Foundations and Trends® in Information Retrieval

Understanding and Mitigating Gender Bias in Information Retrieval Systems

Suggested Citation: Shirin Seyedsalehi, Amin Bigdeli, Negar Arabzadeh, Batool AlMousawi, Zack Marshall, Morteza Zihayat and Ebrahim Bagheri (2024), "Understanding and Mitigating Gender Bias in Information Retrieval Systems", Foundations and Trends® in Information Retrieval: Vol. XX, No. XX, pp 1–175. DOI: XXX.

Shirin Seyedsalehi

Amin Bigdeli

Negar Arabzadeh

Batool AlMousawi

Zack Marshall

Morteza Zihayat

Ebrahim Bagheri

now
the essence of knowledge
Boston — Delft

Contents

Understanding and Mitigating Gender Bias in Information Retrieval Systems

Shirin Seyedsalehi[1], Amin Bigdeli[2], Negar Arabzadeh[2], Batool AlMousawi[3], Zack Marshall[3], Morteza Zihayat[1] and Ebrahim Bagheri[1]

[1] *Toronto Metropolitan University*
[2] *University of Waterloo*
[3] *University of Calgary*

ABSTRACT

Gender bias is a pervasive issue that continues to influence various aspects of society, including the outcomes of information retrieval (IR) systems. As these systems become increasingly integral to accessing and navigating the vast amounts of information available today, the need to understand and mitigate gender bias within them is paramount. This book provides a comprehensive examination of the origins, manifestations, and consequences of gender bias in IR systems, as well as the current methodologies employed to address these biases.

Theoretical frameworks surrounding gender and its representation in artificial intelligence (AI) systems are explored, particularly focusing on how traditional gender binaries are perpetuated and reinforced through data and algorithmic processes. Metrics and methodologies used to identify and

Shirin Seyedsalehi, Amin Bigdeli, Negar Arabzadeh, Batool AlMousawi, Zack Marshall, Morteza Zihayat and Ebrahim Bagheri (2024), “Understanding and Mitigating Gender Bias in Information Retrieval Systems”, Foundations and Trends® in Information Retrieval: Vol. XX, No. XX, pp 1–175. DOI: XXX.

measure gender bias within IR systems are then analyzed, offering a detailed evaluation of existing approaches and their limitations.

Subsequent chapters address the sources of gender bias, including biased input queries, retrieval methods, and gold standard datasets. Various data-driven and method-level debiasing strategies are presented, including techniques for debiasing neural embeddings and algorithmic approaches aimed at reducing bias in IR system outputs. The book concludes with a discussion of the challenges and limitations faced by current debiasing efforts and provides insights into future research directions that could lead to more equitable and inclusive IR systems.

This book serves as a valuable resource for researchers, practitioners, and students in the fields of information retrieval, artificial intelligence, and data science, providing the knowledge and tools needed to address gender bias and contribute to the development of fair and unbiased information systems.

1

Introduction

1.1 Information Retrieval (IR) Systems

Information Retrieval (IR) systems are fundamental to the digital era, and crucial for navigating the vast data landscape of today's world. From simple web searches to sophisticated data analytics in corporate environments, IR systems are integral to modern life and provide the tools necessary for personal and professional decision-making. IR systems do not just facilitate over 1.2 trillion searches per year on a platform like Google (Internet Live Stats, 2024) but also significantly impact various sectors such as:

- **Healthcare.** In healthcare, IR systems manage extensive patient records and research databases, enabling medical professionals to access vital information swiftly. For instance, databases like PubMed offer access to medical research, facilitating better patient care and fostering the rapid development of medical knowledge (Medicine, 2024).
- **Finance and Banking.** Financial sectors utilize IR to analyze market trends and monitor transactions. Tools provided by Bloomberg and Reuters help professionals sift through large

datasets to find critical information on market developments, economic reports, and investment analytics, supporting quick and informed financial decisions (L.P., 2024; Reuters, 2024).

- **Legal.** IR systems such as LexisNexis and Westlaw are indispensable in the legal arena. They allow legal professionals to efficiently search through vast quantities of legal documents, case law, and statutes, essential for case preparation, conducting due diligence, and ensuring comprehensive legal research (LexisNexis, 2024; Westlaw, 2024).
- **Academic Research.** IR systems are also crucial in academia, where platforms like Google Scholar and JSTOR enable researchers to navigate through countless scholarly articles and publications. This access supports various academic disciplines, enhancing research capabilities and fostering educational advancement (Google, 2024; ITHAKA, 2024).

Such systems have deep impacts on different aspects of society. The **economic implications** of IR systems are vast, influencing sectors from e-commerce to online advertising. They drive consumer behavior, facilitate transactions, and are instrumental in strategic business decisions, impacting billions in daily commerce. **Technological advancements** in IR have paralleled the rapid evolution of computing power and data science methodologies. Today's IR systems employ sophisticated algorithms and machine-learning techniques to improve accuracy and user experience. Furthermore, IR systems profoundly shape societal interactions and access to information, influencing education, politics, and social dynamics. In **education**, IR systems provide students and academics access to a wide array of resources, transforming how knowledge is acquired and shared. The availability of digital libraries and online courses has democratized education, making learning more accessible globally. **Politically**, IR systems play a critical role in shaping public opinion and electoral outcomes by controlling the flow of news and information. Their ability to highlight or suppress information can alter perceptions and influence decisions on a large scale. **Culturally**, IR systems facilitate the global exchange of ideas and values, promoting cross-cultural understanding and cooperation (Taksa and Flomenbaum, 2009). They have become platforms for cultural expression and identity exploration, contributing to the global cultural mosaic.

1.2 Biases and IR Systems

Information Retrieval (IR) systems, while immensely beneficial, are not immune to the influence of biases that can skew results and perpetuate societal inequalities. These biases arise from various sources including data, algorithm design, and human factors involved in the development and maintenance of such systems. Biases in IR systems can have profound implications across multiple sectors by reinforcing stereotypes and exacerbating social prejudices. Below, we explore several high-profile examples that illustrate the detrimental effects of these biases.

- **Employment and Job Recommendation Systems** One notable example involves gender bias in job recommendation algorithms. Studies have shown that certain algorithms tend to favor male candidates over equally qualified female candidates. This reflects and perpetuates existing gender disparities in job markets. For instance, a research conducted by Amazon had to scrap their AI recruiting tool because it showed bias against women. The system learned to penalize resumes that included the word "women's," as in "women's chess club captain," and it downgraded graduates of two all-women's colleges (Dastin, 2018).

- **Credit and Loan Approvals** Biases in IR systems also affect financial decisions like credit scoring and loan approvals. An investigation into Apple Card's algorithm revealed it offered higher credit limits to men than to women under similar financial circumstances. This incident sparked a broader discussion about the transparency and fairness of algorithms in financial services (Nicas, 2019).

- **Healthcare Diagnostics** In healthcare, biases in IR systems can lead to life-threatening consequences. Research has indicated that certain diagnostic algorithms prioritize the care of white patients over equally sick patients from minority groups due to biases in the training data. For example, a widely used healthcare algorithm was found to be less likely to refer Black patients than white patients for higher-quality care, even when they were equally ill (Obermeyer *et al.*, 2019).

- **Law Enforcement and Judicial Systems** In law enforcement, predictive policing systems have come under scrutiny for perpetuating racial biases. These systems often target minority-heavy areas more aggressively, leading to a disproportionate number of arrests and convictions in these communities. Similarly, algorithms used to predict future criminal behavior for parole decisions have been criticized for being biased against people of color (Angwin *et al.*, 2016).

Tackling biases in IR systems is not only a technological imperative but also a moral obligation. Given the critical role that these systems play in shaping perceptions and decision-making processes in society, ensuring fairness, equity, and justice in digital interactions becomes paramount.

Several studies have explored bias in practical, applied industry contexts, highlighting both challenges and potential solutions. For instance, in (Bogen and Rieke, 2018), the authors provide recommendations to increase transparency and oversight in hiring technologies to reduce the potential harm these tools can cause. They advocate for independent audits by vendors and employers and suggest that regulators update laws to address the capabilities and risks of modern hiring technologies. The report emphasizes that without intentional intervention, these technologies could reinforce existing inequalities. Nevertheless, it argues that predictive tools also present opportunities to improve diversity if they are actively designed to address historical inequities. This balance between innovation and accountability is crucial as these technologies increasingly influence employment opportunities.

In the realm of recommendation systems, the authors in (Wu *et al.*, 2021) introduce FairRec, a model designed to reduce bias in news recommendations while maintaining performance levels. Traditional recommendation systems often amplify biases by capturing patterns linked to sensitive attributes like gender. FairRec mitigates this by decomposing user interests into two components: a bias-aware embedding that captures attribute-specific biases and a bias-free embedding focused on neutral interests. The model employs adversarial learning to minimize bias in the bias-free embedding and uses orthogonality regularization to keep the two embeddings distinct. Only the bias-free embedding is used in the final ranking, ensuring recommendations are independent of sensitive attributes.

Lastly, the authors in (Binns *et al.*, 2018) explored perceptions of justice in algorithmic decision-making. Through lab and online experiments, the study explored how various explanation styles—such as case-based, demographic, input influence, and sensitivity—affect people's sense of fairness, dignity, and accountability in scenarios like loan approvals and insurance pricing. Results indicate that people's perceptions of justice are shaped by their understanding of the decision-making process and whether they view the factors considered as appropriate. However, repeated exposure to a single explanation style led participants to focus more on scenario details than on specific explanation types. This study highlights the complexities involved in designing explanations that foster a sense of fairness and accountability in algorithmic systems, emphasizing that no single explanation style fits all needs and that users may be reluctant to assign justice or moral responsibility to machine-based decisions.

In addition, the importance of addressing biases in IR systems has been significantly recognized by the research community, prompting a vigorous response aimed at understanding and mitigating these biases. This response has been multi-faceted, focusing on various aspects of bias in IR systems—from identifying the sources of biases and understanding how they are injected into the systems, to exploring ways in which these biases are amplified and spread through societal interactions. Researchers have investigated the mechanisms through which biases are introduced into IR systems. This often originates from the data used to train algorithms, where historical inequalities or skewed data representation lead to biased decision-making processes (Barocas and Selbst, 2016; Mehrabi *et al.*, 2021). Studies have shown how machine learning algorithms can inadvertently learn and perpetuate these biases if not properly checked (Zhao *et al.*, 2017). Moreover, the research focuses on how once biases are injected, they can be intensified by the algorithms through their iterative nature. For example, feedback loops where biased outputs are used as new training data can further entrench and exacerbate these biases (Baeza-Yates, 2018). Understanding these dynamics is crucial for developing effective mitigation strategies (Friedman and Nissenbaum, 1996).

A significant portion of recent research has been devoted to developing methodologies to prevent the spread of biases. These include algorithmic fairness approaches, bias audits, and the use of fairness-enhancing interventions in the algorithmic design (Chouldechova, 2017; Holstein *et al.*, 2019). Researchers are exploring both technical solutions, such as the redesign of algorithms, and policy-based approaches, such as regulatory frameworks and transparency guidelines (Barocas *et al.*, 2020; Binns, 2018).

1.3 Chapter Breakdown

This book aims to contribute significantly to this ongoing discourse by providing a comprehensive overview of how biases in IR systems can be understood and addressed. Each chapter is dedicated to exploring a different aspect of bias in IR, from theoretical underpinnings to practical applications and case studies, thus offering a holistic view of current strategies and future directions in bias mitigation. The textbook is structured to provide a holistic approach to understanding and mitigating gender bias in Information Retrieval (IR) systems. It is composed of a series of chapters that progressively investigate various dimensions of gender bias, ranging from theoretical frameworks to practical de-biasing methods.

1.3.1 Chapter 2: Framing Sex, Gender, and Gender Diversity

Having outlined the biases present in information retrieval (IR) systems, we take the first step toward addressing these issues by looking at how AI systems interpret concepts like sex and gender. This next chapter explores how these interpretations can often reinforce social biases, helping us build a clear foundation for understanding gender bias in IR.

1.3.2 Chapter 3: Gendered Information Retrieval Systems: Metrics and Measurements

Metrics and measurements used to identify and quantify gender biases in IR systems are outlined in this chapter. The chapter discusses various approaches to assess how these systems handle fairness in algorithmic processing and result ranking.

1.3.3 Chapter 4: Understanding the Sources of Gender Bias in IR Systems

This chapter explores the origins of gender biases in IR systems. It analyzes how biases are integrated into algorithms through data training processes and the design of algorithms themselves. The chapter discusses both inadvertent and systematic insertion of biases during the development phases of IR systems.

1.3.4 Chapter 5: Data-Driven De-Biasing Methods

Focusing on practical approaches, this chapter introduces methods for data-driven bias mitigation. It covers techniques such as data augmentation, modification of training datasets, and algorithmic adjustments aimed at reducing the gender bias inherent in IR systems.

1.3.5 Chapter 6: De-biasing of Neural Embeddings

Specific techniques for de-biasing neural network embeddings are covered. This chapter offers the details and the technical aspects of neural networks that process, providing insights into how these can be adjusted to mitigate biases.

1.3.6 Chapter 7: Method-Level De-biasing

This chapter extends the discussion on bias mitigation by focusing on specific methodologies that can be applied at different levels of IR system development. It includes case studies and examples where these methods have been successfully implemented.

1.3.7 Chapter 8: Challenges, Limitations, and Future Directions

The concluding chapter discusses the ongoing challenges in fully addressing gender bias in IR systems, the limitations of current approaches, and the potential future research directions that could lead to more comprehensive solutions.

The structure of this textbook is designed to equip researchers, practitioners, and students with a thorough understanding of the complex nature of gender biases in IR systems and provides a detailed guide on existing strategies to address these biases. Each chapter builds on the previous one, ensuring a comprehensive learning path for the reader.

2

Framing Sex, Gender, and Gender Diversity

To address gender bias effectively in information retrieval (IR) systems, it is crucial to first understand the underlying concepts of sex and gender and how they are often misrecognized and misrepresented in artificial intelligence (AI). This chapter explores these constructs, providing context for how biases in IR systems may arise from simplistic or incorrect assumptions about gender. By examining the ways AI systems interpret and apply these constructs, we set a foundation for understanding the broader impact of gender bias on IR outcomes and the need for frameworks that account for diversity and complexity. This exploration is key to building IR systems that address risks of harm while supporting fair and accurate information access.

2.1 Sex and Gender in the Context of AI

The United Nations Educational, Scientific, and Cultural Organization's International Research Centre on Artificial Intelligence (UNESCO IR-CAI) recently underlined the ways that AI systems perpetuate and amplify human, structural, and social biases (UNESCO, 2024). These biases, which include sexism and gender binarism, serve to reinforce systems of domination and marginalization that structure inequities

(Krieger, 2020; Shrestha and Das, 2022). These systems stand to actively harm those impacted by both discriminatory practices and legislation, resulting in experiences of embodied harm (Krieger, 2020). To best understand how AI is both complicit and actively involved in engendering harm, we invite our readers to explore how AI conceptualizes sex and gender, the limitations of these conceptualizations, and potential solutions for developing equity-informed and gender-aware (Pinney *et al.*, 2023) AI systems moving forward.

2.1.1 Exploring the Definitional Minefield of Sex and Gender: Key Concepts and Definitions

To begin our discussion, it is helpful to explicitly define the constructs of sex and gender in the context of AI. Of course, the distinction between sex and gender is subject to definitional and conceptual ambiguity that varies tremendously. In their book Artificial Knowing: Gender and the Thinking Machine, Alison Adam (1998) describes the distinction between sex and gender as a "definitional minefield" arguing that "... trying to tie gender to an ideal of supposedly uncontroversial biological sex is problematic in many ways" (p. 22). This section focuses in on these nuances, exploring differences that complicate the adoption of universal constructions of sex or gender in AI systems.

