effect of restricting to attributes with a given grammatical gender (McCurdy and Serbetci 2020). Several studies modify the computation of the bias metric to reflect grammatical gender (Zhou et al. 2019; Papakyriakopoulos et al. 2020; Zhao et al. 2020; Lewis and Lupyan 2020), whereas two studies attempt to directly measure *grammatical gender bias*, i.e., to what extent words are more similar exclusively on the basis of grammatical gender (Gonen et al. 2019; Sabbaghi and Caliskan 2022).

Sometimes, pronouns were removed from the *male vs. female terms* lists, since in some languages pronouns can have an ambiguous meaning taken out of context, e.g., *sie* can mean either *she* or *they* in German (Kurpicz-Briki 2020; Pestova 2021; Mulsa and Spanakis 2020).

Translating WEAT, region- and language-specific adaptations

Kurpicz-Briki (2020) adapts WEAT 5–8 for German and French by translating existing wordlists, replacing American first names with corresponding analogs from France, Germany and Switzerland. For German, the authors adapted WEAT 5 with *Swiss German vs. Foreign first names*, using most common names in the German part of Switzerland for men and women respectively. WEAT 6 was adapted for German in two ways, once using the same set of names from the adapted WEAT 5, and once using the most common names of adults living in Germany. For French, the authors followed a similar procedure. In both languages, all pronouns were removed from the *male vs. female terms* attribute sets. Several further language-specific choices regarding translations and omissions in the adaptation of WEAT are detailed in the text.

Wambsganss et al. (2022) expands the German WEAT translations from Kurpicz-Briki (2020) . The authors conduct WEAT tests on the corpus level, in addition to both static and contextual word embeddings. The corpus-level WEAT analysis follows Spliethöver and Wachsmuth (2020) and Caliskan et al. (2017); where the notion of cosine-similarity is replaced by counting co-occurrences of target and attribute words within the same sentence.

Lauscher and Glavaš (2019) adapts WEAT for German, Spanish, Italian, Russian and Turkish. In the case of words that required the specification of a semantic gender in the target language (e.g., *Freund* vs. *Freundin* in German) both options were taken. Tests 5 and 10 were eliminated because they include American names, which were neither adapted nor

Table 4 Table summarizing all WEAT-like metrics

Paper	Languages	Tests	Targets	Attributes	Notes
Kurpicz-Briki (2020)	DE, FR	WEAT 5-DE	German (CH) vs. Foreign first names	–	
		WEAT 6-DE1	German (CH) vs. Foreign first names	–	
		WEAT 6-DE2	German vs. Foreign first names	–	
		WEAT 5-FR	French (CH) vs. Foreign first names	–	
		WEAT 6-FR1	French (CH) vs. Foreign first names	–	
		WEAT 6-FR2	French vs. Foreign first names	–	
		WEAT 7–8	–	–	Pronouns removed
		WEAT KB1-DE	Male vs. female fields of study	Male vs. female terms	
		WEAT KB2-DE	Rational vs. emotional	Male vs. female terms	
Wambsganss et al. (2022)	DE	WEAT 6	–	–	Added corpus WEAT
Lauscher and Glavaš (2019)	DE, ES, IT, RU, TR	WEAT 1, 2, 6–9	–	–	Both grammatical gender translations taken when possible
		XWEAT 6–9	–	–	Targets from language 1, attributes from language 2
Lauscher et al. (2020)	AR	WEAT 1, 2, 7, 8	–	–	
Goldfarb-Tarrant et al. (2020)	ES	XWEAT-mod	–	–	Improved translations from original XWEAT and replaced nouns with both gender forms of corresponding adjectives
		WEAT 5-ES	Migrant vs. non-migrant first names	–	

Table 4 (continued)

Paper	Languages	Tests	Targets	Attributes	Notes
Pestova (2021)	RU	WEAT 6–8	–	–	For WEAT 7–8, targets and attributes swapped. WEAT 6 first names replaced with male vs. female terms. Pronouns dropped
		WEAT P1-RU	Male vs. female terms	Intelligence vs. appearance	
		WEAT P2-RU	Male vs. female terms	Physical vs. emotional strength	
Lewis and Lupyan (2020)	AR, HR, DA, NL, FI, FR, DE, HE, IT, NO, PL, PT, RO, ES, SV, TR	WEAT 6-mod	male vs. female terms	-	Only same gender target and attributes compared
Toney-Wails and Caliskan (2020)	DE, PL, PT, ES, TR	WEAT 8	–	–	
Mulsa and Spanakis (2020)	NL	WEAT 6, 7, 8	–	–	
		SEAT	–	–	
Zhou et al. (2019)	FR, ES	MWEAT	Male form vs female form of occupations	Male vs. female terms	
Sahlgren and Olsson (2019)	SE	SO1	Male vs. female names	Male vs. female occupations	Wordlists based on SE statistics
Papakyriakopoulos et al. (2020)	DE	Pap1	Male vs. female terms	Male vs. female occupations	List of 1600 occupations
		Pap2	German vs. foreigner	Male vs. female occupations	
		Pap3	Straight vs. gay	Male vs. female occupations	
		Pap4–6	" "	Positive vs. negative sentiment	
Zhao et al. (2020)	DE, ES, FR	inBias	Male vs. female terms	Male vs. female form of occupations	

Table 4 (continued)

Paper	Languages	Tests	Targets	Attributes	Notes
Gonen et al. (2019)	DE, IT	GRAM-GO	Similar same gender noun pairs	Similar different gender noun pairs	Cosine-similarity based, but not computed like WEAT. Measures how much grammatical gender accounts for similarity in words
McCurdy and Serbetci (2020)	DE, ES, NL	WEAT 6-mod	Male vs. female terms	–	
		WEAT 6-masc	Male vs. female terms	Career vs. family (masc. gender only)	
		WEAT 6-fem	Male vs. female terms	Career vs. family (fem. gender only)	
		GG-WEAT-MS	Male vs. female terms	Similar masculine vs. feminine inanimate nouns	DE and ES only
Sabbaghi and Caliskan (2022)	DE, FR, IT, ES, PL	WEAT 6, 8	–	–	First names adapted according to language
		GG-WEAT	Similar feminine vs. masculine inanimate nouns	Female vs. male terms	

The use of translations of the original wordlists from (Caliskan et al. 2017) or templates from (May et al. 2019) is indicated by "–". Adaptations during the translation process are described in the Notes column. When available, we refer to the metrics by the name given in the paper

present in sufficient quantity in the vocabulary of the embeddings in other languages. They also define cross-lingual WEAT (XWEAT), which is computed by first using alignment to create a bilingual embedding space and then performing WEAT with target embeddings from one language and attribute embeddings from the other. Using XWEAT, they show that bias levels in the bilingual embeddings lie in between those of the corresponding monolingual embeddings. In follow-up work Lauscher et al. (2020), WEAT is extended to Arabic and compared to an alternative bias measure called the *Embedding Coherence Test (ECT)* Dev and Phillips (2019).[5]

In Goldfarb-Tarrant et al. (2020), the authors modify the Spanish-language XWEAT wordlists from Lauscher and Glavaš (2019) because of two problems stemming from the original XWEAT being too literally translated from English. The first is that many terms do not make sense in the Spanish-speaking community, e.g., the list of names in WEAT 6 or translations such as *arma de fuego* for *firearm*, which is not commonly used in Spanish. The second problem is that the translated nouns on the wordlists for math, science and art concepts are almost all grammatically female, which is shown to distort WEAT measures in McCurdy and Serbetci (2020) and Gonen et al. (2019). The authors balance the wordlists

[5] The ECT uses the gender word-pairs and profession list introduced by Bolukbasi et al. (2016) (see Sect. 5.2.2 below) to measure bias by looking at the coherence between the nearest neighbor professions of male words and those of female words.

by replacing abstract nouns with corresponding adjectives, which can take both a male and female form (e.g, replacing *ciencia* with *científico* and *científica*).

Pestova (2021) extends WEAT to Russian, measuring gender bias by using the *male vs. female terms* as attributes and the *career vs. family*, *math vs. arts* and *science vs. arts* sets as targets. Words that were uncommon in Russian were removed or modified. All pronouns were removed from the *male vs. female terms* attribute set.

Lewis and Lupyan (2020) conduct WEAT 6 in 25 languages, but replaces *male vs. female first names* with *male vs. female terms*, since proper names do not translate well across languages. Wordlists were translated by native speakers, except in the case of Tagalog. The authors also make modifications for grammatically gendered languages: When computing the relationship between target and attribute words in the family category, only targets and attributes of the same gender were compared (e.g., *hombre* (man) to *niños* and *mujer* (woman) to *niñas*). When there were multiple translations for a word or the translation was a multi-word phrase not contained in the word embedding model's vocabulary, the corresponding embeddings were averaged.

Although social biases are not the main topic of interest in Toney-Wails and Caliskan (2020), the authors do perform WEAT 8 in five non-English languages: German, Polish, Portuguese, Spanish and Turkish. The target and attribute lists were translated and verified by native speakers.

In Mulsa and Spanakis (2020), the authors translate WEAT wordlists into Dutch, removing ambiguous pronouns. Additionally, the authors adapt SEAT to Dutch by translating the template sentences originally presented in May et al. (2019).

Sabbaghi and Caliskan (2022) extend WEAT to German, French, Italian, Spanish and Polish. Lists of common names were adapted on a language-by-language basis.

New wordlists for additional tests

Several of the studies we collected introduce new wordlists intended for use in new WEAT-like tests. Some of these wordlists, such as those presented in Kurpicz-Briki (2020), Sahlgren and Olsson (2019) and Goldfarb-Tarrant et al. (2020) are more well-suited than existing WEAT wordlists for evaluating bias in a more region-specific manner, in the sense that they are derived from studies and statistics particular to the region of interest.

Kurpicz-Briki (2020) creates new tests and corresponding wordlists to measure additional gender stereotypes in German, the first with targets *stereotypically male vs. female fields of academic study in Switzerland*, the second with targets *rational vs. emotional*. Both tests use *male vs. female terms* as attributes.

Pestova (2021) creates two new target sets in Russian for testing gender bias: *intelligence vs appearance*, *physical vs. emotional strength*, *STEM vs. humanities* and *rationality vs. emotionality* derived and translated from similar lists in Kurpicz-Briki (2020) and Garg et al. (2018).

Similar to Kurpicz-Briki (2020), Goldfarb-Tarrant et al. (2020) also introduce a modified WEAT 5 test for bias against migrants in Spanish using *migrant vs. non-migrant first names*. Names were selected from Salamanca and Pereira (2013), which allowed the authors to control for socioeconomic class in the choice of names. The *pleasant vs. unpleasant* attributes were balanced for grammatical gender by taking both masculine and feminine forms.

Working in Swedish, Sahlgren and Olsson (2019) follows a methodology resembling WEAT, essentially considering *male vs. female names* as targets and *stereotypically male vs. female occupations* as attributes. The 100 most common Swedish male and female

names were collected from Statistics Sweden[6], along with the 14 most typically male and 14 most typically female occupations. Although such wordlists fit directly into the WEAT framework, it should be noted that the authors instead compute $s(w, A, B)$ for every name w, and then evaluate gender-occupation bias by computing the percentage of, e.g., male names more similar to male occupations, female names more similar to female occupations etc. This is a much less fine-grained measure than WEAT, as it is effectively equivalent to replacing $s(w, A, B)$ with

$$\tilde{s}(w, A, B) = \begin{cases} 1, & \text{if } s(w, A, B) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

and computing, e.g., $\text{mean}_{x \in X} \tilde{s}(x, A, B)$.

Modifying WEAT metrics for gendered languages

The WEAT test was originally developed for English, which as not a gendered language. Complications arise in adapting WEAT for gendered languages. First, attributes often consist of adjectives, which must be modified according to the gender of the noun they describe. Additionally, nouns used to represent concepts such as math, science and the arts also have a grammatical gender. Some of the studies collected for this survey demonstrate that these grammatical gender complications distort WEAT measures (Gonen et al. 2019; McCurdy and Serbetci 2020). As mentioned above, one solution is to modify WEAT on the wordlist-level in order to better account for grammatical gender (Goldfarb-Tarrant et al. 2020; Lewis and Lupyan 2020). In this section, we describe studies that also make modifications to the computation of WEAT for gendered languages.