Sex

Across discourses, sex is often reduced to a purely biological construct, encompassing the cellular, anatomical, chromosomal, hormonal, and genetic differences between males and females (Johnson and Repta, 2012). Defaulting to genetic essentialism is complicated by the tapestry of genetic processes that inform sex determination (Ainsworth, 2015). Moreover, while binary female-male categorizations are useful for certain lines of inquiry, they are limited in their ability to generalize to all females and males (Richardson, 2022). Understandings of sex as a biological construct must encompass the full range of human variation and diversity present within populations; binary conceptualizations fail to do this.

Chromosomal configurations outside of XX and XY, including XXX, XO, XXY, and XYY, occur at a rate as high as 1 person out of every 100 people (Arboleda *et al.*, 2014). Individuals with chromosomal configurations outside of XX and XY are often labelled under the intersex umbrella within the common vernacular, however, increased attention has been placed on developing a modern nomenclature in opposition to pejorative, controversial diagnostic labels. In collaboration with the participants of the International Consensus Conference on Intersex, Lee *et al.* (2006) released a consensus statement on the topic of care for individuals experiencing chromosomal atypicality. Regarding the nomenclature ascribed to individuals, authors of the Consensus Statement on Management of Intersex Disorders recommend the use of the term "disorders of sex development" alongside descriptive accounts of the relevant genetic etiology when possible (Lee *et al.*, 2006). However, since the development of these recommendations, InterACT (2018), a non-profit organization advocating for the rights of intersex youth, has rejected the term due to its emphasis on pathologization. InterACT almost exclusively uses the term "intersex" to refer to individuals with differences in sex traits, however, they maintain the position that terminology in this arena is primarily a matter of individual choice. With this in mind, we will adopt the term "intersex" in this publication, mirroring InterACT's recommendation.

Binary sex categorizations do not capture the full range of biological diversity present within humans and often exclude intersex and trans people. However, it is also important to recognize that binary sex categorizations reduce much of the diversity within categorizations of females and males, in addition to differences across categorizations. Johnson and Repta (2012) argue that variation within binary sex designations, including differences in metabolism, lung capacity, stress response, brain function, and bone size necessitates broader conceptualizations of sex.

References to external genitalia and physical characteristics continue to inform common constructions of sex, gender, and related sex differences. For example, the emergence of secondary sexual characteristics (including body hair, muscle mass, and breast development) during puberty primarily signals the beginning of reproductive age (Richards and Hawley, 2011). Evidence suggests that body composition varies by sex, such that females on average carry more fat mass relative to males and are more likely to accumulate fat around their hips and thighs (Bredella, 2017). Moreover, despite having many of the same hormones (androgens and estrogens), the concentration of these hormones typically found within the body and their relative interactions with organ systems demonstrate sex-based variation (Svechnikov and Söder, 2008).

Among physical characteristics, faces are an attribute of particular interest given their ability to subtly indicate sex or gender. Researchers have described the importance of faces in providing sex and gender-related information, arguing that social conventions and gender presentation may play a role in cueing sex or gender (González-Álvarez and Sos-Peña, 2022). However, even in the absence of sociocultural cues, humans demonstrate remarkable accuracy in determining sex from isolated faces; some estimates put that accuracy as ranging from 96-98% (Bruce *et al.*, 1993). Binary sex categorization of faces devoid of sociocultural sex markers is a function of two elements: facial morphology and surface reflectance (Meinhardt-Injac *et al.*, 2013; Russell *et al.*, 2006). In recent experiments, researchers identified the presence of both these components as central to accurate sex categorizations, with surface reflectance being associated with more categorical sex perception (González-Álvarez and Sos-Peña, 2022).

A study on sex differences in facial morphology led by Bannister *et al.* (2022) found that on average, male faces were 7.30% larger than female faces. Moreover, the researchers (Bannister *et al.*, 2022) describe significant sex differences in facial shape, especially in the brow, nose, cheek, and jaw regions. Russell (2009) describes facial contrast (the relative degree of facial luminance and the resulting contrast it creates) as influencing the perception of facial gender. It is thought that females generally have higher facial contrast than males; thus, isolated androgynous faces can be made to read more female or male by increasing or decreasing facial contrast. Russell (2009) elaborates on the ways cosmetics increase facial contrast, amplifying gendering by exaggerating this attribute. Taken collectively, this evidence suggests that faces are marked by some degree of sexual dimorphism, however, it is important to recognize that the magnitude of these differences varies across populations (Kleisner *et al.*, 2021). According to Kleisner *et al.* (2021), cultural facial preferences, variations in facial shape, and differences in body height cannot explain differences in facial dimorphism around the world, and should not be essentialized.

These examples highlight how sex may be reflected in various physical body characteristics. However, in focusing exclusively on inter-sex differences, intra-sex variation is undermined. Reducing the diversity within sex categorizations has real-world applications. For example, in 2023, World Athletics imposed new hormonal regulations for female track and field athletes (Bowman-Smart *et al.*, 2024). These regulations require some athletes to artificially reduce their body's natural testosterone levels to compete within the female category (Bowman-Smart *et al.*, 2024). Among the athletes impacted are people assigned female at birth and raised as women. The Canadian Broadcasting Corporation and National Public Radio (2024) discuss these stories in their podcast Tested, highlighting the century-long history of policing the bodies of female athletes in women's sports and encouraging audiences to contemplate fairness and equity.

Sex contextualism provides a useful, ethical framework to attend to sex-related variables without defaulting to a sex ontology (Richardson, 2022). Richardson (2022) argues that the relevancy (or lack thereof) and meaning of sex-related variables is context-specific. Moreover, male and female categorizations can be operationalized variably across disciplines. As such, it is possible to explore male-female differences in some works, collapsing the categorizations in others, and further differentiating categorizations where necessary. According to DiMarco *et al.* (2022), in practice, sex contextualism is "...defining and analyzing sex-related biological variables within well-specified contexts, including individual life history, social and physical environment, laboratory and technological constraints, species, strain, developmental stage or age, and level of biological analysis. This reflects what biologists already well know: that factors associated with sex-differentiated biological pathways cut across many different forms of biological organization, from chromosomes to tissues, hormones, and organs." (p. 2). To expand on this idea further, Richardson (2022) raises an example, in which researchers explored the impact of estrogen injections in the treatment of wound sepsis. In this study, Bösch *et al.* (2018) stratified their mouse model into four "sexes" based on estrogen levels. Within the wound sepsis research program, contextualizing sex variation in specific reference to estrogen levels enabled investigators to explore additional variables that mediated the effect of the estrogen treatment on wound sepsis, enhancing scientific precision in turn (Bösch *et al.*, 2018). In their study, it was neither helpful nor adequate to contextualize sex as binary; thus, alternative classification schemes were employed. Pape *et al.* (2024) further emphasize the merits of sex contextualism in enhancing scientific rigor and precision, arguing that researchers interested in these values might choose to operationalize sex by specifying measurable attributes, including hormone levels or chromosomal configurations, as opposed to proxy binary categorizations meant to encompass diverse biological pathways. In a sex contextualist view, binary male-female categorizations are not the only sex categorizations that could exist, and any reported sex differences and variations are approached with enhanced specificity and transparency. Broadly, sex contextualization advocates for reflective practice that validates more diverse understandings of sex (Richardson, 2022).

As a final point of interest, it is important to diverge from strictly biological accounts of sex and acknowledge dimensions of social construction. Johnson and Repta (2012) describe how conceptualizations of sex exist relative to place and time; specifically, they draw attention to how refinements in measurement techniques have allowed us to better understand the full range of phenotypic diversity present within humans. Laqueur (1992) identified a pervasive, single-sex perspective originating in Antiquity Greece that conceptualized females as undeveloped males. This interpretation was first contested in the Western world during the 18th and 19th centuries driven by an increased interest in determining sex differences based on reproductive organs (Laqueur, 1992). Contemporary bioessentialist notions about sex can be traced to Enlightenment Europe when philosophy began to fixate on the natural rights of men and women (D'Ignazio and Klein, 2020). Zimman (2014) reflects on this history, asserting "...it is clear that even the basic idea that the penis and vagina are different (let alone opposite) body parts, rather than external and internal versions of the same organ, is the product of a particular culture at a particular point in time. So is the belief that the body comes primarily or exclusively in two types-female and male" (p. 17). This construction is relevant because it informs current understandings of sex assigned according to external genitalia.

Taken collectively, sex is a multidimensional construct that exists independent of gender but is also a component of gender. Positioning sex as an exclusively "precultural or pre-social" state of being emboldens biological essentialism to enforce the gender binary (Zimman, 2014).

Gender

Gender is understood as a multidimensional psychosocial concept encompassing gender identity, gender roles, gender relations, and institutionalized gender (Tadiri *et al.*, 2021). Gender identity includes a felt sense of self as feminine, masculine, both, or neither, and the ways people express their gender (or gender expression) (Chang and Wildman, 2017). Bolte *et al.* (2021) describe gender self-concepts using a similar framing, taking time to differentiate between sex assigned at birth and current sex phenotype and emphasize the role of internalized gender roles in gender construction. The conceptualization of gender in this text attends to these nuances, recognizing that gender is both socially constructed and has material consequences (Namaste, 2000). Socially, the distribution of power, resources, and opportunity influences gendered norms and expectations. Experientially, gender is subjectively embodied, such that self-identification with a given gender identity can occur for a variety of reasons. This section is dedicated to further exploring concepts related to gender identity and gender roles.

Gender Identity and Gender Subjectivity

Ashley (2023) defines gender identity as including "...the grounds of gender self-categorization, the strength of one's identification, and the totality of feelings about self-categorization. Gender identity can be felt strongly or weakly, notably because of the apprehended strength and cohesiveness of gender subjectivity" (p. 1054). Under this account, a person may choose to identify in any number of ways depending on the degree of their embodiment. Common identifications include cisgender man/woman, which encompasses people whose sex assigned at birth aligns with their gender identity, transgender (trans) man/woman, which comprises people whose sex assigned at birth does not align with their gender identity, and non-binary individuals, who do not (necessarily) identity within the gender binary and may or may not identify with sex assigned at birth. These identifications only begin to capture the full diversity present within populations; after all, all people have a gender identity.

In contrast, gender subjectivity is defined as "The totality of our gendered experiences... and forms the basic substrate of gender identity" Ashley (2023) (p. 1059). Postmodern scholars conceptualize gender subjectivity as socially constructed and informed by embodied cultural discourses (Kruks, 1992). However, feminist scholars also create space for individual agency and consciousness within the subjective experience (Kruks, 1992). In so doing, two incongruent accounts of gender can simultaneously co-exist. To illustrate this point, Ashley (2023) describes how a cisgender woman may "... understand herself as a woman because of the body she was born in without suggesting that transgender women are any less women because of the bodies they were born in" (p. 1054). Ashley (2023) argues that our gendered experiences do not define us; it is our interpretations of gendered experiences (which vary in strength, consistency, and ambiguity) that give form to gendered beings.

This theory of the phenomenological constitution of gender identity presents a novel constructivist account of gender identity that reconciles the seemingly incompatible experiences of gender that vary from one individual to another. Ashley (2023) employs an architectural analogy to position gender subjectivity as the materials that inform constructions of gender identity. Uniquely, this theory authenticates and validates the experiences of a diversity of people, including trans and gender-diverse people, without defaulting to an "othering" rhetoric that entrenches marginalization.

Gender Expression

A further concept of interest relates to gender expression: the broad presentation of masculinity, femininity, or androgyny (Beltz *et al.*, 2021).

Beltz *et al.* (2021) describe gender expression as encompassing an individual's knowledge of, understanding of, and relative prescription to gender norms, expectations, and attitudes. The expression of gender may align with an individual's gender identity, but it may also diverge. For example, a cisgender man may opt to adopt a more stereotypically feminine presentation without it impeding his self-concept as a man. Importantly, gender expression, and gender norms, are also fluid, changing over time and across contexts. One study suggests that expressions of femininity increase with age; interestingly, this holds for both men and women (Hyde *et al.*, 1991). Moreover, a person may choose to present in a way that more closely mirrors their gender identity when in certain environments, but present more androgynously in others. Ghabrial (2019) and Frost (2011) describe how strategic conformity to hegemonic dogmas, especially for sexual and gender minorities, may be potentially protective against social stigma, discrimination, and violence, especially in environmental contexts marked by hostility and threat. However, Anderson (2020) emphasizes how this conduct, even when selectively protective, carries significant social, psychological, and emotional costs: a topic in need of greater academic attention.

There are several demographic and environmental factors associated with gender expression. For instance, one study describes greater gender non-conformity among people with higher education and household income (Sandfort *et al.*, 2021). Further, the authors found higher levels of gender non-conformity in gay men and bisexual women, relative to heterosexual men, women, and lesbian women (Sandfort *et al.*, 2021). Gender expression may also interact with other intersectional factors, including class, rural/urban environments, and culture (Frable, 1997; Sandfort *et al.*, 2021). For example, some research on sexual orientation perception (sometimes called "gaydar research") suggests that humans can judge sexual orientation at levels that exceed chance; evidence points to differences in adornment, behaviors, speech, and appearance as important indicators in making these judgments (Rule, 2017). Concerningly, AI neural networks demonstrate accuracy that exceeds humans in making judgments about sexual orientation from static facial images (Wang and Kosinski, 2018). This is congruent with the assumption that gender expression is, in part, informed by the sociocultural norms and understandings that outline parameters for its expression and that these concepts can be embedded within AI systems (Sandfort *et al.*, 2021). Gender and its related expressions are multidimensional constructs informed by dynamic social identities, and it is important to attend to these constructs with nuance (Frable, 1997).

Gender Roles

In Western societies, gender roles encompass societal-level beliefs and stereotypes related to differing expectations and understandings of bioessentialist roles of women (and the nature of femininity) and men (and the nature of masculinity) (Alesina *et al.*, 2013; Blackstone, 2003). Femininity is devalued, therefore people who express or embody femininity are also devalued. This is a particular problem for many women. These roles reflect society's values about sex and gender, often perpetuating rigid binary conceptualizations of both in turn (Blackstone, 2003). Importantly, these roles are non-essential, and are dynamic across time, culture, and geography (Adam, 1998). Gender roles are institutionalized through education, media, religion, politics, and social systems however, individuals may interact with these roles in a diversity of environmentally mediated ways (Blackstone, 2003; Johnson and Repta, 2012). For example, a society may expect women to disproportionately shoulder the domestic responsibilities associated with maintaining a household, while also barring them from joining the workforce. Other societies may adopt more egalitarian roles, advocating for balanced expectations. A woman in either society may reject or embody these roles, the extent to which either occurs varying from one individual to another in response to their interpretations of gendered experiences alongside structural constraints impacting the amount of individual choice one might have.

A central tenet of feminist scholarship centers around the idea that social attributions related to femininity and masculinity are charged entities (Adam, 1998). In Western societies, the Abrahamic tradition ascribed rationality a masculine trait and irrationality as a feminine one. This extends to attributions of power and status with masculinity, and weakness and negative status with femininity (Adam, 1998). These designations help perpetuate deep-rooted gender hierarchies that impact all of society by privileging people seen as masculine or who express masculinity, while simultaneously marginalizing people characterized as feminine or who express femininity.

2.1.2 Factors Influencing Interactions with Sex and Gender

Conceptualizations of sex and gender do not exist in a vacuum; they are informed by language and cultural norms. We will briefly touch on these factors now.

Language

The Whorf-Sapir hypothesis proposes that differences between cultures can be attributed to differences in language use (Sapir, 2023; Whorf, 2012). The concept of linguistic relativity is linked to this hypothesis, which argues that our perceptions of the world around us are shaped by the language we use.

The Dubenko (2022) study of nine (gendered) languages within the Indo-European family found that cases of gender universality within nouns were often motivated by binary masculine/feminine connotations, cultural traditions, and symbolic gender archetypes. Gender universalities that enforce gender binaries may pattern archetypical mental representations. This research contributes to the growing body of psycholinguistic works that elucidates the impact of grammatical gender on perception in both two-gender and three-gender languages (Dubenko, 2022). Grammatical gender is not the only factor involved in these distinctions. For example, languages such as English use a system of pronominal agreement where “different third-person singular pronouns are used for male and female humans” (Jakiela and Ozier (2020), p. 9). Interestingly, there is research to suggest that even in nongendered languages both children and adults attribute gender to nouns, and that these attributions develop further with age (Atagi *et al.*, 2009). Taken collectively, there appears to be some tendency to ascribe gender to language.

The impact of language on conceptualizations of gender further extends to cultural norms. Mazzuca *et al.* (2023) found distinct differences in how Dutch and Italian participants conceptualized gender. Their results suggested that Dutch participants presented more essentialist constructions of gender relative to their Italian counterparts (Mazzuca *et al.*, 2023). These cultural influences meditate and inform interactions with gender.

Culture

More traditional cultures may adopt more binary conceptualization of sex and gender that are institutionalized through policy. However, there is tremendous variation in cultural conceptualizations of gender around the world. To share just one example, South Asian cultures, including India, Bangladesh, and Nepal, recognize a third gender of Hijra. The Hijra are marked by their gender fluidity (Khan *et al.*, 2009). Within the Hijra, researchers have identified 15 subgroups differentiated by a variety of characteristics, including social identity, sexuality, genitalia, and proportion of embodiment of masculine-feminine characteristics (Khan *et al.*, 2009). The Hijra historically occupied socially recognized roles grounded in religious and spiritual practice, however, British colonial forces criminalized both cultural and Hijra-related practices during the 1800s, setting a precedent for discrimination and social exclusion (Ghosh, 2018; Gill-Peterson, 2024). This provides one example of how cultural understandings of gender vary and have the potential to inform gender identity and expression.

Intersectionality

As elaborated above, an individual's gender self-concept may be informed by factors including, but not limited to, culture, language, and age. The broader contexts that shape perceptions and interpretations of gender may also serve to marginalize individuals who exist outside of binary constraints of sex and gender. However, it is important to understand this marginalization through the lens of intersectionality. Intersectionality theory (Crenshaw, 1991; Crenshaw, 2017) describes how people embody multiple, complex social identities. These identities interact and intersect with each other, weaving together to form the material body of our experiences. Drawing on this theory, we understand that inequities can manifest across different dimensions of who we are. Dimensions of inequity are not purely additive; they are compounded and amplified. They feed into each other to shape experiences of marginalization, oppression, and domination (Crenshaw, 1991; Crenshaw, 2017).

Applying intersectionality theory to AI, it becomes clear that a cisgender, older, white, woman may potentially experience marginalization within the AI workforce as a result of underrepresentation. Here, the hypothetical woman may be subjected to sexism and ageism. However, we also understand that this experience will diverge from the experiences of a trans woman of colour within the same setting. A trans woman of colour will navigate sexism in addition to gender binarism and racism. Krieger (2020) outlines several systemic "isms" (including heterosexism, gender binarism, sexism, and racism) that structure inequities. "Isms" comprise insidious systems of oppression and domination; they inform the ways in which people interact with each other and are interacted with. These "isms" are not absent from AI spaces. While this chapter focuses on gender biases in AI (with specific allusions to sexism and gender binarism), gender biases are best investigated in the context of the other social-isms that structure society.

2.2 AI Conceptualizations of Sex and Gender: A Breeding Ground for Bias

Scheuerman *et al.* (2019) describes the process of integrating gender diversity into AI gender categorizations and classifications as a laborious task. To our understanding, AI does not yet demonstrate the capacity for organic thought. Thus, when discussing conceptualizations of sex and gender within AI, it is helpful to frame this as a reflection of two primary components: the developers behind AI, and the datasets used to train AI. That is to say, AI mirrors and amplifies the biases present within the developers of AI systems and the datasets used to train AI models.

2.2.1 A Brief Look at the Developers Behind AI

The 2023 Global Gender Gap Report found that only 30% of the AI workforce is comprised of women (Caira *et al.*, 2023). Another review from 2019 suggests that women comprise only 18% of C-Suite leaders in global AI enterprises (University, 2019). Outside of development spheres, AI research suggests similar trends. The OECD reports that only 25% of AI researchers are women (Caira *et al.*, 2023). Further, they note that men are the sole authors 55% of AI publications and that the figure decreases to 11% when looking at sole-publications written by women (Caira *et al.*, 2023). Taken collectively, these findings paint a compelling picture of male dominance and gender homogeneity in AI spaces, echoing similar trends in STEM. Despite advocacy for disaggregated data, there remains a noted absence of high-quality information about gender, race, and disability in AI (Young *et al.*, 2023; Zhang *et al.*, 2021). The absence of granular data further prevents us from assessing the contributions of trans and gender-diverse people to the development of AI systems. A noted exception is the work of the grassroots organization Queer in AI, which administers a yearly survey to document the demographics and barriers faced by sexual and gender minority researchers in AI. In their most recent survey from 2021-22, Ovalle *et al.* (2023) (N=225), 22% identified as transgender.

Keyes (2018) explains that "Researchers within Human-Computer Interaction have long studied the way that design processes dominated by men produce gendered, material differences in the usability of the resulting artifacts for women" (p. 88). By this logic, embedding diversity within developer teams might offset the gender discrepancies in artifact usability. Indeed, the presence of diverse perspectives is thought to enhance a group's problem-solving capabilities (Hong and Page, 2004). Boinodiris (2024) further emphasizes the merits of diversity in AI, arguing for the necessity of creating spaces for a range of multidisciplinary practitioners to pool their expertise and begin to develop responsible AI that reduces the potential for harm. In curating this environment, AI developers can also begin to navigate hard questions related to dataset availability and usage; this is central because bias in AI systems reflects historical and ongoing inadequacies in training data that perpetuate inequities (Bernstein and Turban, 2018; Boinodiris, 2024; Buolamwini, 2024; Chen, 2023).

2.2.2 The Training Data Problem

By using training data that defaults to dichotomized sex (male/female) and gender (man/woman) representations, AI systems are prone to sex and gender biases that amplify inequities for women, trans, non-binary, and gender-diverse people. Despite growing awareness about the ramifications of women's underrepresentation in AI spaces, Keyes (2018) identifies a gap in this area stemming from pervasive gender binarism. Specifically, Keyes (2018) argues that models of gender binarism "...fundamentally erase[s] transgender people, excluding their concerns, needs, and existences from both design and research" (p. 88). There is an apparent absence of high-quality training data that meaningfully attends to comprehensive dimensions of sex and gender. For example, Roberts and Fantz (2014) describe how electronic health records may not serve trans people by failing to accurately attend to, share, and store information related to sex and gender. Indeed, even when gender and sexual orientation data are collected, sex assigned at birth often trumps any other information, demonstrating what Albert and Delano (2022) refer to as sex obsession. When these electronic health records are later accessed and used as training data for AI systems, these gaps endure and create tangible harm. Cao and Daumé (2021) and Albert and Delano (2022) explain that failure to attend to gender complexity results in systems that, particularly for binary and non-binary trans people, deliver a worsened quality of service, reinforce stereotypes, and emphasize over/underrepresentation.

Lack of data (Buslón *et al.*, 2023) and incomplete or skewed data (Feast, 2020) are areas for strategic action in alleviating disparities related to sex and gender bias in healthcare-related AI. Buslón *et al.* (2023) propose several solutions to this challenge, including a) ensuring that data collection forms in healthcare attend to gender, sex, and sexual orientation, b) expansion of data collection sources and promotion of data integration, c) improving dataset quality and balance by designing unbiased data collection strategies, improving data representativeness and balance, and defining the parameters of data reutilization, d) inclusivity strategies, especially in relation to trans people, and e) intersectional approaches to data collection. These recommendations echo Albert and Delano (2022) who identify the following additions: a) education to raise awareness about sex and gender, and the experiences of trans people inside and outside medical contexts, b) working in diverse teams where people bring a range of skills and experience, c) being deliberate about whether sex/gender data are relevant to the research, d) documenting the use of sex/gender variables clearly, including how the data was collected, e) carry out data quality checks and identify strategies for managing missing data, f) consider additional sources of bias beyond sex and gender variables, and g) "audit model performance for subgroups without presuming or essentializing differences" (p. 7). These considerations may help to improve data availability moving forward, allowing for the development of training datasets that better attend to gender and sex-related diversity. However, current deficits in training data stand to increase misrecognition and misclassification rates (Katyal and Jung, 2021). The next section will provide three focused examples of bias and deficits in AI systems, highlighting how these deficiencies perpetuate inequities and cause harm.

2.3 Case Studies: Where Gender Binaries and Biases in AI Fail People

2.3.1 Automatic Gender Recognition and Other Facial Recognition Technology

Scheuerman *et al.* (2019) describe how facial analysis software applies binary gender categorizations to people, and image labelling software uses components of gender expression and presentation to ascribe gender. Both technologies consistently fail to accurately identify trans and non-binary individuals. They further describe how labelling in computer vision, understood primarily through gender performance and expression, reduces gender diversity to a male-female binary and directly contrasts with how users conceptualize gender. These scholars (Scheuerman *et al.*, 2021) have further developed their analysis to take into account the impact of colonial histories. Introducing the concept of auto-essentialization they argue that "the contemporary auto-essentialization of gender via the face is both racialized and trans-exclusive: it asserts a fixed gender binary and it elevates the white face as the ultimate model of gender difference" (p. 1).

Automatic Gender Recognition (AGR), a subfield of facial recognition technology, represents one algorithmic technology particularly informed by embedded gender representations. Keyes (2018) explored how gender is operationalized within AGR by performing a content analysis on papers on the topic. Their analyses revealed that of the 58 AGR papers in their sample, 94.8% of papers conceptualized gender as binary, 72.4% of papers conceptualized gender as immutable (gender as static), and 60.3% of papers conceptualized gender as physiological (gender as dependent on external presentation). Keyes (2018) argues AGR technology designed to see gender as something physiologically-rooted is more likely to misclassify and discriminate against trans, non-binary, and gender diverse people. Interviews conducted with trans people about AGR technology support this idea Hamidi *et al.* (2018). In these interviews, trans participants questioned whether AGR could accurately classify trans people, especially given that gender is subjectively experienced and expressed. Moreover, concerns about privacy and safety, especially in relation to incorrect gender attributions, also emerged (Hamidi *et al.*, 2018). These concerns are reinforced through multiple examples of surveillance and misrecognition in Katyal and Jung's analysis of the "gender panopticon" (Katyal and Jung, 2021). Misgendering has been associated with reduced felt authenticity, self-esteem, and increased feelings of stigmatization; higher feelings of stigmatization are also linked to increases in negative affect (McLemore, 2014). Limited gender representations informed by social, historical, and cultural biases are exclusionary to the real-world diversity present within humans; this exclusion stands to entrench experiences of marginalization including intersecting oppressions of racism, sexism, and binarism (Hamidi *et al.*, 2018).

2.3.2 Word Embeddings

In lay terms, word embedding is the process of creating a vector-based representation of a word (Bolukbasi *et al.*, 2016). Evidence suggests that word-embeddings are androcentric, biasing men over women (Petreski and Hashim, 2022). (Bolukbasi *et al.*, 2016) draw attention to the fact the many embeddings demonstrate gender bias and sexism, arguing that within word2vec (embedding system) trained on a dataset encompassing 3 million English words drawn from Google News, gender stereotypes were pervasive. For example, this embedding system returned X in the analogy "he to doctor is as she to X" as nurse (Bolukbasi *et al.*, 2016). This trend illustrates gender biases in the training dataset that perpetuate stereotypes regarding the roles and expectations of women. Other studies suggest similar problems arise using different embedding systems and training datasets. For example,Caliskan *et al.* (2022) describe how their embedding systems trained on internet-derived datasets identified different concepts associated with men and women. Concepts associated with women included references to roles in domains such as technology, engineering, sports, and religion. Conversely, concepts associated with women included female-specific slurs, sexual content, appearance, and kitchen items (Caliskan *et al.*, 2022).

While evidence highlights the social biases present and amplified by word embeddings, there lacks consensus on how to tackle this problem (Basta *et al.*, 2019; Bolukbasi *et al.*, 2016; Caliskan *et al.*, 2022). Bolukbasi *et al.* (2016), for instance, describe a debiasing strategy to remove gender associations from gender neutral words such as nurse. Persistent gender associations present even after debiasing, however, give rise to contextualized word embeddings strategies. These strategies involve creating word vector representations relative to the sentence a word is sourced from. Basta *et al.* (2019) suggest that the resulting gender conceptualizations arising from contextualized word embeddings are less biased than debiased word embedding strategies. While efforts to debias word embeddings stand to combat sexism within the industry, it is important to also recognize how the act of embedding itself enforces and upholds existing gender binaries.

2.3.3 It's not a Coincidence that All your Robots are Female, it's a Design Choice

It is easy to reduce computers and AI to purely technological artifacts aimed at bolstering human pursuits. However, doing so fails to acknowledge the nuances of how humans and computers interact. Human-Computer Interaction (HCI) is an interdisciplinary field focused on how machines balance considerations of functionality and usability to optimize service delivery (Karray *et al.*, 2008; Te'eni *et al.*, 2007). Within the field of HCI, Nass *et al.* (1994) introduced the Computers are Social Actors paradigm, which proposes that humans interact with computers socially and ascribe social norms onto computers. Using experimental results to ground this paradigm, Nass *et al.* (1994) further describe that humans perceive computers as gendered actors, in which gender-based social rules are applied to computers, and gender stereotypes were applied to computers using gendered characteristics such as vocal tone. The Computers are Social Actors paradigm presents a unique explanatory lens for the rapid emergence of female robots.

Borau *et al.* (2021) conducted several online studies highlighting how users perceive female robots as more human (humanness including the capacity for communicating warmth and emotional depth) than their male counterparts. This perception of humanity leads to more favorable user interactions with female robots, however, may stand to catalyze the objectification and dehumanization of real-world women. In this case, women are instrumentalized as a tool to imbue AI machines with humanness (Borau *et al.*, 2021). This hesitancy is amplified by Perugia and Lisy (2023), whose scoping review on gendering robots revealed that while gendered robots elicited gender stereotypes from users, gender did not yield significant effects of perceived robot competence, likeability, or acceptance. With this effect in mind, Perugia and Lisy (2023) argue that developers should continuously gather user feedback about design cues that engender stereotypes during multiple stages of the robot development process. In alignment with Nass *et al.* (1994), Tolmeijer *et al.* (2021) also found that stereotyping occurs when employing gendered voice assistants. Interestingly, Tolmeijer *et al.* (2021) emphasize the viability of gender-ambiguous voice assistants, finding no significant differences in user trust formation between gendered and non-gendered voice assistants. Taken collectively, these studies advocate for a nuanced approach when choosing to gender robots given the applicability (and further perpetuation) of gender biases and stereotypes.

The influx of female virtual assistants, including Apple's Siri, Microsoft's Cortana, and Amazon's Alexa, in technology spheres demonstrates another gendered development trend of interest. Loideain and Adams (2020) describe these assistants as "unbodied," using the term to highlight how their absence of form still evokes female designs via characterization, voice, and name. They argue that "As representations of an unbodied female, however, Alexa, Siri and Cortana also symbolise the ideal laborer of the service economy, whose labor is unseen" (p. 4). Ultimately, by this account, the servitude and staunch obedience of virtual assistants, combined with their invisible labor, mirrors the commodification of real-world women (Loideain and Adams, 2020). Designing virtual assistants who embody these gender norms stands to entrench and amplify existing inequities, including perspectives on gendered labor.