Working in French and Spanish, Zhou et al. (2019) elect to measure word-level bias differently depending on the type of word. Let A and B be the translated *male vs. female terms* from WEAT 7 and 8. The gender bias of inanimate nouns such as $w =$*agua* (water, feminine), which should not be semantically leaning towards one gender or the other, is measured via $b_w = |s(w, A, B)|$. For animate nouns with two grammatical gender forms, e.g., *mathematicien* and *mathematicienne*, a slight modification is made to measure word-level bias via:

$$b_w = \big||s(w_m, A, B)| - |s(w_f, A, B)|\big|, \tag{6}$$

where w_m and w_f denote the masculine and feminine forms of the word. The authors measure overall bias by aggregating over a set of 58 occupation words in Spanish and 23 in French collected for the study. To aggregate the word-level measures of gender bias in occupations, a *modified word embedding association test* (MWEAT) is introduced:

$$B_{MWEAT} = \Big|\Big|\sum_{x \in X} s(x, A, B)\Big| - \Big|\sum_{y \in Y} s(y, A, B)\Big|\Big|, \tag{7}$$

where X consists of the masculine forms of the occupations and Y consists of the feminine forms.

Papakyriakopoulos et al. (2020) define a bias metric that is very similar to WEAT and apply it for German. As in WEAT, wordlists are used to represent different concepts and bias

[6] www.scb.se.

is measured by calculating the average relative cosine similarities between corresponding word embeddings. The authors study gender, ethnicity and sexual orientation bias and construct lists of word-pairs, e.g. (*Mann*, *Frau*), for each type of bias: *Man-Woman, German-Foreigner, Straight-Gay*. Although the bias metric presented in Papakyriakopoulos et al. (2020) is more general, we present it here for the specific case of gender bias in the occupational context. Let X be a list of pairs of gender terms, e.g. (*Mann*, *Frau*), and let A be a list of gender form pairs of occupations, e.g. (*Arzt*, *Ärztin*). The authors measure the overall bias via:

$$B_{Pap} = \frac{1}{|X||A|} \sum_{x \in X} \sum_{a \in A} |\cos(x_m, a_m) - \cos(x_f, a_f)|, \tag{8}$$

where the subscripts *m* and *f* denote the male and female forms respectively. Although this bears a strong resemblance to Eq. 7, it should be noted that B_{MWEAT} measures bias in terms of how closely each gender is associated to the specified group of occupations overall, whereas B_{Pap} computes the gender bias of each occupation one-by-one and then aggregates these occupation-specific biases. Papakyriakopoulos et al. (2020) create a list of 1600 professions for measuring B_{Pap}. The authors use the same equation to compute gender bias with respect to sentiment, where A is replaced with *positive-negative* sentiment pairs developed by Remus et al. (2010).

Working in Spanish, German and French, Zhao et al. (2020) introduce a bias metric called inBias, which is very similar to the metrics defined in Zhou et al. (2019), Papakyriakopoulos et al. (2020), with the principal difference consisting of replacing *cosine similarity* with the inverse notion of *distance*, where $\text{dis}(w_1, w_2) = 1 - \cos(w_1, w_2)$. The authors use inBias to measure gender bias in the context of occupations. The target sets consist of *male vs female terms*. Similar to Zhou et al. (2019), Papakyriakopoulos et al. (2020), a list of 257 word-pairs consisting of the male and female versions of occupation words, e.g., (*doctor, doctora*) is compiled. Using the same notation as in Eq. 8 and simplifying the notation presented in Zhao et al. (2020), inBias is measured via:

$$B_{inBias} = \frac{1}{|X||A|} \sum_{x \in X} \sum_{a \in A} \left| \text{dis}(x_m, a_m) - \text{dis}(x_f, a_f) \right| \tag{9}$$

Like the metrics in Zhou et al. (2019), Papakyriakopoulos et al. (2020), inBias does not indicate the direction of bias or to what extent encoded bias matches stereotypical bias.

Measuring grammatical gender bias

The notion of *grammatical gender* is distinct from *semantic gender*, however Gonen et al. (2019) demonstrate that grammatical gender affects WEAT tests. The authors assert that the most reasonable explanation for this behavior is that word vectors corresponding to nouns of the same grammatical gender are expected to be closer together than those corresponding to nouns with different gender. This is highly relevant for certain WEAT tests. For example, in French, many of the math-, science- and art-related WEAT words have grammatically feminine translations, which could result in bias measurements that indicate less bias than what is actually encoded in the embedding space. In this section, we describe efforts to precisely analyze the contribution of grammatical gender to bias measurements.

Working in Italian and German, Gonen et al. (2019) study how grammatical gender affects word embeddings using inanimate noun pairs extracted from the SimLex-999 dataset (Hill et al. 2014), which is available in English as well as German and Italian (Leviant and Reichart 2015). German and Italian nouns were manually associated to their grammatical gender. This yields 529 similar noun pairs respectively for each language. The pairs were divided into two sets: (1) pairs of nouns with the same gender, denoted by A, and (2) pairs of nouns with different gender, denoted by B. For example, in German, the pair *Apfel* and *Orange* (both masculine) would belong to A, while the pair *Mond* (masculine) and *Sonne* (feminine) would belong to B. For comparison, the English pairs were also split according to the gender of the nouns in the gender-marked language.

The authors then compute the difference between the average similarity of same gender pairs and the average similarity of different gender pairs:

$$B_{Gram-Go} = \text{mean}_{a \in A} \cos(a_1, a_2) - \text{mean}_{b \in B} \cos(b_1, b_2). \tag{10}$$

Large values for $B_{Gram-Go}$ suggest that grammatical gender plays a significant role in the similarity of the corresponding word embeddings; similar nouns with the same grammatical gender are closer together than similar nouns with different genders. It is shown that values of $B_{Gram-Go}$ for Italian and German were much larger than corresponding values for English.

McCurdy and Serbetci (2020) use variants of WEAT to study the role of grammatical gender in WEAT gender-bias measurements in German, Spanish and Dutch. All tests were performed using the *male vs. female terms* from WEAT 7–8 as target sets. For gender bias, the authors used the *career vs. family* attributes from WEAT 6, comparing the results using the full attribute lists to those obtained from only the male- or female-gendered attribute words. The authors also attempt to measure grammatical gender bias in German and Spanish embeddings directly, using attribute lists derived from similar inanimate object pairs with opposite gender (Phillips and Boroditsky 2013). It was found that social gender bias measurements can interact with and be surpassed in magnitude by grammatical gender associations.

Sabbaghi and Caliskan (2022) also study 4 the interaction between WEAT and grammatical gender, creating GG-WEAT to analyze the contribution of grammatical gender to WEAT measures. In GG-WEAT, the attribute sets are *female vs. male terms*, while the target sets are *inanimate grammatically feminine vs. masculine nouns*, similar to McCurdy and Serbetci (2020). The authors note that care must be taken in constructing the target sets to omit inanimate nouns with stereotypical semantic gender associations, e.g., *la moda* and *el baloncesto* (fashion and basketball). Similar to Gonen et al. (2019), the target sets are constructed by taking pairs of semantically similar nouns with opposite grammatical gender from the Simlex-999 dataset (Leviant and Reichart 2015) (e.g. *molecula* and *atomo* in Italian).

5.2.3 DirectBias: measuring the gender component of word embeddings

The second main line of methods for measuring bias in word embeddings, often referred to as *DirectBias*, is descended from the work of Bolukbasi et al. (2016). Their work famously demonstrates that word embeddings trained on a general corpus complete the analogy *man is to computer programmer as woman is to...* with *homemaker*. The basic idea of DirectBias

is to begin with a list of word pairs that differ primarily in terms of their gender association, e.g., (*man*, *woman*), in order to identify the *gender axis* of the embedding space.

In theory, the gender axis or *gender direction*, is a 1-dimensional vector subspace that parametrizes the gender association of the word embeddings. In other words, the gender bias of a word embedding is proportional to the magnitude of its projection onto the gender axis, and the overall bias of the model is equal to the average gender bias over a prescribed set of words, such as a list of occupations.

To illustrate this in more detail, let $m = \overrightarrow{man} \in \mathbb{R}^d$ and $w = \overrightarrow{woman} \in \mathbb{R}^d$ denote the word-embeddings for *man* and *woman* respectively. In an ideal setting, the only difference between the vectors *m* and *w* is gender, so the the vector $g = m - w$, that is, the information that modifies the concept of woman into the concept of man, perfectly captures the notion of semantic gender. The gender axis is then simply the line $\ell_g \subset \mathbb{R}^d$ spanned by the vector *g*.

In practice, word embeddings do not exhibit such idealized behavior. There are many other differences between the words *man* and *woman* besides pure semantic gender. This procedure is thus repeated several times with different sets of *defining pairs*, e.g., $D =$(*he*, *she*). One thereby obtains a collection of gender vectors $G = \{g_1, \ldots, g_n\}$ associated to a set of *defining pairs* $D_1, D_2 \ldots D_n$, and the *gender axis* *g* is defined as the first principal component of the vectors in *G*. The principal component defines the direction of highest variance in the gender vectors g_i, which can be interpreted as the vector capturing the main difference between the male and female concepts. It is thus expected to capture semantic gender. Figure 10 contains a simplified depiction of the gender direction with two defining pairs. Bolukbasi et al. (2016) use a set of ten defining pairs to determine the gender axis of any given English word-embedding space.[7]

Once the gender axis has been identified, it is possible to compute a numerical bias measure. Given an arbitrary word vector *w*, the (gender) bias of the word *w* is given by:

$$b_w = \cos(w, g). \tag{11}$$

The sign of b_w depends on whether the vector *g* was chosen to point in the male or female direction. In the former case, a positive value for b_w indicates that the word *w* is biased in the male direction. In Fig. 10, the word vector for nurse points toward the negative direction on the gender axis. In other words, $b_{\overrightarrow{nurse}} < 0$, indicating a bias toward the female gender.

In order to construct an aggregate measure of bias in the embeddings, one identifies a set of *domain words* *N* that represent the domain(s) in which one would like to measure bias. Bolukbasi et al. (2016) use a set of 327 occupations.[8] The *DirectBias* of the embedding space is defined as

$$db(N, g) = \sum_{w \in N} |\cos(w, g)|^c = \sum_{w \in N} |b_w|^c, \tag{12}$$

where *c* is a parameter that can be adjusted according to the desired level of strictness in measuring bias. Choosing $c = 0$ is highly strict; *db*(*N*, *g*) would then be equal to the percentage of words in *N* that demonstrate any bias at all. On the other hand, setting $c = 1$

[7] https://github.com/tolga-b/debiaswe/blob/master/data/definitional_pairs.json.

[8] https://github.com/tolga-b/debiaswe/blob/master/data/professions.json.

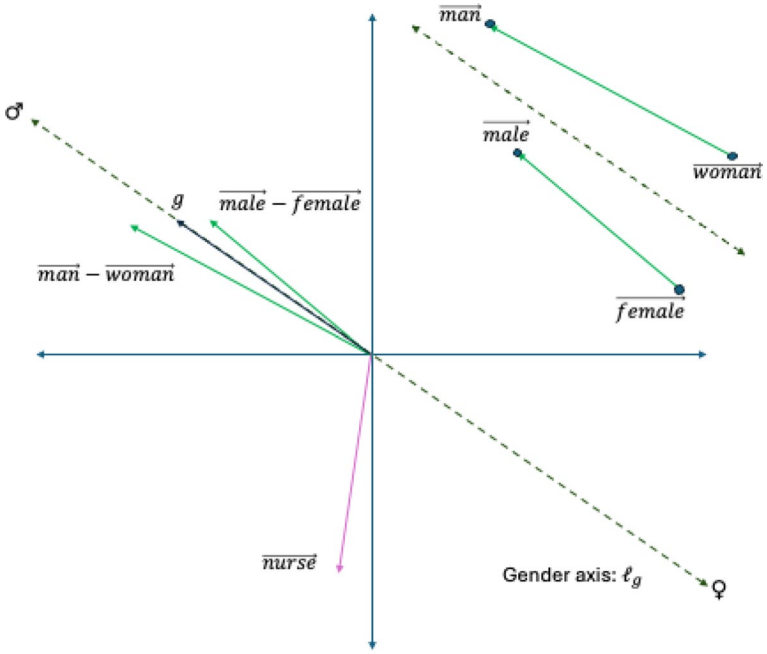

Fig. 10 The gender axis for a set of two defining pairs. The gender axis captures the main difference between the male terms and female terms. The word vector for *nurse* is pointing more toward the negative direction on ℓ_g, indicating a bias toward the female gender

makes $db(N, g)$ a weighted average that accounts for the *degree* of bias measured for each individual word in N. Unlike WEAT, DirectBias as defined above does not a priori provide any indication of whether the measured bias is stereotypical or anti-stereotypical or which gender, if any, the bias tends toward.