Pragmatists may argue that the inherent humanness of female robots may necessitate the integration of gender (Borau *et al.*, 2021). However, other considerations, including considerations of trust, gender bias, and user-acceptability, are also of relevance when developing AI-based systems. Being intentional (as opposed to relying on pejorative stereotypes) about the choice to gender robots or not requires developers to adequately consider dimensions of sexism and gender bias. Shea (2023) points to the example of Ai-Da, an artistic humanoid robot, as a case of intentional gendering in AI. Shea highlights how Ai-Da's persona and appearance is inspired by computer programmer Ada Lovelace. Quoting Zevi, the head of operation for the Ai-Da project, Shea describes how the explicit gendering of Ai-Da stands to give voice to the women who are often underrepresented in technology and art sectors (Shea, 2023). Ai-Da presents just one encouraging example of robot gendering done with caution and intention; a trend we hope to see continue as developers reflect on how best to model human-computer interactions.

2.4 Debiasing AI Moving Forward

2.4.1 Human Centered Design

Advocates and scholars have continued to champion AI design philosophies centered around human experiences; this movement is encompassed by the "human-centered" design umbrella that fundamentally seeks to develop safe, reliable, and trustworthy AI systems (Shneiderman, 2020). Despite the supposed human-focus of human-centered AI, there continue to be gaps in practice. For example, Bingley *et al.* (2023) describe how AI-users prioritize the social impact of AI more than developers; they argue that to better embody the human-centered approach, developers ought to focus on the needs and values of users. Moreover, several guidelines have been established by academic, industry, and government stakeholders to center considerations of human-centered AI in the development of AI systems, however, these guidelines often lack concrete mechanisms to bridge the principle-practice divide (Tannenbaum *et al.*, 2019).

2.4.2 Data Transparency Measures

The data that is fed into an AI system influences how that system behaves. Moving towards using datasets that take an inclusive, rather than a reductive approach to gender is beneficial. Indeed, some work seeks to achieve this. For example, Krieg *et al.* (2023) introduced the dataset Gender Representation-Bias for Information Retrieval (Grep-BiasIR) to aid in alleviating gender bias in information retrieval. However, there are areas in which gold standard comprehensive data or debiasing strategies are unavailable. In these cases, it is in the interest of socially responsible developers to document the creation and intended uses of a dataset (Gebru *et al.*, 2021). By approaching the dataset developmental endeavor with a principle of transparency, those who later wish to use datasets for their own purposes are better positioned to make informed decisions about computing model data sources, increasing developer accountability. There are several ways of doing this, but for the scope of this chapter, we focus on datasheets and nutrition labels.

Gebru *et al.* (2021) propose using datasheets to record the dataset development process, and provide several prompts aimed at allowing AI practitioners to better understand several important things, including the motivation behind the dataset, elemental constituents of the dataset, data collection process, and recommended uses. These datasheets are intended to provide detailed, descriptive accounts about the dataset from the perspectives of dataset developers. The proposed questions posed to dataset developers were continuously iterated upon in response to preliminary feedback from diverse stakeholders, including researchers, developers, and policy makers. Gebru *et al.* (2021) emphasize that these sheets are intended to be completed carefully and reflectively.

Elsewhere, the Data Nutrition Project offers a different way to document dataset characteristics. Chmielinski *et al.* (2022) propose a unique digital interface consisting of three panels that present the following information: Overview, Use Cases and Alerts, and Dataset Information. This constitutes a dataset nutrition label (second generation). A data nutrition label aims at increasing data transparency and combatting problematic datasets that replicate biases and harm. The second generation of the data nutrition label arose in response to interviews with data practitioners, who emphasized the value of "intended use", as a panel of particular interest. These interviews suggested that "intended use" include information related to a dataset's applicability to a particular context, and therefore, was a source of considerable interest (Chmielinski *et al.*, 2022).

Both the datasheets and nutrition labels provide ways for dataset developers to make known the intended parameters, considerations, and limitations that underpin their datasets. Moving towards this practice encourages both data and AI practitioners to make informed decisions about the data they use, fostering a greater sense of accountability in turn.

2.4.3 Designing for Fairness

In addition to creating more inclusive datasets and labeling existing datasets to make visible the intended uses, strengths, and limitations of data, developers may seek to mitigate biases by design algorithms informed by fairness principles. Broadly, fairness in AI refers to the act of active bias mitigation and/or the dissolution of unjust processes to alleviate systematically discriminatory practices (Li *et al.*, 2022; Memarian and Doleck, 2023; Shin *et al.*, 2021). The overall goal of fairness strategies is to ensure group fairness (statistical parity) and individual fairness (similar individuals are classified similarly) (Zemel *et al.*, 2013). Fairness strategies can be integrated at the pre-processing, in-processing, or post-processing stages.

Pre-processing fairness techniques involve addressing biases in training data via several strategies, including, but not limited to, data resampling, augmentation, and reweighting (Chen, 2023). According to Chen, resampling involves equalizing group proportional representations, reweighting increases the significance of underrepresented incidents, while data augmentation involves generating synthetic data to improve data diversity. Kamiran and Calders (2012)emphasize the use of preferential sampling, as opposed to uniform sampling, as effective in reducing discrimination without sacrificing accuracy.

In-processing fairness techniques introduce fairness constraints during the model training phases of algorithm development (Chen *et al.*, 2023b). One common method includes adversarial training and debiasing, which reduces the influence of protected attributes on a model's predictive capabilities (Chen *et al.*, 2023b). Zhao *et al.* (2018a) use adversarial debiasing to develop a predictive model that exhibits both accuracy and reduced attribute stereotypes. Zemel *et al.* (2013) adopt a different approach, presenting a high-accuracy model that reduces discrimination by mapping individuals to a probability distribution. Both approaches highlight how in-processing techniques may help mitigate bias and discriminatory tendencies within AI systems.

Finally, post-processing fairness techniques adjust algorithmic outputs post-model training (Chen, 2023). Pleiss *et al.* (2017) describe calibration, a process that seeks to maintain accuracy in predicted probabilities to avoid bias. However, Pleiss *et al.* (2017) describe specific cases for which calibration enables non-discrimination and cases where it doesn't, ultimately asserting that calibration is not compatible with low-error solutions. Another example of post-processing fairness is reject options classification, which allows for the rejection of biased predictions (Chen *et al.*, 2023b). Lohia *et al.* (2019) explain how their post-processing algorithm engages this technique to improve group fairness, highlighting how their developed bias detector enables for the prioritization of certain data samples to reduce bias.

Taken collectively, the above techniques begin to form the tapestry of technical tools used to mitigate bias in the development of AI systems and ensure fairness in their development.

2.4.4 A Call for a Comprehensive Approach

UNESCO's IRCAI report on systemic biases in large language models presents several recommendations for bolstering gender equity in AI (UNESCO, 2024). Among these recommendations is a call for a comprehensive approach to tackling gender-based inequities that integrates both the direct and social-origins of biases. This integrated approach acknowledges how broader socials norms and gendered rhetoric are immortalized in AI systems, replicating inequities in turn. To engender equity, we must reflect on the social consequences of gender biases, but also acknowledge how data collection and model development efforts must be tailored to address existing disparities. Specifically, UNESCO IRCAI calls for developers to approach the development of AI from an ethical lens from the outset, advocating for pre- and post-market bias audits and evaluations. As part of this ethical approach, developers should perform threat modelling evaluating the potential impact an AI system will have on vulnerable populations. As part of this ethical development strategy, UNESCO IRCAI further emphasizes the necessity of embedding diversity within the development team (UNESCO, 2024).

UNESCO IRCAI provides further recommendations for policy-level intervention in the AI space. Specifically, UNESCO's IRCAI urges policy makers to promote or mandate training dataset transparency. By elucidating data deficits or other forms of under-representation, AI practitioners can make informed decisions about data usage. Moreover, they call for policy makers to manage and continuously verify that AI practitioners are working within the confines of equitable performance via regular human impact assessments and the creation of localized benchmark datasets. This call reflects the need to coordinate policy makers and AI practitioners to foster an environment conducive to the creation of equitable AI that is responsive to real-world inequities.

2.4.5 Critical Next Steps

Beyond these strategies, more critical approaches have been identified by those who have the most to lose from bias in AI, including racialized and non-racialized women, queer, trans, and non-binary people, and people with intersecting marginalized identities. This work draws attention to current harms and anticipated risks associated with broader deployment of biased AI systems linked to identification and surveillance (Katyal and Jung, 2021). When challenges linked to AI impacting marginalized groups collide with government laws targeting these populations, there are important human rights implications that cannot be ignored (Buolamwini, 2024; Castets-Renard and Lequesne, 2023; Dellinger and Pell, 2024; Katyal and Jung, 2021; Ovalle *et al.*, 2023). A number of grassroots organizations including Queer in AI, the Algorithmic Justice League, the Digital Defense Fund, the Electronic Frontier Foundation, and the Surveillance Technology Oversight Project are at the forefront of advocating for change (Buolamwini, 2024; Mort, 2023; Ovalle *et al.*, 2023; Surveillance Technology Oversight Project, 2022). Key to this approach is the need for more decentralized participatory strategies that directly engage people who have been (or will be) directly impacted by gender bias in AI (Deng *et al.*, 2023; Lam *et al.*, 2022; Ovalle *et al.*, 2023). The current political climate, past experience, and potential for serious harm make it clear that the time for action is now.

3

Gendered Information Retrieval Systems: Metrics and Measurements

Chapter 2 examined the constructs of sex and gender, providing insight into how these categories are framed and interpreted in AI-driven systems. Building on this theoretical basis, we now explore how these biases emerge within IR systems, focusing on the metrics, evaluation frameworks, and methodologies developed to detect and address gender bias in search outputs. This chapter bridges theory with practice, setting the stage for an in-depth analysis of gendered interactions within IR systems.

This chapter is organized to provide an exploration of gender in information retrieval (IR), beginning with basic concepts and progressing to practical methodologies for analysis and mitigation. We first introduce fairness in IR systems, focusing on how gender intersects with established principles of group *fairness and individual fairness*. These concepts provide a theoretical basis for understanding the role of gender in IR and establish a framework for addressing bias. Following this, we examine how gender is reflected in queries and documents, exploring perspectives such as lexical (the language used), subject-based (the focus of the content), and authorial (the creator's identity).

Building on this foundation, the chapter transitions to empirical approaches for evaluating and addressing gender bias. We present metrics, datasets, and benchmarks from the literature designed to assess gender bias in queries and retrieved documents, providing actionable methods to analyze and improve fairness in IR systems. The chapter is structured to ensure a logical progression: from defining fairness concepts and identifying gender representations to applying practical tools for bias mitigation. This structure creates connections between theoretical principles and practical evaluation mechanisms, offering a possible useful framework for understanding and addressing gender bias in IR. By linking these components, the chapter aims to provide a structured pathway for exploring gender fairness in information retrieval.

3.1 Gender Fairness in Ranking

Fairness in ranking can be defined in various ways, depending on the perspective from which the problem is analyzed (Zehlike *et al.*, 2022; Diaz *et al.*, 2020; Biega *et al.*, 2018). In this manuscript, we focus on fairness in terms of gender equality. There are two broad definitions for fairness, namely *Group Fairness*, and *Individual Fairness*. In this section, we define gender fairness in information retrieval as a subset of *group* and *individual* fairness.

Definition 3.1. Group Fairness: Group fairness aims to ensure that groups of individuals with different protected sensitive attributes receive comparable treatments statistically.

One of the most common definitions of group fairness is **statistical parity** (Dwork *et al.*, 2012). Statistical parity requires the prediction y to be independent of the sensitive attribute s, denoted as $y \perp s$. This can be mathematically represented for binary classification and binary attributes as follows:

$$P(y = 1 \mid s = 0) = P(y = 1 \mid s = 1) \tag{3.1}$$

To measure statistical parity, we can use the difference in probabilities:

$$\Delta_{SP} = |P(y = 1 \mid s = 0) - P(y = 1 \mid s = 1)| \tag{3.2}$$

A lower value of Δ_{SP} indicates a fairer classifier. Statistical parity can be extended to multi-class and multi-category sensitive attributes by ensuring that the prediction y is independent of s.

One approach to conceptualize group fairness in IR argues that fairness may be achieved when the probability of a document being retrieved for a query is independent of its gender attribute, particularly for gender-neutral queries. This interpretation is rooted in the idea that documents with identical content but differing gender attributes should have equal chances of being retrieved, as their relevance to the query remains unaffected by gender. For instance, given a gender-neutral query q_n, and two documents d_m and d_f with identical content but associated with male and female attributes respectively, the probability of retrieving either document should not differ significantly.

This perspective aligns with the principle of *statistical parity*, which can be mathematically expressed to measure the degree of fairness in the retrieval process. Statistical parity ensures that the likelihood of retrieval is balanced across gender attributes, reducing disparities introduced by implicit or explicit biases:

$$\Delta_{SP} = |P(d_g, q_n) \mid g = m) - P(d_g, q_n) \mid g = f)| \tag{3.3}$$

where, $P(d_g, q_n)$ is the probability of the document d_g being the top-retrieved document for the gender-neutral query q_n. Therefore, group fairness in information retrieval is addressed if and only if:

$$\Delta_{SP} \to 0 \tag{3.4}$$

Definition 3.2. Individual Fairness: Individual fairness focuses on ensuring that similar individuals receive similar algorithmic outcomes.

Individual fairness for gender focuses on ensuring that algorithmic outcomes treat similar entities equitably, regardless of their gender-related attributes. In information retrieval, this means that documents or queries with similar intrinsic relevance or difficulty should receive similar treatment by the system, independent of their association with a particular gender. This approach addresses disparities that may arise from systemic biases, ensuring that gender does not influence the outcomes for otherwise comparable items. By grounding fairness in unbiased metrics of similarity or merit, individual fairness for gender promotes equitable treatment at the level of individual entities while avoiding overgeneralization based on group characteristics.

An example of individual fairness is the amortized attention framework (Biega *et al.*, 2018) that ensures that items with similar relevance receive equitable exposure over a series of rankings, making it a valuable approach for addressing gender bias in information retrieval. This framework can help mitigate imbalances in attention, such as clicks or visibility, between gender-associated documents or content. For example, in a scenario where equally relevant documents feature female-associated content (e.g., biographies of women) and male-associated content (e.g., biographies of men), systemic biases in ranking algorithms might lead to disproportionately higher exposure for male-associated content. By applying the amortized attention framework, the cumulative exposure of female- and male-associated documents can be aligned with their relevance across multiple rankings, ensuring balanced representation and preventing the consistent underrepresentation of any gender.

Another example of individual fairness frameworks is the Expected Exposure Model by Diaz *et al.* (2020), which evaluates fairness in rankings by examining how attention (exposure) is distributed across items of the same relevance grade, making it particularly effective for addressing gender bias in information retrieval. The model ensures that documents or items associated with different characteristics (in this case gender) but of equal relevance receive comparable exposure across stochastic rankings. For instance, in response to a query like ‘top scientists in history,’ documents about male scientists may consistently rank higher than those about female scientists due to systemic biases, resulting in unequal exposure. The expected exposure model addresses this by aligning exposure with relevance, ensuring that female-associated content receives exposure comparable to equally relevant male-associated content, thereby fostering fairness in gender representation.

Given the definitions of group fairness, and individual fairness in information retrieval, a fair information retrieval system satisfies both group and individual fairness to ensure that the model has fair behaviour towards different demographic groups, and also the individuals with different genders.

With the fairness definitions in mind, we are going to take a deeper look into the neural rankers, and the existing gender biases in them. Rekabsaz et al. are among the first researchers who discovered that neural rankers not only exhibit gender biases but also reinforce the existing biases (Rekabsaz and Schedl, 2020). During their extensive experiments, they revealed that neural rankers exhibit gender biases in the sense that for a gender-neutral query, most of the retrieved documents exhibit inclination toward males. This contradicts group fairness defined earlier.

On the other hand, Seyedsalehi *et al.* (2022b) revealed another category of gender biases in information retrieval systems. Their research shows that the neural rankers perform better when applied to male queries, as compared to the female queries. This contradicts individual fairness that states every individual should be treated the same by an information retrieval system, regardless of their gender attribute. This research led to the following research question in terms of the origin of these gender biases as well as different datasets, and metrics for measuring gender bias in information retrieval systems.

3.2 Evaluating Gender Fairness

The initial step in mitigating bias involves quantifying and measuring gender biases. Therefore, we first explore existing methods for assessing gender bias and evaluating the fairness of information retrieval models.

Various metrics and frameworks are proposed for measuring gender bias within Natural Language Processing (NLP) and IR (Basta *et al.*, 2019; Chaloner and Maldonado, 2019; Dev *et al.*, 2020). In this section we review the evaluation frameworks for gender bias in 1) word embeddings as a fundamental component of many IR systems and 2) information retrieval systems.

3.2.1 Gender Bias in Large Language Models

The increasing presence of gender bias in machine learning models and information retrieval (IR) systems has raised concerns about equitable representation across various gender identities. Traditional IR systems often reflect societal biases by disproportionately favoring certain identities, perpetuating stereotypes, or even marginalizing non-binary individuals. There are very limited works, that address gender bias for non-binary communities (Felkner *et al.*, 2023). To address these issues, GenderCARE (Tang *et al.*, 2024) metrics offer a framework for quantifying and mitigating gender bias, aiming to create fairer and more inclusive IR systems. This section describes the key metrics used to assess gender bias for both binary and non-binary identities and explores their application in information retrieval.

1. Bias-pair ratio (BPR) is a lexical metric that quantifies the extent of bias in generated content based on the frequency of biased descriptors. It measures how often a model uses biased terms when responding to gender-targeted prompts. For IR systems, BPR helps to identify and reduce the likelihood that the system will disproportionately represent certain genders with stereotyped or biased descriptors.

The formula for BPR is given by:

$$BPR = \frac{N_{biased}}{N_{total}}$$

where N_{biased} represents the number of biased descriptors selected by the model, and N_{total} is the total number of descriptors (both biased and anti-biased) used in the responses.

This ratio, ranging from 0 to 1, reflects the model's propensity to use biased language, with values closer to 1 indicating higher bias. By monitoring BPR across binary (male/female) and non-binary groups, IR systems can adjust descriptor selection to achieve a more balanced portrayal.

BPR can be employed to audit IR systems to ensure that search results, especially those related to professions or identity-sensitive topics, are not skewed by gendered language. For example, a search for "successful leaders" should return content that represents male, female, and non-binary leaders without biased language.

2. Toxicity Metric is a sentiment-based metric that measures the level of harmful or offensive language in a model's output. Specifically, it assesses whether responses from different gender groups contain language that could perpetuate harmful stereotypes or negative sentiments.

Toxicity is especially relevant when examining non-binary representations, as marginalized identities are more likely to experience biased or derogatory language. The Toxicity score ranges from 0 to 1, with higher values indicating greater levels of harmful content.

Measuring toxicity is useful in IR systems that return user-facing content, such as personalized news or educational material. For queries related to non-binary identities, the IR system should minimize toxicity to prevent the amplification of harmful stereotypes. An IR system could, for instance, use this metric to refine responses to queries like "non-binary identity in the workplace" to ensure that content avoids reinforcing negative views.

3. Regard Metric evaluates the sentiment—positive, neutral, or negative—expressed in a model's output toward specific gender groups. For both binary and non-binary identities, Regard is crucial in assessing whether the IR system is consistently representing all gender identities in a balanced and respectful manner. Each sentiment category (positive, negative, or neutral) is scored from 0 to 1, with higher scores indicating stronger associations within that category.

The Regard metric is particularly valuable when comparing sentiment differences across gender groups to uncover any bias in the tone or sentiment conveyed by the IR system. For example, a model that associates positive descriptors primarily with male identities but negative descriptors with non-binary identities would exhibit sentiment bias.

Regard helps IR systems ensure that search results related to gender-sensitive topics portray all gender identities positively or neutrally, depending on context. For instance, in content recommendations about careers, Regard can help verify that non-binary and binary genders are equally associated with positive sentiments like "intelligent" or "accomplished."

4. Pair-Based Evaluation for Non-Binary Identities

The GenderCARE framework offers a specialized approach for evaluating bias specific to non-binary identities. The benchmark includes Pair Sets composed of descriptors and anti-biased descriptors for binary (male/female) and non-binary gender groups. Each Pair Set is structured as:

(Gender Target, Biased Descriptor, Anti-Biased Descriptor)

The benchmark evaluates how often the IR system selects biased versus anti-biased descriptors for each gender group, including non-binary identities. This pair-based construction captures nuanced biases that may not be visible through single descriptors alone and allows the IR system to adjust based on these insights.

This method is essential for auditing datasets used in training IR models to ensure the inclusivity of non-binary descriptors. For instance, a media-focused IR system could use this benchmark to verify that non-binary identities are not only represented but also associated with positive descriptors (e.g., “resilient” instead of “confused”).

By employing these metrics, information retrieval systems can assess and mitigate biases, leading to more inclusive content delivery. The GenderCARE metrics can be applied in IR across several scenarios:

1. **Fair Representation in Search Results**: These metrics can be used to audit search results for fair gender representation, ensuring that results for gender-related queries represent male, female, and non-binary identities equitably.

2. **Bias Detection in Content Recommendations**: IR systems that provide personalized recommendations, such as video or article suggestions, can use GenderCARE metrics to ensure recommendations do not disproportionately favor or misrepresent any gender. For instance, in a news recommendation system, the metrics help prevent the reinforcement of stereotypes, ensuring that non-binary users see content that aligns with a balanced gender representation.

3. **Enhanced User Sentiment Analysis**: GenderCARE’s sentiment-based metrics (Toxicity and Regard) enable IR systems to offer context-sensitive responses by evaluating sentiment and reducing the likelihood of inadvertently harmful content, particularly for sensitive queries related to identity or support resources.

4. **Evaluating Fairness in Query-Driven Systems**: GenderCARE metrics help IR systems respond to open-ended or question-answering queries (e.g., “What challenges do non-binary individuals

face in STEM?") without reinforcing stereotypes, instead presenting balanced, positive language across gender groups.

5. **Dataset Curation and Audit**: GenderCARE metrics are valuable in dataset curation for training IR systems. By using these metrics, curators can ensure that datasets used for training and fine-tuning IR systems fairly represent all gender identities, fostering equitable search and retrieval outcomes.

3.2.2 Gender Bias in Word Embeddings

The pervasive issue of gender bias in word embeddings has significant implications for IR systems. This section provides a comprehensive examination of gender bias within word embeddings. We begin by exploring two different approaches to understanding and measuring biases in word embeddings: first, through the geometric properties that reveal the presence of gender bias, and second, through natural language inference.

Geometry of Gender and Bias in Word Embeddings

Gender bias in word embeddings is a critical issue that arises from the way these embeddings are constructed. This bias is not only a reflection of societal norms but also a potential amplifier of these biases when utilized in machine learning and natural language processing tasks. To address and mitigate these biases, it is essential to understand the geometric properties of the embeddings that capture gender bias. This section delves into the geometry of gender and bias, identifying the gender subspace in word embeddings and introducing metrics to quantify the biases (Bolukbasi *et al.*, 2016).

Identifying the Gender Subspace. Word embeddings represent words as vectors in a high-dimensional space. To identify gender bias within this space, we first need to locate the gender subspace—a direction or set of directions that captures the gender differences in the embeddings. This subspace can be found by examining the differences between pairs of gendered words. The procedure can be explained as follows:

1. Pairs of gendered words, such as ‘he’ and ‘she’ or ‘man’ and ‘woman,’ are selected. Each pair provides a difference vector, which is hypothesized to represent the gender component in the embedding space.

2. To robustly estimate the gender direction, differences of multiple gender pairs are aggregated and PCA is applied. The first principal component (PC) captures the direction that explains the most variance in these differences, which is interpreted as the *gender subspace*.

3. This subspace is validated by comparing it against human-judged lists of gendered words. For instance, crowd workers provide words associated with males or females, and the alignment of the identified gender direction with these associations is checked.

Based on this gender subspace, two different types of bias can be defined.

- **Direct bias** refers to the association between gender-neutral words and the gender subspace. For example, if the word ‘nurse’ is closer to the gender vector associated with ‘female’ than ‘male’, this indicates direct bias. The direct bias of an embedding for a set of gender-neutral words N and the gender direction g is defined as:

$$\text{DirectBias}_c = \frac{1}{|N|} \sum_{w \in N} |\cos(\vec{w}, g)|^c \tag{3.5}$$

  Here, N is the set of gender-neutral words, w is the vector representation of the words, g is the gender direction learned from above, and c is a parameter that determines the strictness of the bias measurement. For $c = 0$, any deviation from the gender direction counts as bias. For $c = 1$, bias is measured more gradually.

- **Indirect bias** manifests in the relationships between gender-neutral words that arise due to their association with the gender subspace. For instance, if ‘receptionist’ is closer to ‘softball’ (a word associated with females) than ‘football’ (a word associated with males), this indicates indirect bias. Each word vector w is decomposed into its gender component w_g and the remainder $w_\perp$:

$$w = w_g + w_\perp \tag{3.6}$$

The gender component w_g is the projection of the word vector w onto the gender subspace g. The gender component of the similarity between two word vectors w and v is then defined as:

$$\beta(w, v) = \frac{(w \cdot v - w_\perp \cdot v_\perp)}{w \cdot v} \tag{3.7}$$

where $w_\perp$ and $v_\perp$ are the components of w and v orthogonal to the gender subspace g.

This metric quantifies how much the similarity between two words changes when the gender subspace is removed. A high β value indicates significant indirect bias.

In summary, understanding and addressing gender bias in word embeddings is crucial for creating fair and unbiased machine learning systems. By identifying the gender subspace and applying debiasing algorithms, these biases can be mitigated, moving towards more equitable AI applications.

Detecting Biases in Word Embeddings Using Natural Language Inference

The gender bias measurement introduced in (Dev *et al.*, 2020) relies on the task of natural language inference (NLI) to detect biases within word embeddings. The principle behind this approach is that biased word embeddings can lead to invalid inferences when used in downstream NLP models.

Natural Language Inference (NLI) involves determining whether a given hypothesis sentence can be logically inferred from a given premise sentence. The goal is to categorize the relationship between these sentences as entailment, contradiction, or neutral. The researchers use this task to uncover biases by constructing sentence pairs that should ideally result in a neutral relationship but, due to biased embeddings, often do not.

To illustrate, consider the following pairs of sentences:

- **Premise**: The accountant ate a bagel.
- **Hypothesis 1**: The man ate a bagel.
- **Hypothesis 2**: The woman ate a bagel.

In an unbiased system, the premise should neither entail nor contradict either hypothesis since the gender of the accountant is not specified. However, biased embeddings might yield higher probabilities of entailment or contradiction based on gendered assumptions.

Mathemat ically, bias is quantified by calculating model probabilities for entailment e_i, neutral n_i, and contradiction c_i for each sentence pair s_i. Three aggregate measures are defined:

1. Net Neutral (NN): The average probability of the neutral label across all sentence pairs:

$$\mathrm{NN} = \frac{1}{M}\sum_{i=1}^{M} n_i \tag{3.8}$$

2. Fraction Neutral (FN): The fraction of sentence pairs labeled as neutral:

$$\mathrm{FN} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}[n_i = \max\{e_i, n_i, c_i\}] \tag{3.9}$$

where $\mathbf{1}[\cdot]$ is an indicator function.

3. Threshold $\tau(T : \tau)$: The fraction of examples with a neutral probability above a threshold τ:

$$\mathrm{T}{:}\tau = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}[n_i > \tau] \tag{3.10}$$

These measures are intended to be close to 1 in an ideal bias-free scenario. By generating a large number of sentence pairs targeting specific biases, the degree of bias in word embeddings can be effectively measured.

For instance, when evaluating gender bias, occupations are paired with gendered words to assess neutrality. Templates are used to generate sentences systematically, ensuring comprehensive coverage of potential biases. This methodology reveals substantial gender biases across different embeddings, such as GloVe(Pennington *et al.*, 2014), ELMo(Peters *et al.*, 2018b), and BERT(Devlin *et al.*, 2018a).

By identifying the extent of bias through invalid inferences, this approach provides a robust mechanism to measure and subsequently mitigate biases in word embeddings, thereby improving the fairness and reliability of NLP models.

3.2.3 Gender Bias in Recommendation Systems

Recommendation systems operate within a multi-stakeholder environment, often needing to balance the interests of content providers and consumers. As recommendation systems have become pervasive, their potential for unintended harms, such as feedback loops, gender stereotyping, and racial bias, has grown. To address these issues, a plethora of fairness definitions and metrics have emerged (Chen *et al.*, 2021; Chen *et al.*, 2023a; Saxena and Jain, 2024), but this diversity has led to confusion for practitioners attempting to implement these metrics effectively. Smith and Beattie (2022) discusses how the complex structure of recommendation systems, which often include multiple sequential components (e.g., content generation, retrieval, and ranking), complicates fairness measurement. Fairness must often be assessed within each component, and conflicting fairness demands can arise between different metrics or stakeholders. Practitioners must consider parameters such as proxy variables and fairness thresholds, which further complicate metric selection. For instance, one metric may prioritize fairness within groups while another emphasizes fairness across groups, requiring practitioners to decide which aligns better with their system's objectives and context.

Beattie *et al.* (2022) explore the gap between ethical guidelines in AI research and practical implementation within industry recommendation systems, highlighting Spotify's experience. Their work emphasizes the complexity of defining, measuring, and mitigating algorithmic bias in recommendations, which involve multi-stakeholder considerations and lack standardized tools. The authors outline a four-step framework: defining the scope of fairness evaluations, identifying appropriate metrics, implementing these metrics, and establishing ongoing monitoring to flag biases. Each step presents unique challenges, such as selecting fair metrics compatible with recommendation algorithms, translating research methods into scalable code, and setting actionable thresholds for bias intervention. The authors argue that meaningful progress requires more concrete, standardized guidance, as well as collaboration between industry and academia to develop robust, adaptable auditing frameworks for AI ethics in recommendation systems.

Quantifying Gender Bias in Recommendation Systems

Gender bias can be regarded as a subset of popularity bias, where the imbalance in recommendations or exposures extends to favor certain gender-associated items or content. Just as popularity bias in recommender systems amplifies already popular items, gender bias results in an unequal representation of items associated with particular gender groups. This gender-skewed exposure is a form of popularity bias because it tends to favor items popular within specific gender demographics, thereby limiting fair exposure across genders. By adapting existing popularity bias metrics (Abdollahpouri *et al.*, 2021), we can measure and address gender bias in recommendation systems more precisely.

Below are the modified metrics to capture this aspect of gender bias within recommendation systems:

1. Average Recommendation Popularity (ARP) (Yin *et al.*, 2012): Traditionally, ARP calculates the average popularity of items in recommendations by measuring the average number of interactions with these items. For gender bias, we redefine ARP to measure the average gendered popularity of recommended items:

$$\mathrm{ARP}_{\text{gender}} = \frac{1}{|U|} \sum_{u \in U} \frac{\sum_{i \in L_u} \phi_{\text{gender}}(i)}{|L_u|}$$

where $\phi_{\text{gender}}(i)$ denotes the number of interactions or ratings from users associated with a specific gender for item i. This metric highlights whether items popular among certain gender groups are disproportionately recommended, indicating potential gender bias in exposure.