Although the methods described above assume that a single dimension captures the concept of semantic gender, it is important to note that Bolukbasi et al. (2016) provide a more general framework. Using further principal components, one can define a higher dimensional *gender subspace* that theoretically captures further semantic gender information. One can also modify the framework and use defining word sets D_i consisting of more than two words. Both generalizations could be the foundation for similar methods that move beyond binary gender or extend to other attributes such as race and sexuality.

5.2.4 Non-English bias detection using DirectBias-inspired methods

For simplicity, we henceforth use DirectBias to refer to the entire methodology of Bolukbasi et al. (2016), not just the metric computed in Eq.12. The meaning will be clear from context. Nine of the papers reviewed for this survey employed bias measurement methods related to DirectBias. These are summarized in Table 5.

Translating defining pairs and occupations

Escouflaire et al. (2023) adapts DirectBias for French, in order to analyze the evolution of gender stereotypes in words representing occupations and social groups in Belgian French articles from 2008-2021.

In joint-work, Chen et al. (2021) and Matthews et al. (2021) adapt DirectBias for eight languages–Chinese, Spanish, Arabic, German, French, Farsi, Urdu and Wolof. Substantial changes are made to both the defining pairs and profession list from Bolukbasi et al. (2016) in order to generalize well across the languages. Six of the ten original defining pairs, such as (*guy, gal*) and (*John, Mary*), are removed because of inconsistent translation across languages and replaced with three better generalizing pairs, e.g. (*king, queen*). The original

Table 5 Studies adapting the methods of Bolukbasi et al. (2016) to non-English languages

Paper	Languages	Defining pairs	Domain words	Notes
Escoufflaire et al. (2023)	FR	-	-	
Chen et al. (2021); Matthews et al. (2021)	ES, AR, DE, FR	Modified	32 of original 327 professions	Some defining pairs do not translate well
Wagner and Zarrieß (2022)	DE	Kinship words	764 role nouns	Use RIPA instead of DirectBias
Katsarou et al. (2022)	SE	(He, she)	-	Extend to CWEs via templates extracted from STS-B dataset
Basta et al. (2021)	ES	-	Both gender forms of occupations	Extend to CWEs via sampling from WMT13 corpus. Use Zhou et al. (2019) methods for grammatical gender
Alshahrani et al. (2022)	AR	-	-	Examine issues with use of original seed words in different context
Gonen et al. (2019)	DE, IT	-	None	Only for mitigation
Zhou et al. (2019)	ES, FR	-	Both gender forms of occupations	Only for mitigation

A hyphen denotes straightforward translation of the define pairs and/or domain words from the original work

327 professions are narrowed down to 32 professions appearing in all eight languages, with attention to variety, salary-bracket and cultural context. In translating to grammatically gendered languages, all gender forms of a given profession were included. Word-level bias is computed on both the defining pairs and profession set. Interestingly, the authors report counter-intuitive findings for the defining pairs, such as the word husband having a female bias in every language. The authors compute DirectBias scores in two ways. The first is the standard average over profession words as in Eq. 12 above with $c = 0$, while the second additionally weighs each profession word by the number of times it occurs in the Wikipedia training corpus. Notably, the authors find that for some languages, the first principal component does not capture a large majority of the variance in the gender-pair vectors.

Constructing new defining pairs and domain words

Wagner and Zarrieß (2022) adapt a method called the *Relational Inner Product (RIPA)* (Ethayarajh et al. 2019) to German. RIPA is almost identical to DirectBias, except that the cosine similarity in Eq. 12 is replaced with the inner product $\langle w, g \rangle$. In practical terms, this means that RIPA accounts for the length of the word embedding vector w, unlike DirectBias, which may contain important information. Wagner and Zarrieß (2022) use a set of nine defining pairs related to kinship, e.g., (*Frau, Mann*), (*Tante, Onkel*), to obtain the gender direction g. The authors then compile a list of 764 role nouns[9], 636 male, 128 female, with 71 occurring in both gender forms. The list of role nouns was extracted from the *Gebrauchsliteratur* section of the German-language fiction corpus,[10] in a process aimed at obtaining all role nouns present in the corpus. To evaluate gender bias in the corresponding word embeddings, the authors study the distribution of their RIPA scores.

Extending DirectBias to contextual word embeddings

In order to adapt Bolukbasi et al. (2016) for contextual word embeddings (CWEs), it is necessary to specify how CWEs for all *seed words* (defining pairs and occupation words) should be obtained. We observed two different strategies in the collected studies. Katsarou et al. (2022) creating a list of template sentences into which the seed words may be inserted as contextual information for CWEs. Alternatively, Basta et al. (2021) collect all sentences containing seed words from a given text corpus. Given a seed word, one obtains a CWE using a context sentence randomly sampled from the collected sentences.

Katsarou et al. (2022) adapt and modify DirectBias for Swedish CWEs. In order to compute the gender direction, a list of 149 sentence templates is extracted from the STS-B dataset[11], which consists of sentence pairs labeled with a scalar measuring their level of similarity. All sentences from STS-B beginning with *A man* or *A woman* are extracted. Of these, only sentences with no further gendered words (*his, hers*, etc.) are retained. The final 149 templates are obtained by simply replacing *A man* or *A woman* with a [MASK] token, e.g., [MASK] *is walking*. In place of using all defining pairs, the authors use only the pair (*he, she*), obtaining 149 pairs of CWEs by replacing the mask token with *he* or *she* is each template. Notably, rather than using principal components, the authors define the gender direction to be the average $g = \frac{1}{149} \sum_{i=1}^{149} (\overrightarrow{she}_i - \overrightarrow{he}_i)$. The word-level bias (Eq. 11) is then computed for nine selected occupations, each of which is either stereotypically male or stereotypically female. Multiple contextual embeddings for each occupation are obtained from the same template sentences by replacing [MASK] with an occupation, e.g., *The nurse*

[9] nouns denoting someone's activity or occupation, e.g., *Zuhörer/Zuhörerin* (listener).

[10] https://www.deutschestextarchiv.de/download.

[11] http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark.

is walking. This yields a distribution of 149 bias values b_w for each occupation w. The presence of stereotypical bias in large transformer models is then studied by comparing these bias distributions over the selected occupations.

Basta et al. (2021) translate the seed words from Bolukbasi et al. (2016) to Spanish and verify the translations by native speakers. In order to compute the gender direction in the setting of CWEs, the authors sample all sentences from the Spanish version of the WMT13 news corpus[12] that contain at least one of the seed words, yielding a list of context sentences S (examples). A CWE for each word in the defining pairs is obtained by randomly sampling a sentence containing that word from S. Following this, the initial gender direction is computed in exactly the same manner as in Bolukbasi et al. (2016). However, the authors modify this gender direction and seed words to account for grammatical gender (see below).

Accounting for grammatical gender

Basta et al. (2021) also attempt to compute the *grammatical gender direction* g_g, for which all nouns from the WMT13 corpus are extracted, approximately 7000 per gender. Contextualized representations are again obtained by random sentence-sampling. Since these nouns do not form natural defining pairs, the grammatical gender direction g_g is computed using linear discriminant analysis (LDA), following Zhou et al. (2019). Ensuring that all vectors are normalized and have consistent gender polarity, the *semantic gender direction* used to measure bias is defined as

$$g = g_{PCA} - (g_{PCA} \cdot g_g)g_g. \tag{13}$$

To compute the overall DirectBias $db(N, g)$, an analogous process is used to obtain contextual representations for each profession. In order to ensure balance, a gender-swapped version of every profession sentence extracted from the WMT13 corpus was added. The authors compute three separate DirectBias measures: once using only the male forms of professions, once using only the female forms, and once using all forms.

Uses beyond bias detection

Some of the papers we collected draw from the methods in Bolukbasi et al. (2016), but do not use DirectBias for bias detection. Alshahrani et al. (2022) critique the use of the original defining pairs and professions in new languages and contexts. Gonen et al. (2019) and Zhou et al. (2019) use WEAT-like measures for bias detection, but also compute the gender direction based on DirectBias methods. This is because Bolukbasi et al. (2016) also provide a method for bias mitigation via subtracting the gender component from word vectors. Since bias mitigation is beyond the scope of this survey, we do not go into full detail here.

Alshahrani et al. (2022) analyze the use of DirectBias for Chinese and Arabic, looking specifically at corpora from several different time periods and studying the limitations of the methods in Bolukbasi et al. (2016) when applied to such corpora. Although they do not compute DirectBias scores, the authors highlight issues with both the gender defining pairs and occupation words used in the original study when applied to this context: The authors identify 50 profession words from the original set that would not exist in corpora from older time periods. In translating to Arabic, the authors take both gender forms of each profession, but note that there are some cases where a male or female version does not exist, or a gender neutral form is used. In addition, several Arabic profession words have other meanings not related to professions. Furthermore, some professions from the original set, such as

[12] https://www.statmt.org/wmt13/translation-task.html.

bartender, are forbidden in the religion of Islam. The authors provide similar findings for Chinese, but note that even words from the gender defining pairs, such as *woman*, can be written in different ways and that usage changes over time.

Gonen et al. (2019) use Bolukbasi et al's methods directly to compute the gender direction in Italian and German. Zhou et al. (2019) make a slight modification for French and Spanish in order to better account for grammatical gender. This is described in Eq. 13 and the text preceeding it above.

5.2.5 Probability-based metrics for masked language models (MLMs)

WEAT, DirectBias and related bias metrics were initially developed for static word embeddings, and, as detailed above, can be adapted for contextual word embeddings (CWEs) in various ways. There are further bias metrics designed specifically for CWEs that use characteristics intrinsic to such embeddings. Recall that contextual word embeddings are context-aware word vectors extracted from the final hidden layer(s) of deep learning based language models. In this section, we describe a set of bias metrics designed specifically for masked language models (MLMs) such as BERT, which use the information encoded in contextual word embeddings to compute probability scores which predict the true value of masked tokens in a given input sequence. The methods we discuss here make direct use of these probability estimates to evaluate the model's bias. This is in contrast working purely at the level of word embeddings, but is very strongly correlated with the information encoded in those embeddings.

Log Probability Bias Score

Many probability-based metrics for contextual word embeddings are derived from Kurita et al. (2019). Inspired by WEAT, the main insight of this work is to replace the use of cosine-similarity as a measure of the level of association between a target (e.g., *man*) and an attribute (e.g. *programmer*). Instead, one uses the *probability* the model estimates for the target and attribute to appear together in the same sentence. This method is only applicable to models trained using a masked language modeling objective, i.e., to predict masked tokens, denoted as [MASK], given the token's context. For example, given an input of the form $x =$ [MASK] *is a programmer*, the model will output a probability estimate $p([\text{MASK}] = w|x)$, the probability that the masked token is given by the word w, for every word w in the model's vocabulary.