2. Aggregate Diversity (Agg-Div) (Adomavicius and Kwon, 2011): Originally, this metric measures the breadth of unique items recommended across all users to promote variety. To track gender bias, we redefine it to assess the diversity of items across different gender associations in the recommendation pool:

$$\text{Agg-Div}_{\text{gender}} = \frac{\left|\bigcup_{u \in U} L_u^{\text{gender}}\right|}{|I|}$$

where $|I|$ represents the total number of unique items in the entire catalog, and L_u^{gender} includes items in user u's list associated with different genders, thereby assessing if items linked to each gender are equitably represented in recommendations.

3. Gini Index (Eskandanian and Mobasher, 2020): The Gini Index traditionally quantifies inequality in item recommendation frequencies, where a higher Gini Index reflects greater disparity. For gender bias, we modify the Gini Index to measure imbalance in gender-based item recommendations across users:

$$\text{Gini}_{\text{gender}} = 1 - \frac{1}{|I| - 1} \sum_{k=1}^{|I|} (2k - |I| - 1)\, p(i_k | L, \text{gender})$$

where $|I|$ represents the total number of unique items in the entire catalog, and $p(i_k|L, \text{gender})$ is the probability of item i_k being recommended within a specific gender demographic. A lower value suggests a fairer distribution of gender-related items, ensuring no single gender's associated content is overemphasized.

3.2.4 Gender Bias in Information Retrieval Systems

Quantifying Bias in Search Engine Rankings

Gender stereotypes in word embeddings (WEs) are captured by calculating the genderedness of words (Fabris *et al.*, 2020). As explained in the previous section, this process begins with identifying a gender subspace using principal component analysis (PCA) on intrinsically gendered word pairs (e.g., she-he, woman-man). The first principal component, which captures the most variance, is considered the gender direction (wg). The genderedness score of a word w is then computed as its projection onto the gender direction using the formula

$$g(w) = \frac{w \cdot wg}{|w||wg|} \tag{3.11}$$

This score indicates the association of the word with male or female stereotypes, where positive values are associated with female stereotypes and negative values with male stereotypes.

To model stereotypes in query-document pairs, the genderedness scores for both queries and documents need to be computed. For a query q, the genderedness score is determined by averaging the genderedness scores of its terms after removing stop words:

$$g(q) = \frac{1}{|q|} \sum_{w \in q} g(w) \tag{3.12}$$

Similarly, for a document d retrieved for query q, the genderedness score is calculated by averaging the genderedness scores of its terms, excluding stop words and terms present in the query:

$$g_q(d) = \frac{1}{|d \setminus q|} \sum_{w \in d \setminus q} g(w) \tag{3.13}$$

Where, $d \setminus q$ is the set of words appearing in the documents, excluding the query terms. The genderedness of a ranked list L of documents retrieved for a query q is computed as a weighted average of the genderedness scores of the documents, with weights decreasing logarithmically by rank. The weight for each document d_k at rank k is calculated using:

$$w_k = \frac{1}{\log_2(k+1)} \tag{3.14}$$

The genderedness of the ranked list L is then determined by

$$g_q(L) = \frac{1}{W} \sum_{k=1}^{K} w_k \cdot g_q(d_k) \tag{3.15}$$

where W is the normalization factor

$$W = \sum_{k=1}^{K} w_k \tag{3.16}$$

To measure the tendency of a search engine (SE) to reinforce gender stereotypes across a set of queries, Gender Stereotype Reinforcement (GSR) is introduced, which involves calculating the correlation between the genderedness of queries and the genderedness of the corresponding ranked lists. For a set of queries Q and their corresponding ranked lists L, a linear fit is performed between the genderedness of queries $g(q)$ and the genderedness of ranked lists $g_q(L)$. The GSR is defined as the slope of the linear fit:

$$ms(Q, D) = \frac{1}{\sigma^2_{g(q)}} \cdot \frac{1}{N} \sum_{i=1}^{N} (g(q_i) - \mu_q)(g_{q_i}(L_i) - \mu_{q,L}) \tag{3.17}$$

where $\sigma^2_{g(q)}$ is the variance of the genderedness of the queries, μ_q is the mean genderedness of the queries, and $\mu_{q,L}$ is the mean genderedness of the ranked lists.

To summarize, the GSR measure involves several key steps. First, the genderedness score for each word in the query and documents is calculated. Next, the average genderedness for queries and documents is computed. Then, the weighted average genderedness for ranked lists is determined. Following this, a linear fit between the genderedness of queries and ranked lists is performed. Finally, the GSR is calculated as the slope of the linear fit. The GSR measure provides a quantitative assessment of how SEs reinforce gender stereotypes through their ranked results. By leveraging word embeddings to detect gender bias, the measure offers a way to audit and improve the fairness of SEs, ultimately aiming to reduce the perpetuation of harmful gender stereotypes in information retrieval.

Retrieval Bias Measurement Framework

In the following, we provide a detailed explanation of a framework designed to measure gender bias in information retrieval (IR) systems introduced by Rekabsaz and Schedl (2020). This framework evaluates the degree of gender-related bias present in the documents retrieved by various IR models. The framework is structured into two main components: Document Gender Magnitude Measurements and Retrieval of Gender Bias Metrics.

Document Gender Magnitude Measurements. The first step in the framework involves quantifying the presence of concepts belonging to different gender identities within a document. This process is known as measuring the document's gender magnitude. Here's a detailed breakdown of this process:

1. Defining Gender Concepts: To measure gender-related content, we start by defining gender concepts using a set of highly representative words, known as gender-definitional words. For female-related concepts, the set includes words such as 'she', 'woman', and 'her', and for male-related concepts, the set includes words such as 'he', 'man', and 'him'. This selection is not exhaustive but serves as a practical approximation using highly representative terms.

2. Calculating Gender Magnitude: The gender magnitude of a document can be measured using two methods:

   - **Term Frequency (TF) Method:** This method calculates the sum of the logarithm of the occurrences of gender-definitional words within the document. Mathematically, for female-related words, it is expressed as:

   $$mag_f(d) = \sum_{w \in G_f} \log(\#\langle w, d \rangle) \tag{3.18}$$

   where G_f is the set of female definitional words, and $\#\langle w, d \rangle$ denotes the number of occurrences of word w in document d.

   - **Boolean Method:** This method checks for the presence of any gender-definitional word in the document. If any such word is found, the magnitude is set to 1; otherwise, it is set to 0. Mathematically, it is expressed as:

   $$mag_f(d) = \begin{cases} 1 & \text{if } \sum_{w \in G_f} \#\langle w, d \rangle > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.19}$$

Similarly, the male magnitude is calculated using the male definitional words set G_m

3. Given the document gender magnitudes, the next step is to measure the overall gender bias in the ranking lists produced by an IR model for a set of queries. This is done using two metrics: Rank Bias (RaB) and Average Rank Bias (ARaB).

   - **Rank Bias (RaB)**: This metric evaluates the average gender magnitudes of the top t retrieved documents for each query. For a given query q, the Rank Bias for female-related words at cutoff t (denoted as $qRaB^t_f(q)$) is calculated as:

     $$qRaB^t_f(q) = \frac{1}{t}\sum_{i=1}^{t} mag_f(d(q)_i) \tag{3.20}$$

     where $d(q)_i$ represents the document at the ith position in the ranking list for query q. The overall Rank Bias for a model over all queries Q is then defined as:

     $$RaB_t = \frac{1}{|Q|}\sum_{q\in Q}(qRaB^t_m(q) - qRaB^t_f(q)) \tag{3.21}$$

     A positive RaB value indicates a bias towards male concepts, while a negative value indicates a bias towards female concepts.

   - **Average Rank Bias (ARaB):** This metric incorporates the ranking positions of documents, similar to the Average Precision metric. For a given query q, the ARaB for female-related words at cutoff t (denoted as $qARaB^t_f(q)$) is calculated as:

     $$qARaB^t_f(q) = \frac{1}{t}\sum_{x=1}^{t} qRaB^x_f(q) \tag{3.22}$$

     The overall Average Rank Bias for a model over all queries Q is defined as:

     $$ARaB_t = \frac{1}{|Q|}\sum_{q\in Q}(qARaB^t_m(q) - qARaB^t_f(q)) \tag{3.23}$$

This framework provides a systematic approach to measure and compare gender bias in the retrieval results of various IR models. The inclusion of both term frequency and boolean methods for calculating document gender magnitude, along with rank-based and average rank-based metrics for retrieval gender bias, ensures a comprehensive evaluation of gender bias in IR systems. However, the limitations of the framework can be listed as follows. The framework does not account for the inherent bias present in the document collection itself. However, since the goal is to compare different IR models using the same collection, this limitation does not impact the comparative analysis. In addition, the metrics calculate average bias values per query, which does not reflect the distribution of these values. Further research could explore additional metrics that consider the distribution of bias values across queries.

3.2.5 Linguistic Inquiry and Word Count Toolkit

The Linguistic Inquiry and Word Count Toolkit (LIWC) is widely recognized as a powerful text analysis tool for categorizing words into various psychological and linguistic dimensions. Developed by James W. Pennebaker, Martha E. Francis, and Roger J. Booth, LIWC (Pennebaker *et al.*, 2001b) has been extensively applied in psychological and linguistic research to analyze both written and spoken language.

In the context of evaluating gender affiliation, LIWC uses the social referents category, which includes subcategories for male and female references. These subcategories consist of words commonly associated with distinct gender roles and identities. By counting the frequency of male- and female-related reference words in a given text, LIWC provides a measure of gender affiliation within the analyzed content.

As outlined by Bigdeli (2021), the difference between male and female affiliations can serve as an indicator of gender bias in a text. A positive difference suggests a stronger association with male references, while a negative difference indicates a stronger association with female references. This method can be extended to evaluate overall gender bias across a collection of texts by averaging the gender bias scores of individual documents, enabling the assessment of bias at a broader level.

A lower LIWC-derived score for gender bias is generally desirable, reflecting a more balanced representation of gender in the text. When applied systematically, this approach provides valuable insights into how gender is represented and can help identify areas where text may exhibit disproportionate bias. This aligns with similar metrics like ARaB in addressing gender bias within linguistic content.

Measuring Fairness in Query-Document Relations.

This section presents an evaluation framework designed to assess the fairness of an information retrieval system by examining the fairness of the ranked list of documents for each query. A detailed explanation of the measurement process is provided to ensure clarity and understanding. In the first step, document neutrality is assessed to determine whether a document is neutral or balanced regarding protected attributes such as gender. A document is considered neutral if it does not indicate any member of the protected attribute or represents all members equally. To establish a balanced representation, an expected proportion (J) for each protected member, such as male or female, is determined. For instance, in a binary gender setting, $J_{\text{female}} = 0.5$ and $J_{\text{male}} = 0.5$ are set. Then the magnitude of existence (mag) of representative words for each protected member in the document is calculated as:

$$\text{mag}_a(d) = \sum_{w \in V_a} \#(w, d) \tag{3.24}$$

where $\#(w, d)$ is the number of times word w appears in document d. The neutrality score ω of a document d is then computed as:

$$\omega(d) = \begin{cases} 1 & \text{if } \sum_{a \in A} \text{mag}_a(d) \leq \tau \\ 1 - \sum_{a \in A} \left| \frac{\text{mag}_a(d)}{\sum_{x \in A} \text{mag}_x(d)} - J_a \right| & \text{otherwise} \end{cases} \tag{3.25}$$

A threshold τ is set to reduce noise, ensuring documents with few representative words are considered neutral. The neutrality score $\omega(d)$ ranges from 0 to 1, where 1 indicates full neutrality, and 0 indicates dominance of one member of the protected attribute.

Fairness in a ranked list is evaluated by considering the neutrality of documents and their positions in the list. Higher-ranked positions have more influence. The fairness metric (FaiRR) is defined using the formula:

$$\text{FaiRR}_q(L) = \sum_{i=1}^{t} \omega(L_q^i) p(i) \tag{3.26}$$

where $p(i)$ represents position bias, typically using the Discounted Cumulative Gain (DCG) formula:

$$p(i) = \frac{1}{\log_2(1+i)} \tag{3.27}$$

To ensure comparability across queries, FaiRR is normalized using an ideal fairness score (IFaiRR), which represents the best possible fairness by reordering the documents in a set. The normalized fairness (NFaiRR) is calculated using the formula:

$$\text{NFaiRR}_q(L, S) = \frac{\text{FaiRR}_q(L)}{\text{IFaiRR}_q(S)} \tag{3.28}$$

The overall NFaiRR for an IR model is the average of NFaiRR scores across all queries:

$$\text{NFaiRR}(L, S) = \frac{1}{|Q|} \sum_{q \in Q} \text{NFaiRR}_q(L, S) \tag{3.29}$$

Ranker-agnostic fairness is measured to evaluate fairness independent of specific ranking models by averaging the neutrality scores of all possible document permutations. The ranker-agnostic fairness metric (SetFaiRR) is defined using the formula:

$$\text{SetFaiRR}_q(S) = E_{L \sim \Psi(S_q)} \left[\text{FaiRR}_q(L) \right] \tag{3.30}$$

which can be simplified to:

$$\text{SetFaiRR}_q(S) = t \times \left(\frac{1}{|S_q|} \sum_{d \in S_q} \omega(d) \right) \sum_{i=1}^{t} p(i) \tag{3.31}$$

This is normalized similarly to NFaiRR:

$$\text{NFaiRR}_q(S, S) = \frac{\text{SetFaiRR}_q(S)}{\text{IFaiRR}_q(S)} \tag{3.32}$$

The described framework provides a robust method for measuring gender biases in IR systems. By defining document neutrality, establishing fairness metrics for ranked lists, and implementing comprehensive fairness evaluations, more equitable IR models can be developed. This methodology can be extended to include other protected attributes, further enhancing the fairness of IR systems. Future research will refine these measurements and explore their impact on user perceptions and real-world applications.

TExFAIR: Assessing Group Fairness in Ranked Lists

In the following, we review a metric called TExFAIR (term exposure-based fairness)(Abolghasemi *et al.*, 2024) to assess fairness in the representation of groups in a ranked list. This approach is designed to address the limitations of existing metrics like NFaiRR (Rekabsaz *et al.*, 2021), which measure bias at the document level and can miss nuanced group representations across the entire ranked list.

The metric TExFAIR is grounded in two key extensions to a generic fairness evaluation framework known as attention-weighted ranking fairness (AWRF). These extensions include a probabilistic term-level association of documents to groups and a rank-biased discounting factor (RBDF) to account for non-representative documents.

The term exposure (TE) is defined to quantify the amount of attention each term receives in a ranked list for a given query. Formally, the term exposure for a term t in a list of k documents L_q retrieved for a query q is calculated as follows:

$$TE@k(t, q, L_q) = \sum_{r=1}^{k} po(t|d_r^q) \cdot po(d_r^q) \tag{3.33}$$

Here, d_r^q is the document at rank r in the ranked results for query q. The probability of observing term t in document d_r^q, denoted $po(t|d_r^q)$, is estimated using the frequency of term t in d_r^q normalized by the total number of terms in d_r^q. The observation probability $po(d_r^q)$ is assumed to depend only on the rank position and is estimated using the position bias $(\log(r+1))^{-1}$. Therefore, the term exposure can be reformulated as:

$$TE@k(t,q) = \sum_{r=1}^{k} \frac{tf(t,d_r^q)}{|d_r^q|} \cdot \frac{1}{\log(r+1)} \tag{3.34}$$

The representation of each group G_i in the ranked list is then estimated by leveraging the exposure of its representative terms. This is expressed as:

$$p(G_i|q,k) = \frac{\sum_{t \in V_{G_i}} TE@k(t,q)}{\sum_{G_x \in G} \sum_{t \in V_{G_x}} TE@k(t,q)} \tag{3.35}$$

Here, V_{G_i} denotes the set of terms representing group G_i. The denominator represents the total attention on all group-representative terms in the ranked list.

To evaluate fairness, the term exposure-based divergence (TED) is used. TED measures the absolute divergence between the actual group representations and their target representations. Let $\hat{p}_{G_i}$ be the target representation for group G_i. Then TED is defined as:

$$TED(q,k) = \sum_{G_i \in G} |p(G_i|q,k) - \hat{p}_{G_i}| \tag{3.36}$$

Non-representative documents, which do not include any group-representative terms, are addressed using the rank-biased discounting factor (RBDF). This factor discounts the bias by considering the proportionality of representative documents:

$$RBDF(q,k) = \frac{\sum_{r=1}^{k} \frac{1[d_r^q \in SR]}{\log(1+r)}}{\sum_{r=1}^{k} \frac{1}{\log(1+r)}} \tag{3.37}$$

Here, SR represents the set of representative documents, and $1[d_r^q \in SR]$ is an indicator function that equals 1 if d_r^q is a representative document.

Finally, TExFAIR is defined by incorporating the rank-biased discounting factor into TED. The resulting fairness measure, which adjusts for the presence of non-representative documents, is given by:

$$TExFAIR(q,k) = 1 - \left(TED(q,k) \cdot \frac{\sum_{r=1}^{k} \frac{1[d_r^q \in SR]}{\log(1+r)}}{\sum_{r=1}^{k} \frac{1}{\log(1+r)}} \right) \tag{3.38}$$

This metric ensures that fairness is assessed based on the overall representation of groups in the ranked list, accounting for term-level nuances and the position of non-representative documents.

3.3 Benchmarking Gender Fairness

Gendered-ness of queries and documents can be defined in multiple ways, depending on the perspective from which gender is analyzed. One key perspective is the lexical definition of gender, where gender is inferred based on the explicit or implicit presence of gendered words in the text. Explicit indicators include gendered pronouns ('he,' 'she'), gendered nouns ('man,' 'woman'), or professions explicitly marked by gender ('policeman,' 'policewoman'). For example, a query like 'female CEO leadership strategies' or a document discussing 'the contributions of male nurses' clearly signals gender through direct lexical cues. However, lexical gendered-ness can also be implicit, where certain words or topics inherently suggest a specific gender without using explicitly gendered terms. For instance, a query about 'pregnancy nutrition tips' implicitly points towards a female subject, as pregnancy is strongly associated with women. Beyond lexical cues, subject-based gendered-ness refers to the gender of the entities or individuals discussed in a document, such as the main character in a novel, the subject of a biography, or the individuals featured in a news article. A biography titled 'The Life of Marie Curie' would be considered female-gendered due to its subject. Another dimension is authorial gender, where gender is determined by the identity of the author or creator of the text. For example, a memoir written by a female author might be considered female-gendered, regardless of its content. These definitions highlight that gendered-ness in text can emerge from linguistic choices, subject focus, or authorial perspective, and these facets are not mutually exclusive. In this manuscript, given the current predominant focus on the literature (albeit not the sole focus), we review the lexical definition of gender, emphasizing how the textual content of queries and documents — both explicitly and implicitly — reflects gendered-ness. As such and in this manuscript, a query or document is considered gendered if it includes explicit or implicit gendered words. In the absence of such indicators, it is classified as neutral. This approach provides a measurable foundation for analyzing gender bias in information retrieval systems, while acknowledging that other dimensions of gender representation remain valuable for broader contextual analyses.

This approach to defining gendered-ness is limited by its reliance on textual cues, which may overlook the complexities of gender representation and its intersectionality with other social factors. Implicit gendered-ness, for instance, is context-dependent and may vary across cultures or individual interpretations, making it challenging to generalize. Additionally, focusing on lexical definitions risks oversimplifying gender as a binary concept, excluding non-binary and fluid identities. We acknowledge that while this framework provides a measurable starting point, it does not capture the full scope of gender representation, which may extend beyond textual content to societal, historical, and cultural contexts.

3.3.1 IR Datasets

Several datasets have been proposed to identify and measure gender biases in NLP and information retrieval systems (Jha and Mamidi, 2017; Rodríguez-Sánchez *et al.*, 2022; Samory *et al.*, 2021). These datasets incorporate queries labeled with gender information, which can serve multiple purposes, such as analyzing search engine behavior, studying gender bias in human judgment, and analyzing human query generation. In this section, we will provide an in-depth explanation of these datasets, including their samples, statistical characteristics, and limitations. An overview of the datasets is included in Table 3.1.

Gendered Queries

This dataset is proposed by Rekabsaz and Schedl (2020). The creation of the gender-annotated queries dataset involved several key steps aimed at ensuring the inclusion of queries that do not contain any gender-specific elements. This process was essential to accurately measure the gender bias present in the retrieval models. Here is a detailed explanation of how the dataset was created:

1. **Query Selection**: The queries were selected from the test set of the MS MARCO Passage Retrieval collection (Nguyen *et al.*, 2016), a dataset comprising 8,841,822 passages and a large set of informational question-style queries from Bing's search logs. The initial selection focused on queries whose ranked list of documents displayed the highest inclinations towards gender, determined by the retrieval results of seven ranking models.

For all the Information Retrieval (IR) models, the retrieval gender bias of each test set query was calculated using the Term Frequency (TF) gender magnitude measure and the Rank Bias (RaB) approach at a cutoff of 10 (explained in section 3.2.4). This process generated two separate lists of queries for each of the seven IR models studied: one list for queries biased toward females and another for queries biased toward males. Consequently, this resulted in a total of 14 lists of sorted queries. A pooling method introduced by Jones (1975) was applied to these sorted lists with a cutoff of 500, leading to a total of 3,924 unique queries. This method ensures a comprehensive selection of queries, capturing various degrees of gender bias as perceived by different models.

2. **Human Annotation**: The next step involved human annotation to categorize the queries accurately. Three Amazon Mechanical Turk workers were tasked with classifying each query into one of four categories:

- *Non-gendered*: Queries that do not refer to any specific gender.
- *Female*: Queries containing words or phrases related to female concepts (e.g., queen, pregnant).
- *Male*: Queries containing words or phrases related to male concepts (e.g., king, father).
- *Other or Multiple Genders*: Queries that refer to other genders or multiple genders (e.g., transgender, references to both male and female).

The detailed descriptions and guidelines for these categories were provided to the annotators to ensure consistency and accuracy. Based on the annotations, each query was assigned to a category using the majority vote of the annotators. Queries that did not reach an unambiguous majority decision (i.e., each annotator chose a different category) were removed from the dataset. This step was crucial to maintain the reliability of the dataset. The details of the dataset are included in Table 3.1.

MSMARCOFair: Gender-neutral Queries

The process of creating this dataset involves several detailed steps aimed at identifying and annotating fairness-sensitive queries related to gender equality. Rekabsaz *et al.* (2021) began with an initial selection of queries from two prominent datasets: the TREC Deep Learning Track 2019 Passage Retrieval (TRECDL19) (Craswell *et al.*, 2020) and the development set of the MSMARCO Passage Re-ranking collection(Nguyen *et al.*, 2016). Specifically, they selected 1,765 non-gendered queries from the MSMARCO collection, which had been previously annotated by Amazon Mechanical Turk workers.

Next, the researchers employed three Amazon Mechanical Turk workers, all native English speakers, to annotate the queries from TRECDL19 in a similar manner. This crowdsourced annotation ensured consistency across both datasets. Following this, a meta-annotation process was carried out to verify the initial annotations and identify queries where the presence of gender bias in retrieval results would be socially problematic.

During the meta-annotation process, the researchers evaluated each query on two criteria: whether it was non-gendered and whether gender bias in its retrieval results would be socially problematic. Socially problematic queries were identified based on their potential to reinforce existing gender norms and promote gender inequality. The researchers focused on domains such as education, career, health, violence, exploitation, social inequality, and politics. For example, a query like 'how important is a governor?' was marked as fairness-sensitive because bias in this context could reinforce career stereotypes. Another query, 'When do babies start eating whole foods?' was identified as problematic due

Table 3.1: Overview of the datasets for gender bias in information retrieval.

Query set	queries	neutral	male	female	other	human annotator
Gendered Queries	3,750	1,765	1,202	742	41	✓
$MSMARCOFair$	215	215	-	-	-	✓
$TRECDL19_{Fair}$	30	30	-	-	-	✓
BERT Gendered Queries	51,827	48,200	2,222	1,405	-	✗
Grep-BiasIR	118	-	-	-	-	✓

to its potential to reinforce the stereotype of 'women as caretakers', thereby impacting career choices and perpetuating gender norms.

The final step involved compiling the datasets, ensuring only those queries agreed upon by both meta-annotators were included. This resulted in the MSMARCOFair dataset containing 215 queries and the TRECDL19Fair dataset including 30 queries. These datasets were designed to serve as benchmarks for studying fairness in retrieval results, enabling research on fairness alongside utility in information retrieval models.

BERT-annotated Gendered Queries

To begin, Bigdeli (2021) employed a publicly available gender-annotated dataset provided by Rekabsaz et al. (2020), which includes queries labeled as non-gendered (neutral), female, male, or other/multiple genders, as mentioned in section 3.3.1.

On this basis, they trained classifiers using both dynamic and static embeddings to predict the gender of queries. The performance of these classifiers was evaluated using a 5-fold cross-validation strategy. The fine-tuned uncased BERT model outperformed others, showing the highest accuracy and F1 scores for gender identification: 0.856 accuracy, 0.816 for female, 0.872 for male, and 0.862 for neutral queries.

Using the fine-tuned BERT model, the researchers labeled all 51,827 queries in the MS MARCO Dev set, resulting in 48,200 neutral queries, 2,222 male queries, and 1,405 female queries. To create a balanced dataset, they retained all 1,405 female queries and randomly selected 1,405 male and 1,405 neutral queries. These labeled queries, along with their associated relevant judgment documents, were used to investigate the presence of stereotypical gender biases.

Grep-BiasIR dataset

The Grep-BiasIR dataset (Krieg *et al.*, 2023), designed to investigate gender representation bias in information retrieval systems, comprises 118 bias-sensitive queries and 708 associated documents. The creation process began with the categorization of queries into seven gender-related stereotypical concepts based on the gender role dimensions introduced by Behm-Morawitz and Mastro. These categories include Career, Domestic Work, Child Care, Cognitive Capabilities, Physical Capabilities, Appearance, and Sex & Relationship. Each category contains around 15 queries, resulting in a well-rounded dataset that addresses a variety of gender-related topics.

For each query, the dataset includes one relevant and one non-relevant document. The relevant documents were identified by submitting the queries to the Google search engine and selecting documents that fully addressed the query's information need. Non-relevant documents were either taken from the same search results or created by the authors to ensure they did not match the search query. Each document is provided in three variations: male, female, and neutral. The variations maintain the same content, with gender-indicating words modified accordingly. For instance, male indications include words like 'man' and 'he', while female indications use 'woman' and 'she'. Neutral terms like 'person' and 'they' were used to create gender-neutral versions. This thorough and systematic approach ensures that the dataset can effectively facilitate the study of gender biases in information retrieval systems.

The dataset underwent rigorous auditing by two post-doctoral researchers who reviewed each query and document for quality. They judged the items as high, medium, or low quality and only high-quality items were included in the final dataset. This review process also involved checking for ambiguous content and ensuring that gender-neutral documents were properly formulated, such as using only surnames to avoid gender-specific references. Additionally, the reviewers assessed the expected stereotypes for each query based on anticipated gender characteristics and behaviors. This meticulous process of data collection and auditing ensures that the Grep-BiasIR dataset is a reliable and valuable resource for investigating gender representation biases in IR systems.

3.3.2 Gender Bias datasets in Natural Language Processing (NLP)

Similar to information retrieval, detecting and mitigating gender bias in natural language processing (NLP) is a critical task. Several recent datasets have been created to aid in this endeavor, providing annotated examples of various forms of gender bias. Below is a brief overview of some of the most significant datasets in this domain. An overview of the datasets is included in Table 3.2.

(Waseem and Hovy, 2016).

This dataset comprises tweets collected and annotated for instances of racism, sexism, or neither. It was created using various self-defined keywords to filter potentially sexist or racist tweets from the Twitter stream over two months. The dataset is foundational for research into hate speech and gender bias on social media platforms.

(Jha and Mamidi, 2017).

Augmenting the Waseem & Hovy dataset, this collection includes tweets that exhibit benevolent sexism. Benevolent sexism refers to statements with a positive tone that imply women need special treatment and protection from men, thereby reinforcing stereotypes about women's capabilities. Three external annotators cross-validated the tweets to ensure consistency.

AMI@Evalita (Fersini *et al.*, 2020).

Created for the Automatic Misogyny Identification (AMI) task at the Evalita 2020 competition, this dataset includes instances of misogynistic content. It is used to classify texts as misogynous or not misogynous and to identify the targets of misogynous texts. The dataset is available in multiple languages, including English, Spanish, and Italian.

AMI@IberEval (Fersini *et al.*, 2018).

Similar to the AMI@Evalita dataset, this collection was developed for the IberEval 2018 competition. It focuses on identifying misogynistic content in texts, providing annotations to help detect and classify such content. The dataset is instrumental in developing and evaluating models aimed at recognizing and mitigating misogyny online.

EXIST@IberLEF (Rodríguez-Sánchez *et al.*, 2021).

This dataset was created for the EXIST (sEXism Identification in Social neTworks) task at the IberLEF 2021 competition. It includes tweets annotated for sexism and non-sexism, aiding in the detection and analysis of sexist content on social media. The dataset supports efforts to address gender bias and promote gender equality in online communication.

'Call Me Sexist' (Samory *et al.*, 2021).

Collected using the phrase "call me sexist(,) but," this dataset comprises tweets retrieved from Twitter that potentially contain sexist remarks. Annotated via crowd-sourcing, the dataset focuses on the content following the phrase to determine sexism and toxicity. It provides insights into how disclaimers are used in sexist remarks and their impact on online discourse.

(Chowdhury *et al.*, 2019).

This dataset aggregates tweets using the 'MeToo' hashtag to collect personal recollections of sexual harassment. It offers valuable insights into the experiences of sexual abuse shared on social media and helps understand the social media constructs that facilitate such discussions. The dataset is beneficial for clinicians, health practitioners, caregivers, and policymakers to identify at-risk communities and address issues of sexual harassment.

These datasets provide a comprehensive foundation for developing and evaluating models that detect various forms of gender bias in textual

Table 3.2: Overview of the datasets for gender bias in natural language processing.

Dataset	Size	Labels
(Waseem and Hovy, 2016)	16K	Racism, Sexism, Neither
(Jha and Mamidi, 2017)	22K	Benevolent sexism, hostile sexism, others
(Fersini *et al.*, 2020)	10K	Misogynous, not misogynous
(Fersini *et al.*, 2018)	8K	Misogynous, not misogynous
(Rodríguez-Sánchez *et al.*, 2021)	11K	Sexist, not sexist
(Samory *et al.*, 2021)	14K	Sexist, not sexist, toxicity
(Chowdhury *et al.*, 2019)	5K	recollection of sexual harassment, not recollection

data, contributing significantly to the advancement of gender-inclusive NLP systems.

4

Understanding the Sources of Gender Bias in IR Systems

Information retrieval systems consist of three primary components: (1) input query, (2) retrieval method, and (3) gold standard documents. Ideally, the objective of a search engine is to retrieve the gold standard documents using one or more retrieval methods given a specific search query. However, if any of these components harbor biases, the list of retrieved documents presented to users will reflect these biases. Consequently, users are exposed to biased content, which can negatively influence their perceptions and judgments.