To compute the association between the target *male gender* and the attribute *programmer*, one first computes the probability that the sentence [MASK] *is a programmer* will be completed with the word *he*.

$$p_{tgt} = p\big([\text{MASK}] = \text{he} \big| [\text{MASK}] \text{ is a programmer.}\big). \tag{14}$$

Independent of the context, the model may be statistically more or less likely to predict the word *he* than the word *she*, for instance if the corpus the model was trained on contains many more references to male subjects. To account for this difference and isolate the contribution of the word *programmer* to the model's predictions, the probability

$$p_{prior} = p\big([\text{MASK}]_1 = \text{`he'} \big| \text{`}[\text{MASK}]_1 \text{ is a } [\text{MASK}]_2\text{.'}\big) \tag{15}$$

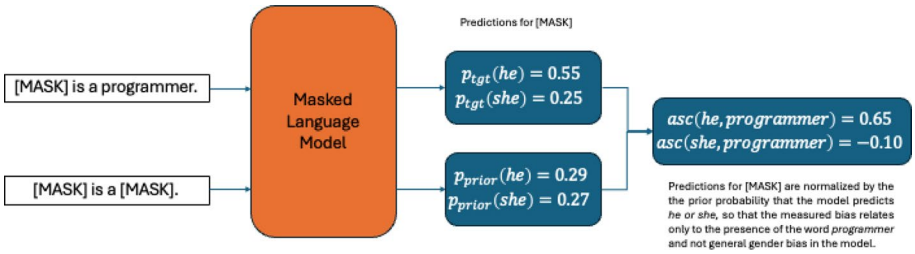

Fig. 11 The masked template is fed into the MLM, both with and without the attribute *programmer*. The predictions for the first masked token are used to compute the increased log probabilities and measure the model's bias. Here the model has slight general gender bias, with a significant bias associated to the word *programmer*

is also computed and used to normalize p_{tgt}.

In general, the association between an arbitrary target x and attribute a is defined as

$$asc(x, a) = \log \frac{p_{tgt}(x|a)}{p_{prior}(x)}, \tag{16}$$

where p_{tgt} and p_{prior} are computed exactly as in the above case with $x = he$ and $a = programmer$. Kurita et al. (2019) refer to $asc(x, a)$ as the *increased log probability score*. For intuition, a positive association means that the probability of the target increased with presence of the attribute, whereas a negative association indicates that the probability of the target decreased when combined with the attribute.

Given an attribute a, the gender bias of a is measured using the *log probability bias score*:

$$b_a^{\log} := asc(\text{he}, a) - asc(\text{she}, a). \tag{17}$$

The steps for computing the increased log probability scores used to measure bias are illustrated in Fig. 11

The increased log probability score $asc(x, a)$ is analogous to the cosine similarity $\cos(\vec{x}, \vec{a})$, and is used to compute an *effect size* completely analogous[13] to WEAT, namely, given an attribute word a and target word lists X and Y, generalizing Eq. 17, one computes:

$$s_{\log}(a, X, Y) = \operatorname*{mean}_{x \in X} asc(x, a) - \operatorname*{mean}_{y \in Y} asc(y, a), \tag{18}$$

One defines $s_{\log}(X, Y, A, B)$ analogously and computes the effect size

$$es_{\log}(X, Y, A, B) = \frac{s_{\log}(X, Y, A, B)}{\operatorname{stddev}_{w \in A \cup B} s_{\log}(w, X, Y)}. \tag{19}$$

For simplicity, we refer to this bias metric as *LPBS*. The authors compute LPBS using the template *TARGET is ATTRIBUTE* and target and attribute lists from WEAT tests 1, 3 and

[13] In the denominator in Eq. 19, the standard deviation is computed over attribute words, as opposed Eq. 3. The main reason for this seems to be that the authors omitted many target words from their tests, so that the resulting distribution does not capture the variance in s_{log} values effectively.

6–8. To test statistical significance, the authors carry out a two-sided permutation test (t-test) over partitions of the combined attributes $A \cup B$, which differs from WEAT. They show that LPBS indicates statistically significant bias in BERT that was not detected by conducting ordinary WEAT on contextual word embeddings extracted from the final hidden layer.

CrowS-Pairs and the Pseudo-log-likelihood Score

The use of templates such as those in Kurita et al. (2019) has been criticized for the limited scope and contrived nature of the resulting sentences. Nangia et al. (2020) address this by compiling the *Crowd-sourced Stereotype Pairs (CrowS-Pairs)* dataset, which consists of 1508 sentence pairs dealing with nine types of social bias: race, gender, sexual orientation, religion, age, nationality, disability, physical appearance and socioeconomic status/occupation. Each pair consists of a stereotypical sentence s_s, such as *Fat people can never really be attractive*, and an antistereotypical counterpart s_a, e.g., *Thin people can never really be attractive.* The sentences s_s and s_a are as semantically similar as possible, only differing in the terms representing the demographic groups being compared. As opposed to template-based methods, it is asserted that the crowd-sourced nature of the dataset results in greater diversity in both sentence structure and the stereotypes expressed.

Similar to Kurita et. al., the authors use sentence pairs to define a probability-score-based bias metric. Given a sentence pair (s_s, s_a), let U be the set of shared words in s_s and s_a, e.g., $U = \{$*people*, *can*, *never*, *really*, *be*, *attractive*$\}$. Rather than using the increased log probability score (Eq. 16) to measure the likelihood of the sentence s_s, the authors adapt the *psuedo-log-likelihood (PLL)* score (Salazar et al. 2019)

$$pll(s_s) := \sum_{u \in U} \log(p([\text{MASK}] \ = u | s_s \setminus u), \tag{20}$$

where $s_s \setminus u$ denotes the sentence s_s with a [MASK] token in place of the word *u*, e.g., *Fat* [MASK] *can never really be attractive*. Using the above example concerning physical appearance, $pll(s_s)$ can be interpreted as the likelihood the model attributes to the remaining part of the sentence given the presence of the word *fat* in the beginning.

The difference

$$b^{p\log}_{s_s,s_a} := pll(s_s) - pll(s_a), \tag{21}$$

is then a bias measure analogous to the log probability bias score (Eq. 17). It measures the degree of the model's preference for the stereotypical sentence over the anti-stereotypical sentence (Fig. 12).

To measure the overall bias of the model, the authors compute the percentage of pairs (s_s, s_a) in the full CrowS-pairs dataset for which the model prefers the the stereotypical sentence s_s over the anti-stereotypical s_a, i.e.,

$$B_{CrowS} := \frac{100}{N} \sum_{(s_s,s_a)} \mathbb{I}(pll(s_s) > pll(s_a)), \tag{22}$$

where $N = 1508$ is the number of pairs in the dataset.

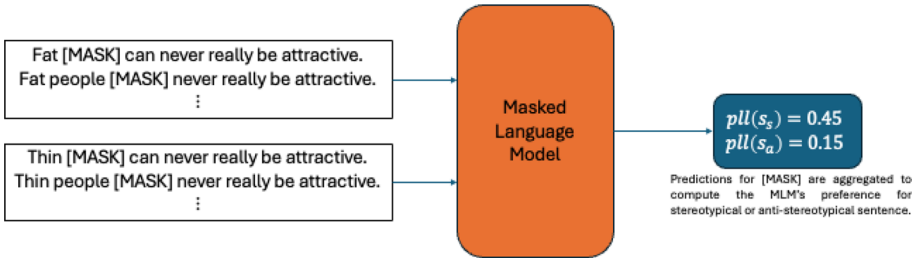

Fig. 12 Likelihood scores are computed for both the stereotypical and anti-stereotypical sentence. Here the model demonstrates preference for the stereotypical sentence

5.2.6 Non-English bias detection using LPBS and CrowS-Pairs inspired methods

Six of the papers collected for this study use metrics based on the probability estimates of masked language models. The studies most similar to Kurita et al. (2019) are summarized in Table 6, while those derived from Nangia et al. (2020) are described in Table 7. Every study we collected modified the original methods of Kurita et al. (2019) in some manner. Zmigrod et al. (2019) use verb-less templates and compare log probability bias scores (Eq. 17) between various gendered animate nouns and a set of four adjectives, while Bartl et al. (2020) and Ahn and Oh (2021) generalize Kurita et al's methods to include more templates and go beyond binary target and attribute categories. Kaneko et al. (2022) generalizes Nangia et al's methods to consider both arbitrary male and female sentences and templates modeled after Kurita et al, instead of only similar sentence pairs. They also replace the pseudo-log-likelihood (PLL) with a similar probability estimate. Fort et al. (2022), Névéol et al. (2022) also build on Nangia et al. (2020) by adapting the original CrowS-pairs dataset to a French cultural context.

Adapting LPBS

Zmigrod et al. (2019) computes gender bias in Spanish, Hebrew, French and Italian using templates of the form *GENDERED ANIMATE NOUN + ADJECTIVE* (e.g., *El ingeniero bueno* and *La ingeniera buena*). Animate nouns[14] in each language were extracted from Wikipedia using WordNet[15] and combined with four adjectives: *good, bad, smart* and *beautiful*. Given an adjective and animate noun in both gender forms, the gender bias is computed as the difference in log-likelihoods between the male form and female form, analogous to the log probability bias score $b_a^{\log}$ from Eq. 17, with the gendered noun as a target and the adjective as the attribute. It should be noted that here these is a slight difference compared to Eq. 17 related to grammatical gender: The adjectives in each template agree with the gender of the corresponding animate noun.

Bartl et al. (2020) apply a modified form of Kurita et al's methods in German. Rather using the single sentence template from Kurita et al. (2019), the authors use a set of five templates constructed with the goal of measuring gender bias in professions, e.g., *TARGET works as a PROFESSION*, where targets are *male vs. female terms*. A list of professions was compiled using real workforce statistics from the U.S. Bureau of Labor Statistics (2020).[16]

[14] Nouns referring to something that is alive and sentient, in this case only those applying to humans.

[15] https://wordnet.princeton.edu.

[16] https://www.bls.gov/opub/mlr/2020/.

Table 6 Table summarizing papers adapting methods from Kurita et al. (2019)

Paper	Languages	Tests	Targets	Attributes	Notes
Zmigrod et al. (2019)	ES, IW, FR, IT	Log-prob-bias	Male vs. female animate nouns	Adjectives (good, bad, smart, beautiful)	Template: GENDERED ANIMATE NOUN + ADJECTIVE
Bartl et al. (2020)	DE	LPBS-Bartl	Male vs. female terms	Male vs. female vs. neutral professions	Uses five templates instead of one. Modified to compare three different attribute categories.
Ahn and Oh (2021)	DE, ES, TR, KO, ZH	Categorical Bias	Ethnicities	Social positions	Modifies LPBS to include more templates and multi-class targets.

The authors selected three groups of 20 professions each: Those with the highest female participation, those with the lowest, and those with a roughly 50-50 gender distribution. In German, these professions were translated into both the masculine and feminine form. To evaluate bias, the authors measure and compare the average associations

$$s(X, A) := \frac{1}{|T|}\frac{1}{|A|}\frac{1}{|X|}\sum_{t\in T}\sum_{x\in X}\sum_{a\in A} asc_t(x, a), \tag{23}$$

where X consists of a list of either male or female terms[17], A is one of the three groups of professions, and T is the set of five template, with $asc_t(x, a)$ being association between x and a (Eq. 16) measured relative to the template t.

[17] Eight gender-denoting person words were taken from Kiritchenko and Mohammad (2018) for both male and female respectively, with the omission of *girl/boy*, which are unlikely to appear in the professional context.

Table 7 Table summarizing papers adapting methods from Nangia et al. (2020)

Paper	Languages	Tests	Sentence pairs	Notes
Kaneko et al. (2022)	JP, RU, DE, AR, ES, PT, IN, ZH	MBE	Male vs. female sentences	Replaces PLL with AULA and adds similarity weight.
	JP, RU	MBE-CrowS	CrowS-Pairs	Same as b_{CrowS} with PLL instead of with AULA.
	JP, RU	MBE-Kurita	GENDERED NOUN is a/an OCCUPATION	Same as MBE-CrowS, with pairs according to Kurita et al. (2019) template.
Fort et al. (2022), Névéol et al. (2022)	FR	CrowS	CrowS-Pairs	Translated and expanded to French cultural context.

In one of the few papers in our survey focused on ethnic bias, Ahn and Oh (2021) generalize Kurita et al's methods for *multi-class targets*, as opposed to the binary classes assumed in the case of gender. The authors work in German, Spanish, Korean, Turkish and Chinese. Based on Kurita et al. (2019), the authors generate ten semantically equivalent templates T, which are designed not to contain any clues regarding ethnicity, e.g., *People from TARGET are ATTRIBUTE*. They use thirty *ethnicities* (represented as country names) as targets X and seventy *social positions* (occupations, legal status) as attributes A. Translations were revised by professional translators. These templates, targets and attributes are used to compute a bias metric that generalizes Kurita et al. (2019). Let $T = \{t_1, t_2, \ldots t_\ell\}$ be a set of template sentences, $X = \{x_1, x_2 \ldots x_m\}$ a set of target ethnicity words and $A = \{a_1, a_2, \ldots a_n\}$ a set of attribute words. The authors define the *Categorical Bias* (CB) score:

$$cb(T, X, A) = \frac{1}{|T|}\frac{1}{|A|}\sum_{t\in T}\sum_{a\in A} Var_{x\in X}\left(asc_t(x, a)\right), \tag{24}$$

where $asc_t(x, a)$ denotes the increased log probability score (Eq. 16) computed using the template sentence *t*. When $\ell = 1$, $m = 2$ and $n = 1$, this reduces to the log probability bias score from Kurita et al. (2019) (Eq. 17). The intuition behind the CB score is the following: Consider a sentence template such as *People from* [MASK] *are ATTRIBUTE*. In an unbiased model, the probabilities of [MASK] being replaced with various ethnicity words should not depend on the attribute. The CB score measures the extent to which the model's predictions deviate from this benchmark.

Adapting CrowS-Pairs

With the goal of defining a bias metric that does not require the demanding work of carefully translating existing wordlists and templates into non-English languages, Kaneko et al. (2022) propose the *multilingual bias evaluation (MBE)* score, which uses only uses English *male vs. female terms* along with existing parallel corpora[18]. The MBE score is used to measure bias in MLMs in Japanese, Russian, German, Arabic, Spanish, Portuguese, Indonesian and Chinese. The authors extract all sentences containing either a male or female term (e.g., *he*, *she*) from the English version of the parallel corpora, thereby obtaining sets of sentences E_m and E_f containing male and female terms respectively. These correspond to sets T_m and T_f of sentences from the equivalent corpus in the target language. Rather than attempting to measure gender bias relative to a specific context, such as occupation, the authors use the relative likelihood of male sentences vs. female sentences as a proxy for gender bias. In other words, the model is biased if it systematically assigns a higher probability to, e.g., male sentences over female sentences. As with the pseudo-log-likelihood (Eq. 20) used in Nangia et al. (2020), this requires assigning a likelihood that the model associates to a given input sentence.

Kaneko et al. (2022) use *All Unmasked Likelihood with Attention weights (AULA)* from earlier work (Kaneko and Bollegala 2022). Given an input sentence $s = w_1, w_2, \ldots w_{|s|}$ made up of tokens w_i, the MLM outputs a probabilty $p(w_i|s)$ at position *i* for each token w_i. AULA is a weighted average of the probabilities $p(w_i|s)$, where the terms are weighted by the *importance* of the corresponding token w_i within the sentence *s*. The importance of w_i is computed using all of the attention weights, special parameters in transformer-based models, associated to w_i. Each such weight a_{ij} represents the strength of the relation between w_i and another token w_j, and the importance of w_i is then defined as $\alpha_i := \frac{1}{|s|}\sum_{j=1}^{|s|} a_{ij}$. The AULA score of *s* is then given by:

$$a(s) := \frac{1}{|s|}\sum_{i=1}^{|s|} \alpha_i \log(p(w_i|s)). \tag{25}$$

Additionally, the last layer in the MLM before computing probabilities consists of a sequence of embedding vectors $v_{w_1}, v_{w_2}, \ldots v_{w_{|s|}}$ corresponding to the tokens in *s*. These can be averaged into a vector $\bar{s}$ which represents the sentence *s*, allowing the similarity between two sentences to be measured via

[18] TED2020 v1 and GlobalVoices.

$$c(s_1, s_2) := \cos(\vec{s_1}, \vec{s_2}). \tag{26}$$

Finally, the MBE bias score is defined to be:

$$mbe(T_m, T_f) := 100 \times \frac{\sum_{t_m \in T_m} \sum_{t_f \in T_f} c(t_m, t_f) \mathbb{I}(a(t_m) > a(t_f))}{\sum_{t_m \in T_m} \sum_{t_f \in T_f} c(t_m, t_f)}. \tag{27}$$

This is similar to b_{CrowS} (Eq. 22). The main difference is that the formula has been adapted for situations where sentences do not come in natural semantically similar pairs. Indeed the MBE score can be interpreted as the percentage of male sentences preferred by the MLM over female sentences, except that the proportion is weighted by the similarity of the corresponding sentences. This makes sense, as comparisons between similar male and female sentences are more meaningful than comparisons between very dissimilar sentences. MBE values above 50 indicate a bias in favor of males.

In the case of Japanese and Russian, Kaneko et al. (2022) compare MBE to other methods. All 262 gender-related sentence pairs from the CrowS-Pairs dataset are translated from English, both manually and via machine translation. This is used to compute a bias score using Eq. 22 with the AULA likelihood in place of the pseudo-log-likelihood. They also compute $mbe(S_m, S_f)$, where S_m and S_f are the manually translated male and female sentences from CrowS-Pairs respectively. In this case, since the sentence pairs are similar by construction, the MBE score is computed with $c(t_m, t_f) = 1$ for all pairs (t_m, t_f). Finally, the authors consider sentence templates following Kurita et al and extract occupation lists from Wikipedia in order to generate sentence pairs of the form, e.g., *GENDER is a/an OCCUPATION,* measuring bias using the same formulae as in the case of the CrowS-Pairs sentences.

In a series of two papers (Névéol et al. 2022; Fort et al. 2022), the authors manually translate, adapt and extend the CrowS-Pairs dataset to French. The translated dataset is extended with sentence pairs expressing stereotypes particular to French sociocultural norms using a citizen science platform called LanguageARC[19]. To measure bias, the authors use the same methods as Nangia et al. (2020).

5.2.7 Measuring bias in sentence completions with extrinsic classifiers

Methods derived from Kurita et al. (2019) and Nangia et al. (2020) are specifically tailored to masked language models such as BERT. However, they are not immediately adaptable to transformer-based causal language models such as those in the GPT-series. Sheng et al. (2019) address this by evaluating sentence completions in place of masked-token predictions. As a starting point, they show that GPT-2 completes the prompt *The white man worked as...*, with *a police officer, a judge, a prosecutor, and the president of the United States.* The authors introduce the concept of *regard* towards different demographic groups, i.e., how those groups are viewed on a societal level.

The authors construct prompt templates designed to measure bias in two contexts, *respect* and *occupation*, with five templates for each context. An example from the respect category is *PERSON was known for...*, and an example from the occupation category is *PERSON worked as...*" Here PERSON is a stand-in for a term indicating a particular demographic.

[19] https://languagearc.com.

The authors consider the categories of gender, race and sexual orientation, resulting a 60 unique prompts (ten per demographic group). These are in turn used to generate a total of 360 completions, which annotators labeled as *positive, negative* or *neutral* in terms of the regard those completions indicate for the demographic group in question. The annotated dataset is then used to train an automatic BERT-based regard classifier.

Both the regard classifier and VADER (Hutto and Gilbert 2014), a rule-based sentiment classifier, are used to analyze bias in sentence completions. For each demographic, 500 completions (one for each of the five prompts) are generated in the respect and occupation contexts respectively. Regard and sentiment scores consist simply of the proportion of completions classified as either positive, neutral, or negative. For example, the negative regard score in the respect context for women would be given by:

$$R^{-}_{resp} = \frac{1}{N|P|} \sum_{p \in P} \sum_{c \in C(p,N)} \mathbb{I}(c \in \text{ negative regard}), \tag{28}$$

where P is the set of five prompts for women in the respect category, $N = 100$ the number of generated completions, and $C(p, N)$ is the set of N completions for the given prompt p. See Fig. 13 for an illustration of this process.

5.2.8 Non-English bias detection based on sentence completions

Two of the papers we reviewed measure bias using methods inspired by Sheng et al. (2019), although both restrict their analyses to the category of gender.

Working in Italian, French, Portuguese, Romanian and Spanish, Nozza et al. (2021) create 420 prompts for each language, generated using 28 identity terms (14 male and 14 female) inserted into 15 templates based on those in Sheng et al. (2019). The authors then use HurtLex (Bassignana et al. 2018), a multilingual lexicon of hurtful words to evaluate how often the model's completions contains a hurtful word. For each template, the top $K = 20$ completions are reviewed for hurtful words. The HONEST score is introduced as a metric for how likely the model is to produce hurtful completions, defined via

$$H = \frac{1}{K|P|} \sum_{p \in P} \sum_{c \in C(p,K)} \mathbb{I}(c \in \text{ hurtful}), \tag{29}$$

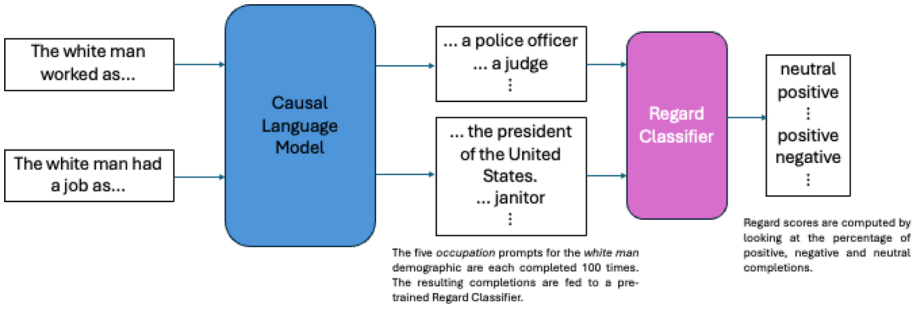

Fig. 13 Computing regard scores in the occupation context for white men

where $c \in \text{hurtful}$ means c contains a word from HurtLex. In other words, the HONEST score is the proportion of top-K completions containing a hurtful word.

Kraft et al. (2022) train a *regard* classifier for German. The classifier is trained on an annotated dataset of 1,157 sentences compiled via survey. Participants were given gender neutral templates derived from Sheng et al. (2019), e.g., *The person is known for...*, and instructed to generate their own completions communicating positive, negative or neutral regard. The authors use translations of the templates from Sheng et al. (2019) to generate 10 prompts each for the *male* and *female* category. The authors evaluate bias by computing the proportion of N completions ($N = 100$ or 50 per prompt) classified as conveying positive, negative or neutral regard respectively.

5.2.9 Further bias measurement methods

The methods presented in the previous sections cover nearly the entirety of the bias metrics appearing in the works collected for this survey. Without going into detail, we also observed the following methods for evaluating bias in contextual word embeddings: Male- and female clustering of CWEs, training classifiers on CWEs of biased words, measuring gender stereotypes in K-nearest neighbors of CWEs corresponding to professions. See Basta et al. (2021) for details.

5.2.10 Comparison of bias metrics

Within the category of WEAT-like metrics discussed in Sect. 5.2.1, many studies compare their adapted metrics to the WEAT baseline (Lauscher and Glavaš 2019; Goldfarb-Tarrant et al. 2020; Mulsa and Spanakis 2020; McCurdy and Serbetci 2020), predominantly in order to demonstrate improvements for gendered languages. Unfortunately, none of the papers collected in this survey compare bias metrics across all four of the main categories identified in Fig. 8. This is an important limitation, as it was demonstrated in Kurita et al. (2019) that some metrics may capture bias that was not detected by others. The question of which bias metrics to use in which situations is critical, and requires careful comparison. Basta et al. (2021) compare DirectBias to a slew of metrics that were not discussed in the scope of this survey, but deliberately exclude WEAT. Zhou et al. (2019) use the WEAT metric to evaluate debiasing techniques derived from Bolukbasi et al. (2016), showing that WEAT scores on DirectBias are at least somewhat correlated.

5.3 Models and corpora

In this section, we explore the models and corpora covered within the surveyed literature. The results of this exploration are detailed in Table 8.

Models

The majority of the studies did not cover contextual word embeddings, which is likely because the most widely adopted bias detection methods, WEAT and DirectBias, were both originally developed for static word embeddings. This indicates a large opening for further research in non-English CWEs. Since transformer-based models have become the default in NLP applications, it is important that this shortcoming be addressed in future research. Additionally, the studies that do include transformer-based models use relatively small

Table 8 Summary of embedding models and corpora studied in the surveyed literature

Paper	Languages	Domain	Embedding model	Corpora	Bias type
Basta et al. (2021)	ES	News	Word2Vec, ELMo	WMT13	Gender
Lauscher et al. (2020)	AR	News	SKIP-GRAM, CBOW, FASTTEXT	Arabic news corpora	Gender
Mulsa and Spanakis (2020)	NL	Various	SWEs and CWEs	N/A	Gender
Papakyriakopoulos et al. (2020)	DE	Social media, Wikipedia	GloVe	Social media, Wikipedia	Gender, nationality, sexual orientation
Ahn and Oh (2021)	DE, ES, KO, TR, ZH	N/A	BERT	Europarl V7 Corpus, UN parallel corpus, Naver Movie Sentiment Corpus, Leipzig Corpora Collection	Ethnic
Bartl et al. (2020)	DE	Various	BERT	Equity Evaluation Corpus (EEC), GAP corpus	Gender
Gonen et al. (2019)	DE, IT	TED talks, news	BERT, multilingual BERT, RoBERTa, ALBERT, DistilBERT, ConvBERT, XLM, Deberta, Word2Vec	TED talks, news corpora	Gender
Kraft et al. (2022)	DE	Various	SentenceBERT	N/A	Gender
Alshahrani et al. (2022)	AR	Religious, scientific, poetic texts	GloVe	Shamela Library, Arabic Poem Comprehensive Dataset	Gender
Escouflaire et al. (2023)	FR	Journalistic articles	Word2Vec	RTBF Corpus	Gender, evaluative
Katsarou et al. (2022)	Various	N/A	T5, mT5	N/A	Gender
Wagner and Zarrieß (2022)	DE	Fiction	Word2Vec	N/A	Gender
Chen et al. (2021)	Various (nine languages)	Wikipedia	Word2Vec	Wikipedia	Gender
Kurpicz-Briki (2020)	DE, FR	Various	GloVe, Word2Vec	N/A	Gender, nationality
Lauscher and Glavaš (2019)	DE, ES, IT, RU, TR	Various	CBOW, GloVe, FastText, Dict2Vec	Common Crawl, Wikipedia, tweet corpus	Gender
Lauscher et al. (2020)	AR	Various	CBOW, SKIP-GRAM, FastText	N/A	Gender
McCurdy and Serbetci (2020)	Multiple languages	Various	CBOW Word2Vec	OpenSubtitles	Gender

Table 8 (continued)

Paper	Languages	Domain	Embedding model	Corpora	Bias type
Mulsa and Spanakis (2020)	NL	Various	6 SWEs and 2 CWEs	OpenSubtitles	Gender
Pestova (2021)	RU	Various	Various	Various	Gender
Sabbaghi and Caliskan (2022)	Multiple languages	Various	FastText	Universal Dependencies datasets	Gender
Wambsganss et al. (2022)	DE	Educational language models	T5, GPT-2, GloVe, BERT	German Educational Peer-Review Data	Conceptual, Racial, and Gender
Zhao et al. (2020)	Multiple languages	Various	M-BERT	N/A	Gender
Zhou et al. (2019)	ES	Various	FastText	N/A	Gender
Goldfarb-Tarrant et al. (2020)	Various	Various	FastText, Skip-gram Word2Vec	Tweets, Wikipedia	Gender, migrant
Lewis and Lupyan (2020)	25 languages	Various	FastText	Wikipedia subtitles	Cultural stereotypes
Sahlgren and Olsson (2019)	SV	Various	Various	N/A	Gender
Wevers (2019)	NL	Newspapers	Word2Vec	Dutch national newspapers	Gender
Névéol et al. (2022)	FR	Various	CamemBERT, FlauBERT, FrALBERT, mBERT	Various	Cultural Stereotypes
Zmigrod et al. (2019)	ES, HE	Various	BPE-RNNLM (Mielke and Eisner 2019)	Various	Gender

models such as the *base* size versions of BERT, GPT-2 and T5, which all possess on the order of 100 million parameters. For comparison, state-of-the-art language models such as GPT-4o have around 200 billion parameters. Such immense models must be studied closely, as they are widely used and there is evidence that larger language models may encode *more* bias than smaller models Gallegos et al. (2024). However, the computational requirements and proprietary nature of state-of-the-art language models make it very difficult to perform bias detection studies similar to those surveyed in this work, likely contributing to increased development of prompt-engineering and sentence completion based methods in the same vein as the those described in Sect. 5.2.4.

Corpora

In the majority of cases, the 'Corpora' column in Table 8 refers to corpora used to train or fine-tune word embeddings. Corpora fall into roughly two categories: (1) *pre-training data* consisting of relatively large and general text datasets, e.g., from sources such as Wikipedia and the CommonCrawl corpus, and (2) *specialized domain corpora*, containing relatively less text overall and pertaining to a particular domain (e.g., news articles, fiction or social media) and language(s).

Pre-trained models and embeddings using large, general corpora were typically used out of the box rather than being trained by the authors of the collected studies. In some cases, the research context motivated the use of specially curated general corpora. For example,

McCurdy and Serbetci (2020) compare embeddings across multiple languages, controlling as much as possible for differences in training data by using the OpenSubtitles dataset to compile a *parallel corpus*[20].

In cases where specialized domain corpora were used, bias detection was applied not only to study bias in the corresponding word embeddings, but also within the corpora in question. For example, Alshahrani et al. (2022) use corpora consisting of Islamic texts and Arabic poetry spanning across different eras, studying how the occupations mentioned in the texts change over time in relation to the occupation lists in Bolukbasi et al. (2016). In another study, Wambsganss et al. (2022) train word embeddings on a German-language corpus of university student peer-reviews, attempting to detect biases within the corpus, corresponding static word embeddings and contextual word embeddings resulting from fine-tuning pre-trained models on the peer-review corpus.

In some cases, corpora are listed in Table 8 that were not used to train word embeddings or language models, but rather for evaluation in downstream tasks. These are included because they consist primarily of textual data and relate to the domain context of the work in question. For instance, Bartl et al. (2020) use the Equity Evaluation Corpus (EEC), which consists of 8640 sentences specially chosen to help identify racial and gender bias.

5.3.1 Limitations and challenges in bias detection in word embeddings

The papers reviewed in this study highlight many challenges in assessing and reducing bias in word embeddings. In this section, we gather some of the obstacles and calls for further research put forth in the surveyed literature. These challenges span the selection and adaptation of training data, templates and word lists, detection and measurement methods, model choice, and language representation.

Data limitations

Some of the collected papers report limitations related to data. For example, Mulsa and Spanakis (2020) found that adapting test data from English to Dutch might miss unique Dutch linguistic features, leading to incomplete bias detection. Pestova (2021) and Kurpicz-Briki (2020) both call for further research on the relation between train corpora and bias in corresponding word embeddings. Sabbaghi and Caliskan (2022) noted that bias testing often assumes two gender classes, overlooking neuter classes in languages like Polish and German, and suggested methods for better disentanglement of grammatical and semantic gender as well as consideration of non-binary representations. Moreover, the study acknowledged a lack of research on the influence of grammatical gender on social biases and human cognition. Wagner and Zarrieß (2022) suggested future research with better-suited corpora, particularly noting the lack of sufficiently large job posting corpora.

Detection and measurement methods

Several studies examine detection and measurement methods for bias. Kraft et al. (2022) face challenges in addressing non-binary gender bias in German and suggested using more natural prompts and considering participant comfort in future studies. Zhou et al. (2019) call for testing both monolingual and bilingual embeddings in downstream tasks and expanding methods to address grammatical gender in languages with a more complex range of gender forms. Chen et al. (2021) suggest exploring the detection of grammatical gender direction in gendered languages and acknowledge the need for corresponding datasets analogous to

[20] Meaning that every text sample is translated into each of the selected languages.

the gender seed words first defined by Bolukbasi et al. (2016). Lewis and Lupyan (2020) highlight the limitations of the Implicit Association Test (IAT) and call for measures more closely related to real-world behavior and further research on the impact of linguistic associations on cultural stereotypes and the effect of bilingualism on implicit bias. Additionally, Kurpicz-Briki (2020) note that WEAT word lists exclude many important contexts in which bias can occur. Goldfarb-Tarrant et al. (2020) attempted to connect intrinsic bias in language models to bias in downstream tasks. However, their tasks were limited to the discriminative context. They called for more studies exploring the effect of intrinsic bias on text generation to address representational harms and suggested future work explore bias in both discriminative and generative contexts. Mulsa and Spanakis (2020) point out that bias testing methods like WEAT and SEAT focus on detecting bias but cannot guarantee the absence of bias, which suggests the necessity to explore both more holistic and more context-specific methods.

Model choice

Model choice is also considered an important domain for further research. Pestova (2021) calls for research examining how hyperparameters and model types relate to bias in Russian word embeddings. Bartl et al. (2020) focus only on one BERT model and suggest that future research explore gender bias across other contextual word embedding models. Moreover, the authors acknowledge limitations in the use of sentence templates and curated word lists, which are also affected by human bias. They recommend that future research include a broader variety of sentences, potentially incorporating random sampling. Katsarou et al. (2022) note sensitivity differences in bias detection between languages and model sizes, pointing out indicators in their own work that encoded gender bias increases with model size, which is alarming given the steady adoption of larger and larger models in real-world applications. Lauscher and Glavaš (2019) find that different models can either amplify or alleviate biases in text, depending on embedding type and training corpora. Sahlgren and Olsson (2019) suggest that future research replicate studies with various embeddings trained on the same data to understand learning algorithms' impact on bias.

Language representation

Language representation challenges are discussed in three studies. Ahn and Oh (2021) suggest that future work should include more languages and more detailed ethnic groups, rather than using nationality as a stand-in for ethnicity. Alshahrani et al. (2022) suggest expanding to more languages in addition to incorporating English for valuable comparison, given that the vast majority of modern NLP methods were developed and optimized for English. Zhao et al. (2020) acknowledge their focus on European languages and encourage extending research to languages with different grammatical gender structures, such as Czech and Slovak.

6 Discussion and future work

In this section, we relate the studies collected within this survey to broader work in bias detection and mitigation in word embeddings and language models, additionally encompassing state-of-the-art research for the English language and surveys of such work (e.g., Sun et al. 2019; Meade et al. 2022; Delobelle et al. 2022). Partially guided by comparison to existing work in English, we indicate several important directions for future research.

6.1 Diversity in terms of languages and types of bias

The majority of studies reviewed in this literature focus primarily on gender bias. While important, this emphasis reveals a gap compared to the state-of-the-art research in English, which is increasingly addressing other types of bias, particularly *intersectional bias* (e.g., Guo and Caliskan 2021). Although various languages were identified (see Fig. 4), not all languages from our search query are covered in the collected works. Therefore, there is considerable scope for improvement in detecting bias and adapting methods and word lists to meet the specific needs of some European languages, which are currently underrepresented in research. A recent survey on the effectiveness of debiasing techniques for pre-trained language models (Meade et al. 2022) highlights that the focus on English-trained models is a limitation of their work. Additionally, their survey points out oversimplified assumptions about bias, such as a binary definition of gender. While there is initial research considering non-binary gender definitions for use in debiasing English word embeddings (e.g., Manzini et al. 2019; Meade et al. 2022), much remains to be explored. Delobelle et al. (2022) also observed that most methods should be extendable to non-binary settings and other biases, but this is often not considered by the authors.

6.2 Language and cultural-specific aspects

We observe that translation is often used as a first step to adapt sentence templates and word lists from English to other languages. There is large potential for more human-centered approaches that consider specific regional or cultural aspects, as only a small portion of the surveyed literature adopt more culturally situated practices. When using translation, a cultural shift to US English can occur, leading to a loss of cultural context for bias. Some studies addressed this by adapting word lists to local circumstances, such as using different types of first names (e.g., local and migrant group names instead of European-American and African-American names). Meade et al. (2022) also acknowledge that their work is skewed towards North American social biases, which is a limitation.

Most of the surveyed research on gender bias focuses on the occupational sector. While some word lists and sentence templates are adapted for specific linguistic or cultural properties, they are not tailored to other contexts. Future work could explore how bias manifests in different use cases. This poses a challenge for current bias detection methods, which typically address one use case at a time in a very limited context, making it difficult to achieve a holistic measure of bias in the models.

Overall, we find that the term *bias* is often used broadly in the literature. Hellström et al. (2020) point out that the term bias in machine learning is used in many different contexts with varying meanings. This is underscored in our discussion on the dimensions of bias in the examined papers, where we observed the use of similar but not equivalent terms to describe various sensitive characteristics (e.g., origin, nationality, cultural differences). In their survey on bias in NLP, Blodgett et al. (2020) also note that *bias* is also used to describe a wide range of system behaviors that could be harmful to different groups in various ways. We suggest that future research precisely define these notions, following Blodgett et al. (2020)'s recommendation to provide explicit statements on why the described system behaviors are harmful, in what way, and to whom.

6.3 Limitations of bias metrics

All of the bias metrics we discuss in this survey are sensitive to the choice of word lists or test sentences. They are all only indicators of bias, limited to the context captured in the word lists, and cannot be considered exhaustive. Limitations are discussed in both the studies included in this survey and the broader literature. WEAT sensitivity to both word lists and corpus word frequency is discussed in Sedoc and Ungar (2019) and Ethayarajh et al. (2019). Corpus word frequency is cited as a crucial limitation in Wambsganss et al. (2022). Ethayarajh et al. (2019) also demonstrate theoretically that both WEAT and DirectBias are prone to overestimate bias in word embeddings, introducing the more theoretically robust RIPA metric, which was used in Wagner and Zarrieß (2022) and mentioned as being of interest for future work in Goldfarb-Tarrant et al. (2020). Kurita et al. (2019) also argue that WEAT is insufficient for measuring bias in CWEs. For these reasons, Basta et al. (2021) opt out of WEAT entirely when evaluating bias in contextual word embeddings. As discussed above, existing gender bias metrics also suffer from entanglement with grammatical gender in the case of gendered languages (Gonen et al. 2019; McCurdy and Serbetci 2020; Zhou et al. 2019; Sabbaghi and Caliskan 2022; Basta et al. 2021; Goldfarb-Tarrant et al. 2020; Lewis and Lupyan 2020; Papakyriakopoulos et al. 2020; Zhao et al. 2020). None of the methods we encountered attempt to study intersectional bias or non-binary notions of gender.

In this survey, we discussed only *intrinsic* bias metrics, i.e., metrics that attempt to measure bias inherently present in word embeddings and language models. This is opposed to *extrinsic* bias metrics, which attempt to measure bias in downstream tasks[21], such as ranking job candidates or translation. Goldfarb-Tarrant et al. (2020) demonstrate that the correlation between intrinsic bias metrics and and bias in downstream tasks is weak, concluding that the relevance of intrinsic metrics depends strongly on the downstream task in question. They recommend that intrinsic metrics be used as descriptive metrics in computational social science and examining bias in human text, but NOT for measuring model bias. Instead, the NLP community should focus on the creation of datasets and challenge sets that allow bias measurement in downstream applications. Delobelle et al. (2022) corroborate Goldfarb-Tarrant et al. (2020)'s findings, although they do report much better correlation of the LPBS and CrowS-Pairs metrics in comparison to WEAT and SEAT with extrinsic metrics on downstream tasks in the occupational context.

Another limitation of the techniques summarized in this survey is that they are not applicable to widely used proprietary models such as GPT-4o, the model underlying ChatGPT at the time of writing; where researchers may not be given access to embeddings or probability scores. Of the studies collected for this survey, only the two discussed in Sect. 5.2.4 attempt to directly detect and measure bias in AI-generated text. Although outside the scope of this survey, it is important to note that bias detection in human- and AI-generated text plays an essential role in curtailing harmful outcomes from AI-aided decision-making; the bias encoded in models originates from biases in the training data, which are likely to be repeated and reinforced AI-generated text. Thus, a holistic approach regarding bias in training data, models and generated text is recommended.

[21] We do not consider sentence completion and masked token prediction to be downstream tasks, since the language models we consider were originally trained precisely on these taks.

6.4 Outlook for future work

Based on our findings, we can encourage future work to consider more diverse types of bias (i.e., going beyond gender), go beyond binary representation, and in particular address the concept of intersectionality. Our survey revealed numerous limitations in terms of the scope and understanding of bias addressed by existing detection methods. For example, WEAT only encompasses 5 dimensions of bias, whereas CrowS-Pairs encompasses 9. Working in English, Smith et al. (2022) construct a more holistic framework for constructing test sentences encompassing 13 dimensions of bias. Future work should explore more both (1) more holistic bias detection methods for non-English languages and (2) intrinsic bias metrics that are more context- and culture- specific, particularly with an eye towards intended downstream applications. Additionally, more work can be done with regard to bias detection for languages (1) included in the search for this survey, but not found, (2) additional languages that have not been covered by this survey.

Interestingly, some of the points identified in this survey, and partially mentioned by previous surveys from the recent past were already identified by Sun et al. (2019) in 2019: the authors encourage researchers to mitigate gender bias beyond English, consider non-binary gender bias, and in particular address interdisciplinary collaboration. In our survey, we have seen room for improvement with regard to these aspects, and encourage researchers to consider them as directions for future work.

Finally, the proposed methods in the state-of-the-art identified in this literature survey rely on the fact that the embeddings are available. Under consideration of the shift in the recent months towards models such as GPT-4o[22], where contextual embeddings or predicted probabilities may not be publicly available, it has become necessary to open up separate avenues of research for bias detection and mitigation in closed-source language models. Early forays into this field use prompt-based approaches such as in Bai et al. (2024), including tools providing modules to apply prompt-based bias testing to a variety of large-language models.[23] These methods are similar in spirit to those described in Sect. 5.2.4.

Acknowledgements This work is part of the Europe Horizon project BIAS, grant agreement number 101070468, funded by the European Commission, and has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

Author contributions AP: conceptualization, methodology, investigation (data collection and selection, execution of review) and writing (original draft). CI: conceptualization, methodology, investigation (data collection and selection, execution of review) and writing (original draft). CR: investigation (execution of review) and writing (original draft). EFV: investigation (execution of review) and writing (review & editing). MK: investigation (execution of review) and writing (review & editing). RS: investigation (execution of review) and writing (review & editing). MK-B: supervision, investigation (data collection and selection, execution of review) and writing (review & editing).

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

[22] https://openai.com/index/hello-gpt-4o/.

[23] https://github.com/SOM-Research/LangBiTe, https://ai-sandbox.list.lu/.

References

Ahn J, Oh A (2021) Mitigating language-dependent ethnic bias in bert. arXiv preprint arXiv:2109.05704

Aloisi A (2023) Regulating algorithmic management at work in the European union: data protection, non-discrimination and collective rights. Int J of Comp Labour Law and Ind Relat 40(1):1–34

Alshahrani S, Wali E, Alshamsan AR et al (2022) Roadblocks in gender bias measurement for diachronic corpora. In: Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pp 140–148

Bai X, Wang A, Sucholutsky I et al (2024) Measuring implicit bias in explicitly unbiased large language models. arXiv preprint arXiv:2402.04105

Bartl M, Nissim M, Gatt A (2020) Unmasking contextual stereotypes: measuring and mitigating Bert's gender bias. arXiv preprint arXiv:2010.14534

Bassignana E, Basile V, Patti V et al (2018) Hurtlex: a multilingual lexicon of words to hurt. In: CEUR Workshop proceedings, CEUR-WS, pp 1–6

Basta C, Costa-Jussa MR, Casas N (2021) Extensive study on the underlying gender bias in contextualized word embeddings. Neural Comput Appl 33(8):3371–3384

Blodgett SL, Barocas S, Daumé III H et al (2020) Language (technology) is power: a critical survey of "bias" in NLP. In: Jurafsky D, Chai J, Schluter N et al. (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485, https://aclanthology.org/2020.acl-main.485

Bojanowski P, Grave E, Joulin A et al (2016) Enriching word vectors with subword information. Trans of the Assoc for Comput Linguist 5:135–146. https://doi.org/10.1162/tacl_a_00051

Bolukbasi T, Chang KW, Zou JY et al (2016) Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Adv Neural Inf Process Syst 29:1–9

Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):183–186

Chen Y, Mahoney C, Grasso I et al (2021) Gender bias and under-representation in natural language processing across human languages. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp 24–34

Crenshaw K (1989) Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. The Univ of Chic Legal Forum 1(8):139–167

Danks D, London AJ (2017) Algorithmic bias in autonomous systems. In: IJCAI, pp 4691–4697

Delobelle P, Tokpo EK, Calders T et al (2022) Measuring fairness with biased rulers: a comparative study on bias metrics for pre-trained language models. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp 1693–1706

Deutsch C, Paraboni I (2023) Authorship attribution using author profiling classifiers. Nat Lang Eng 29(1):110–137. https://doi.org/10.1017/S1351324921000383

Dev S, Phillips J (2019) Attenuating bias in word vectors. In: The 22nd international conference on artificial intelligence and statistics, PMLR, pp 879–887

Devlin J, Chang MW, Lee K et al (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Escouflaire L, Descampe A, Lits G et al (2023) Analyzing the semantic evolution of bias in French news articles using word embeddings. In: Digital Humanities Benelux 2023

Ethayarajh K, Duvenaud D, Hirst G (2019) Understanding undesirable word embedding associations. arXiv preprint arXiv:1908.06361

Fiske ST (2017) Prejudices in cultural contexts: shared stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion). Perspect Psychol Sci 12(5):791–799
Fort K, Névéol A, Dupont Y et al (2022) Use of a citizen science platform for the creation of a language resource to study bias in language models for French: a case study. In: 2nd LREC Workshop on Novel Incentives in Data Collection from People
Fosch-Villaronga E, Drukarch H, Khanna P et al (2022) Accounting for diversity in AI for medicine. Comput Law Secur Rev 47(105):735. https://doi.org/10.1016/j.clsr.2022.105735
Gallegos IO, Rossi RA, Barrow J et al (2024) Bias and fairness in large language models: a survey. Comput Linguist 50(3):1097–1179. https://doi.org/10.1162/coli_a_00524
Garg N, Schiebinger L, Jurafsky D et al (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc Natl Acad Sci 115(16):E3635–E3644
Goldfarb-Tarrant S, Marchant R, Sánchez RM et al (2020) Intrinsic bias metrics do not correlate with application bias. arXiv preprint arXiv:2012.15859
Gonen H, Kementchedjhieva Y, Goldberg Y (2019) How does grammatical gender affect noun representations in gender-marking languages? arXiv preprint arXiv:1910.14161
Grave E, Bojanowski P, Gupta P et al (2018) Learning word vectors for 157 languages. In: Calzolari N, Choukri K, Cieri C et al (eds) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan, https://aclanthology.org/L18-1550
Greenwald AG, McGhee DE, Schwartz JL (1998) Measuring individual differences in implicit cognition: the implicit association test. J Pers Soc Psychol 74(6):1464
Guo W, Caliskan A (2021) Detecting emergent intersectional biases: contextualized word embeddings contain a distribution of human-like biases. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp 122–133
Hardmeier C, Costa-jussà MR, Webster K et al (2021) How to write a bias statement: recommendations for submissions to the workshop on gender bias in NLP. arXiv preprint arXiv:2104.03026
Hellström T, Dignum V, Bensch S (2020) Bias in machine learning-what is it good for? In: International Workshop on New Foundations for Human-Centered AI (NeHuAI) co-located with 24th European Conference on Artificial Intelligence (ECAI 2020), Virtual (Santiago de Compostela, Spain), September 4, 2020, RWTH Aachen University, pp 3–10
High Level Group on, Non-discrimination, Equality and Diversity (2021) Guidance note on the collection and use of equality data based on racial or ethnic origin. https://commission.europa.eu/system/files/2022-02/guidance_note_on_the_collection_and_use_of_equality_data_based_on_racial_or_ethnic_origin_final.pdf
Hill F, Reichart R, Korhonen A (2014) Simlex-999: evaluating semantic models with (genuine) similarity estimation. Comput Linguist 41(4):665–695. https://doi.org/10.1162/COLI_a_00237
Hill Collins P, Bilge S (2020) Intersectionality, Key concepts, 2nd edn. Polity Press, Cambridge Medford, MA
Hovy D, Prabhumoye S (2021) Five sources of bias in natural language processing. Lang Linguist Compass 15(8):e12,432
Howard E (2006) The case for a considered hierarchy of discrimination grounds in EU law. Maastricht J Eur Comparat Law 13(4):445–470. https://doi.org/10.1177/1023263X0601300404
Howard E (2018) EU anti-discrimination law: has the CJEU stopped moving forward? Int J of Discrim and the Law 18(2–3):60–81. https://doi.org/10.1177/1358229118788454
Hutto C, Gilbert E (2014) Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the international AAAI conference on web and social media, pp 216–225
Ireni-Saban L, Sherman M (2020) Incorporating intersectionality into AI ethics. In: Giusti S, Piras E (eds) Democracy and fake news. Information manipulation and post-truth politics. Routledge, Milton Park, pp 41–52
Joshi P, Santy S, Budhiraja A et al (2020) The state and fate of linguistic diversity and inclusion in the NLP world. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 6282–6293
Kaneko M, Bollegala D (2022) Unmasking the mask–evaluating social biases in masked language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 11954–11962
Kaneko M, Imankulova A, Bollegala D et al (2022) Gender bias in masked language models for multiple languages. arXiv preprint arXiv:2205.00551
Katsarou S, Rodríguez-Gálvez B, Shanahan J (2022) Measuring gender bias in contextualized embeddings. In: Computer Sciences and Mathematics Forum, MDPI, p 3
Kaufmann MC, Krings F, Sczesny S (2016) Looking too old? how an older age appearance reduces chances of being hired. Br J Manag 27(4):727–739. https://doi.org/10.1111/1467-8551.12125

Kaufmann MC, Krings F, Zebrowitz LA et al (2017) Age bias in selection decisions: the role of facial appearance and fitness impressions. Front Psychol 8:2065. https://doi.org/10.3389/fpsyg.2017.02065
Kiritchenko S, Mohammad SM (2018) Examining gender and race bias in two hundred sentiment analysis systems. arXiv preprint arXiv:1805.04508
Kraft A, Zorn HP, Fecht P et al. (2022) Measuring gender bias in German language generation. INFORMATIK 2022
Kurita K, Vyas N, Pareek A et al (2019) Measuring bias in contextualized word representations. arXiv preprint arXiv:1906.07337
Kurpicz-Briki M (2020) Cultural differences in bias? Origin and gender bias in pre-trained German and French word embeddings. In: Proceedings of 5th SwissText and 16th KONVENS Joint Conference 2020
Kurpicz-Briki M (2023) More than a chatbot: language models demystified. Springer Nature, Cham
Kurpicz-Briki M, Leoni T (2021) A world full of stereotypes? further investigation on origin and gender bias in multi-lingual word embeddings. Front Big Data 4(625):290
Lauscher A, Glavaš G (2019) Are we consistently biased? multidimensional analysis of biases in distributional word vectors. arXiv preprint arXiv:1904.11783
Lauscher A, Takieddin R, Ponzetto SP et al (2020) Araweat: multidimensional analysis of biases in Arabic word embeddings. arXiv preprint arXiv:2011.01575
Leviant I, Reichart R (2015) Separated by an un-common language: towards judgment language informed vector space modeling. arXiv preprint arXiv:1508.00106
Lewis M, Lupyan G (2020) Gender stereotypes are reflected in the distributional structure of 25 languages. Nat Hum Behav 4(10):1021–1028
Manzini T, Yao Chong L, Black AW et al (2019) Black is to criminal as Caucasian is to police: detecting and removing multiclass bias in word embeddings. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 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, Minneapolis, Minnesota, pp 615–621. https://doi.org/10.18653/v1/N19-1062, https://aclanthology.org/N19-1062
Matthews A, Grasso I, Mahoney C et al (2021) Gender bias in natural language processing across human languages. In: Proceedings of the First Workshop on Trustworthy Natural Language Processing, pp 45–54
May C, Wang A, Bordia S et al (2019) On measuring social biases in sentence encoders. arXiv preprint arXiv:1903.10561
McCurdy K, Serbetci O (2020) Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. arXiv preprint arXiv:2005.08864
Meade N, Poole-Dayan E, Reddy S (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 (vol 1, Long Papers), pp 1878–1898
Mielke SJ, Eisner J (2019) Spell once, summon anywhere: a two-level open-vocabulary language model. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 6843–6850
Mikolov T, Sutskever I, Chen K et al (2013) Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 26
Mulligan DK, Kroll JA, Kohli N et al (2019) This thing called fairness: disciplinary confusion realizing a value in technology. Proc ACM Hum Comput Interact 3(CSCW):1–36. https://doi.org/10.1145/3359221
Mulsa RAC, Spanakis G (2020) Evaluating bias in Dutch word embeddings. arXiv preprint arXiv:2011.00244
Nangia N, Vania C, Bhalerao R et al (2020) Crows-pairs: a challenge dataset for measuring social biases in masked language models. arXiv preprint arXiv:2010.00133
Névéol A, Dupont Y, Bezançon J et al (2022) French crows-pairs: extension à une langue autre que l'anglais d'un corpus de mesure des biais sociétaux dans les modèles de langue masqués. In: Conférence sur le Traitement Automatique des Langues Naturelles (TALN)
Nozza D, Bianchi F, Hovy D et al (2021) Honest: measuring hurtful sentence completion in language models. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics
Ovalle A, Subramonian A, Gautam V et al (2023) Factoring the matrix of domination: a critical review and reimagination of intersectionality in AI fairness. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. ACM, Montréal QC Canada, pp 496–511. https://doi.org/10.1145/3600211.3604705
Page MJ, McKenzie JE, Bossuyt PM et al (2021) The Prisma 2020 statement: an updated guideline for reporting systematic reviews. Int J Surg 88(105):906
Papakyriakopoulos O, Hegelich S, Serrano JCM et al (2020) Bias in word embeddings. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp 446–457

Pennington J, Socher R, Manning C (2014) Glove: global vectors for word representation. In: EMNLP Proceedings, pp 1532–1543. https://doi.org/10.3115/v1/D14-1162
Pestova A (2021) Measuring gender bias in word embeddings for Russian language. In: Proceedings of Dialog 2021
Phillips W, Boroditsky L (2013) Can quirks of grammar affect the way you think? grammatical gender and object concepts. In: Proceedings of the 25th Annual Cognitive Science Society. Psychology Press, pp 928–933
Radford A, Narasimhan K, Salimans T et al (2018) Improving language understanding by generative pre-training. Published online
Remus R, Quasthoff U, Heyer G (2010) Sentiws-a publicly available german-language resource for sentiment analysis. In: LREC
Rigotti C, Fosch-Villaronga E (2024) Fairness, AI & recruitment. Comput Law Secur Rev 53(105):966. https://doi.org/10.1016/j.clsr.2024.105966
Risman BJ (2018) Where the millennials will take us: a new generation wrestles with the gender structure. Oxford University Press, Oxford
Sabbaghi SO, Caliskan A (2022) Measuring gender bias in word embeddings of gendered languages requires disentangling grammatical gender signals. arXiv preprint arXiv:2206.01691
Sahlgren M, Olsson F (2019) Gender bias in pretrained swedish embeddings. In: Proceedings of the 22nd Nordic Conference on computational linguistics, pp 35–43
Salamanca G, Pereira L (2013) Prestigio y estigmatización de 60 nombres propios en 40 sujetos de nivel educacional superior. Universum (Talca) 28(2):35–57
Salazar J, Liang D, Nguyen TQ et al (2019) Masked language model scoring. arXiv preprint arXiv:1910.14659
Sedoc J, Ungar L (2019) The role of protected class word lists in bias identification of contextualized word representations. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pp 55–61
Sheng E, Chang KW, Natarajan P et al (2019) The woman worked as a babysitter: on biases in language generation. arXiv preprint arXiv:1909.01326
Simon P (2015) The choice of ignorance: the debate on ethnic and racial statistics in France. In: Simon P, Piché V, Gagnon AA (eds) Social statistics and ethnic diversity. Series title: IMISCOE Research Series. Springer International Publishing, Cham, pp 65–87. https://doi.org/10.1007/978-3-319-20095-8_4
Smith EM, Hall M, Kambadur M et al (2022) " I'm sorry to hear that": finding new biases in language models with a holistic descriptor dataset. arXiv preprint arXiv:2205.09209
Spliethöver M, Wachsmuth H (2020) Argument from old man's view: assessing social bias in argumentation. arXiv preprint arXiv:2011.12014
Sun T, Gaut A, Tang S et al (2019) Mitigating gender bias in natural language processing: literature review. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 1630–1640
Telles EE (2014) Pigmentocracies: ethnicity, race, and color in Latin America. The University of North Carolina Press, Chapel Hill
Toney-Wails A, Caliskan A (2020) Valnorm quantifies semantics to reveal consistent valence biases across languages and over centuries. arXiv preprint arXiv:2006.03950
Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. Adv Neural Inf Process Syst 30:1–11
Wachter S, Mittelstadt B, Russell C (2021) Why fairness cannot be automated: bridging the gap between EU non-discrimination law and AI. Comput Law Secur Rev 41(105):567. https://doi.org/10.1016/j.clsr.2021.105567
Wagner J, Zarrieß S (2022) Do gender neutral affixes naturally reduce gender bias in static word embeddings? In: Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022), pp 88–97
Wambsganss T, Swamy V, Rietsche R et al (2022) Bias at a second glance: a deep dive into bias for german educational peer-review data modeling. arXiv preprint arXiv:2209.10335
Weerts H, Xenidis R, Tarissan F et al (2023) Algorithmic unfairness through the lens of EU non-discrimination law: or why the law is not a decision tree. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp 805–816
Wevers M (2019) Using word embeddings to examine gender bias in Dutch newspapers, 1950-1990. arXiv preprint arXiv:1907.08922
Wilkinson DY, King G (1987) Conceptual and methodological issues in the use of race as a variable: policy implications. Milbank Q 65(1):56–71
Xenidis R (2020) Tuning EU equality law to algorithmic discrimination: three pathways to resilience. Maastricht J Eur Comput Law 27(6):736–758. https://doi.org/10.1177/1023263X20982173
Zhao J, Mukherjee S, Hosseini S et al (2020) Gender bias in multilingual embeddings and cross-lingual transfer. arXiv preprint arXiv:2005.00699

Zhou P, Shi W, Zhao J et al (2019) Examining gender bias in languages with grammatical gender. arXiv preprint arXiv:1909.02224

Zmigrod R, Mielke SJ, Wallach H et al (2019) Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. arXiv preprint arXiv:1906.04571

Authors and Affiliations

Alexandre Puttick[1] · Catherine Ikae[1] · Carlotta Rigotti[2] · Eduard Fosch-Villaronga[2] · Mark W. Kharas[3] · Roger A. Søraa[3] · Mascha Kurpicz-Briki[1]

✉ Alexandre Puttick
alexandre.puttick@bfh.ch

Catherine Ikae
catherine.ikae@bfh.ch

Carlotta Rigotti
c.rigotti@law.leidenuniv.nl

Eduard Fosch-Villaronga
e.fosch.villaronga@law.leidenuniv.nl

Mark W. Kharas
mark.w.kharas@ntnu.no

Roger A. Søraa
roger.soraa@ntnu.no

Mascha Kurpicz-Briki
mascha.kurpicz@bfh.ch

[1] Applied Machine Intelligence Research Group, Bern University of Applied Sciences, Hoeheweg 80, 2502 Biel/Bienne, Switzerland

[2] eLaw Center for Law and Digital Technologies, Leiden University, Steenschuur 25, 2311 ES Leiden, Netherlands

[3] Department of Interdisciplinary Studies of Culture, Norwegian University of Science and Technology NTNU, 7491 Trondheim, Norway