In this section, we explore the presence of gender biases within each of the aforementioned components and demonstrate the evidence of gender bias in each. Additionally, we provide a comprehensive analysis of how each component can serve as a source of gender bias, ultimately contaminating the fairness and accuracy of the information retrieval system.

4.1 Gender Bias in Input Query

The input query marks the initial point of interaction between the user and the information retrieval system. While users' search queries may appear gender-neutral, they can inherently carry social biases that affect the search engine's responses. This section will cover how these biases manifest and the implications of query reformulation on gender bias in search results.

4.1.1 Algorithmic Query Reformulation

Imagine a user entering the query 'top scientist'. On the surface, this query seems unbiased and straightforward. However, due to intrinsic societal biases, search engines might prioritize documents that highlight male scientists, assuming that scientists are predominantly male. This subtle bias can significantly impact the user's perception and the visibility of female scientists. When users submit queries, these can carry hidden social biases that influence the retrieval process. If the queries are socially problematic, they introduce biases into the retrieved documents. Therefore, understanding the biases in initial queries is crucial for developing fair and unbiased information retrieval systems.

Query reformulation methods, such as Pseudo-Relevance Feedback (PRF), can exacerbate these biases. The RM3 PRF method, for instance, enhances the original query by incorporating terms from the top-ranked documents retrieved by the initial query, assuming these documents are relevant. However, if the top-ranked documents are biased, the expanded query may reflect and amplify these biases. Consider the example of the query 'top scientists'. If the initial top-ranked documents are biased towards male scientists, the RM3 PRF method might expand the query to include terms like 'top scientists men male', leading to a biased set of results. This process illustrates how reformulated queries can perpetuate gender stereotypes and underscores the need for analyzing and addressing these biases.

To investigate the extent to which pseudo-relevance feedback methods, such as RM3, introduce biased terms, Bigdeli *et al.* (2021a) targeted a set of non-gendered news-related queries from different TREC corpora, including Robust04 (Voorhees, 2004), Gov2 (Clarke *et al.*, 2004), ClueWeb09 (Callan *et al.*, 2009), and ClueWeb12 (Clarke *et al.*, 2012). They used the *ARaB* metric (Rekabsaz and Schedl, 2020) to examine the level of bias among the top 10 documents retrieved by BM25 and the PRF model. The authors found that the reformulated queries included gender-specific terms not present in the original queries. For example, the query ‘Cult Lifestyles’ was reformulated to include terms such as ‘student Krishna Kim Chilton lifestyle **she mother** cult **her** car pension,’ or ‘american muslim mosques schools’ changed to ‘mosque muslim american wahhabi my saudi america religion Islam school **he**’ by the PRF model. These additions resulted in increased bias in the retrieved documents, demonstrating how PRF methods can inadvertently introduce and amplify gender bias in search results.

The findings highlight the importance of critically evaluating query reformulation methods to ensure they do not perpetuate existing biases. Researchers and developers must consider the potential for these methods to introduce bias and work towards creating fairer, more balanced information retrieval systems. This includes exploring alternative approaches to query reformulation that mitigate bias and developing metrics to measure and address bias in search results effectively.

4.1.2 User-Driven Query Reformulation

In an observational study by Raj *et al.* (2023), a large-scale search log data from Bing explored how users reformulate their queries to include gender-specific terms, a process called gender-specializing query reformulations (GSQR). The study identified approximately 4.7 million pairs of consecutive queries where the second query was a GSQR of the first. For instance, consider a user initially searching for ‘NCAA scores.’ If their interest lies specifically in women’s basketball, they might reformulate the query to ‘NCAA women’s scores.’ This simple modification demonstrates how users can introduce gender-specific terms to refine their search results. The study aimed to understand the contexts in which users reformulate their queries to include gender-specific terms and the impacts of these reformulations on search results.

The authors defined a query reformulation as specializing if the reformulated query contained all terms from the original query in the same order and included additional contiguous terms related to gender. They calculated the overall frequency of GSQRs and categorized the original queries by topic, revealing higher rates of GSQRs in categories such as shopping and fashion. This categorization provided insights into the contexts where users are more likely to introduce gender-specific terms.

The study also explored the timing and method of query reformulation. Time differences between the original and reformulated queries were analyzed to infer user behavior, showing a median time of 19 seconds between queries, with men-related reformulations occurring slightly more quickly. Additionally, the study examined how users entered their reformulated queries. It was found that most gender-specific query reformulations (GSQRs) were made either by editing the original query directly in the search bar at the top of the search results page (SERP) or by selecting one of the recommended queries provided by the search engine.

Furthermore, the study investigated the genderedness of original queries using average GloVe (Pennington *et al.*, 2014) embeddings of terms. This analysis revealed instances where GSQRs corrected the under-representation of a gender or reinforced existing gender representations. For example, the query 'ADHD symptoms' might be reformulated to 'ADHD symptoms for women' to obtain gender-specific information, while a query like 'NCAA basketball score' might be reformulated to 'NCAA men's basketball score' to emphasize men's results.

These findings highlight the active role users play in shaping search results through their query reformulations. Understanding user behavior in this context is crucial for designing search systems that accommodate user needs while mitigating the introduction of bias. It also underscores the importance of providing users with tools and options to refine their searches in a way that promotes fairness and diversity in the retrieved results.

4.2 Gender Bias in Retrieval Methods

Retrieval methods play a pivotal role in retrieving relevant documents in response to user input queries. However, when these methods lack awareness of bias, they may inadvertently retrieve biased document sets, especially in response to sensitive queries. The algorithms and techniques employed by retrieval methods to match the input query with relevant documents can be a source of gender biases. These methods are often based on Pre-trained Language Models (PLMs) trained on large datasets that contain human-generated content such as online forums, books, articles, web pages, etc. If the training data itself contains gender biases, the retrieval algorithms capture those biases from data as a form of association and likely reproduce and even amplify these biases. For example, if a search engine's algorithm has been trained on data that overrepresents male achievements in science and underrepresents female achievements, it will preferentially retrieve documents about male scientists, even when the query is a gender-neutral one.

4.2.1 Gender Bias in Embedding Models

In the field of Natural Language Processing (NLP), studies have investigated the presence of gender biases in embedding representations of word-embedding models (Bolukbasi *et al.*, 2016; Caliskan *et al.*, 2017) such as Word2Vec (Mikolov *et al.*, 2013) and pre-trained language models (May *et al.*, 2019; Zhao *et al.*, 2019) like ELMo (Matthew, 2018) and BERT (Devlin *et al.*, 2018b). For instance, Bolukbasi *et al.* (2016) provided significant insights into gender biases within Word2Vec and GloVe word embeddings. They introduced a geometric framework to identify gender in the embedding space and evaluated whether the embeddings of occupations exhibit stereotypical gender biases. Additionally, they examined if the embeddings generate analogies that humans perceive as reflecting gender stereotypes. To investigate these, they created a gender subspace by considering 10 pairs of female and male words such as ('she', 'he') and ('mother', and 'father') and subtracting the embedding representation of each pair's word embeddings to form matrics of ten vector embeddings and finally applied singular value decomposition to obtain *she-he* gender subspace (as outlined in Section 3.2.2).

Followed by that, this gender subspace was projected into male-stereotypic and female-stereotypic occupations and it has been shown that it is strongly correlated with the annotations of ten crowd-workers who were asked to annotate the occupations into male, female, or neutral. Examples of extreme male and female occupations are captain and nurse, respectively. To investigate if analogies also reflect stereotypes, the authors modified the standard analogy task to generate pairs of words, allowing the systematic creation of analogies based on seed words like ‘he’ and ‘she.’ Using a scoring metric based on cosine similarity and a semantic coherence threshold, they identified top analogous pairs. Finally, crowd workers evaluated these pairs to determine if they made sense and reflected gender stereotypes. The number of workers identifying a stereotype in each analogy quantified the degree of bias. The results of the experiment showed that 72 out of 150 analogies were considered gender-appropriate, while 29 exhibited gender stereotypes.

In addition to static word embeddings, contextualized embeddings also demonstrate stereotypical gender biases, as evidenced by Zhao *et al.* (2019). In their study, 400 sentences were selected from the OntoNotes 5.0 dataset (Weischedel *et al.*, 2012), each containing at least one gendered word (e.g., ‘he’ or ‘she’). Subsequently, they created a gender-swapped version for each sampled sentence and analyzed the disparities in ELMo embeddings among the occupation words in these paired sentences. Principal component analysis (PCA) was then applied to these differences.

The results revealed the existence of two principal components related to gender in ELMo embeddings: one representing contextual gender information and the other representing gender inherent in the occupation word itself. The first principal component segregates male and female contexts, while the second component clusters male-associated and female-associated occupation words. This analysis underscores the nuanced depiction of gender in ELMo embeddings, encompassing both contextual and lexical gender information.

4.2.2 The Impact of Neural Embeddings on Retrieval Methods

With the emergence of neural embeddings, retrieval methods shifted from term-frequency-based methods to neural-based retrieval methods that leverage different variations of static word embeddings and contextualized word embeddings for the task of retrieving relevant documents given a search query. For instance, retrieval methods such as KNRM (Xiong *et al.*, 2017), Conv-KNRM (Dai *et al.*, 2018), DRMM (Guo *et al.*, 2016), DUET (Mitra *et al.*, 2017), and MatchPyramid (Pang *et al.*, 2016) use Word2Vec or GloVe word embeddings for matching query and documents representation.

Despite the improved performance of neural-based retrieval methods, they often rely on static or contextualized embedding models that have been found to contain and amplify gender biases (Bolukbasi *et al.*, 2016; Zhao *et al.*, 2019). Moreover, these embedding representations, already biased, are fine-tuned based on gold standard collections to learn query-document semantic mapping. During the fine-tuning process, there is a risk of further reinforcing gender biases. For instance, if a relevant document pair for a query is biased towards a specific gender attribute, the model may capture and reflect this biased association during the ranking process.

To investigate how gender biases within static and contextualized embeddings manifest at the ranking stage by retrieval methods, several studies have measured the level of gender biases among the ranked lists of documents for queries using different retrieval methods (Fabris *et al.*, 2020; Rekabsaz and Schedl, 2020; Rekabsaz *et al.*, 2021). Fabris *et al.* (2020) employed term-frequency-based retrieval methods such as TF-IDF and BM25, as well as neural-based retrieval methods including DRMM and MatchPyramid, to compare the level of gender bias among the retrieved lists of documents for a subset of queries using the GSR framework introduced in Section 3.2.4. Based on their experimental results, traditional lexical models such as TF-IDF and BM25 exhibited low GSR, whereas semantic models based on biased word embeddings tended to reinforce gender stereotypes, even with IDF-inspired weighting schemes.

Rekabsaz and Schedl (2020) released a set of gender-neutral queries consisting of 1,765 queries annotated by annotators as being non-gendered, as outlined in Section 3.3.1. Using this gender-neutral set of queries, BM25, static and contextualized embedding retrieval methods such as BERT, KNRM, and MatchPyramid are employed to rank the top 1000 documents retrieved by BM25. They then measured the level of bias within the ranked lists of documents from each of these retrieval methods using term-frequency-based and boolean-based variations of the *ARaB* metrics mentioned in Section 3.2.4. The results of the comparison, in terms of *ARaB* metrics, demonstrated that all retrieval methods, especially neural models and BERT, showed a bias towards male concepts, with neural models consistently increasing retrieval gender bias compared to BM25. Furthermore, the use of pre-trained word embeddings in neural ranking models tends to increase gender bias.

In another study by Rekabsaz *et al.* (2021), the authors investigated the level of fairness within the retrieved list of documents ranked by different retrieval methods across $MSMARCO_{FAIR}$ and $TRECDL19_{FAIR}$ queries introduced in Section 3.3. They proposed NFaiRR metric to measure fairness across the top-k ranked list of documents for the two sets of fair queries by different retrieval methods. They evaluated the fairness metric across ranked lists of documents by retrieval methods such as BM25, KNRM, MatchPyramid, BERT-Tiny, and BERT-Mini. Based on the results, ranker-agnostic document sets reveal that in the $MSMARCO_{FAIR}$ collection, the NFaiRR of SetTop200 (top-200 documents retrieved by BM25) is slightly lower than SetAll (top-1000 documents retrieved by BM25), indicating higher gender bias in the top retrieved documents compared to the entire collection. This suggests that $MSMARCO_{FAIR}$ queries tend to pull documents from biased subspaces. Conversely, in the $TRECDL19_{FAIR}$ collection, SetTop200 is more fair than SetAll, approaching ideal fairness, suggesting that $TRECDL19_{FAIR}$ queries lead to balanced gender representation. For ranking models in $MSMARCO_{FAIR}$, classical retrieval models such as BM25 exhibit the lowest fairness, while neural models show significantly higher fairness scores, with BERT rankers achieving the best results.

In conclusion, the investigation into language models and retrieval methods has highlighted significant gender biases that can affect the relevance and fairness of retrieved document sets. These biases originate from the training data and are further amplified by the algorithms and pre-trained language models used in retrieval systems. Studies have shown that neural-based retrieval methods, despite their improved performance, tend to increase gender bias compared to traditional models like BM25. This is evident in models that use both static word embeddings and contextualized embeddings, which often capture and perpetuate gender stereotypes. Consequently, there is a pressing need to apply de-biasing methods to retrieval algorithms. Implementing such methods can mitigate these biases, ensuring that retrieval systems provide fairer and more balanced results, particularly in response to gender-neutral queries.

4.3 Gender Bias in Gold Standard Datasets

One of the main, if not the main, sources of gender biases in both embedding models and retrieval methods used for ranking relevant documents in search queries is the use of biased training datasets. Numerous NLP research studies have statistically analyzed the ratio of male and female-related terms within the datasets used for training neural models. These studies have shown that such training datasets contain considerable gender bias, leading models to learn and reproduce associations between occupations or other entities with gender. For example, in (Dinan *et al.*, 2019), the authors analyzed six dialogue datasets for gender bias and found a significant bias towards males. Specifically, the LIGHT text adventure world dataset (Urbanek *et al.*, 2019) was identified as the most biased, with male bias reaching 73%. This high level of bias is attributed to the dataset's multiple potential sources of bias, its crowdsourced nature, and its medieval, fantasy setting, which may reflect the gender biases of the crowdworkers.

In another study by Zhao *et al.* (2019), the authors targeted the One Billion Word Benchmark corpus (Chelba *et al.*, 2013) and calculated the occurrences of male pronouns (he, his, him) and female pronouns (she, her) in the corpus, as well as the co-occurrence of these pronouns with occupation words. The set of occupation words and their gender assignments were based on the WinoBias corpus (Zhao *et al.*, 2018a). The findings revealed a significant gender skew in the Billion Word corpus as it could be observed that male pronouns tend to appear three times more frequently than female pronouns, and male pronouns were more commonly associated with occupation words.

Investigating gender biases in gold-standard training datasets for information retrieval methods is imperative. Quantifying these biases is crucial because neural-based retrieval methods are often trained on such data and biased training data can transfer biases into the algorithmic and representational aspects of retrieval methods. This bias can ultimately affect the ranking process by causing models to associate gender with query needs for sensitive queries, leading to biased documents being ranked higher and increasing their exposure to users.

Gold standard datasets for training retrieval methods are typically obtained from crowdworkers' annotations, where workers are tasked with identifying relevant documents to queries. Gender bias within these datasets can arise from two primary sources:

1. **Perceived bias in human judgments:** Crowdworkers assigned to annotate relevant documents may inadvertently inject their own biases into the dataset. These biases can be influenced by various factors such as cultural background, personal beliefs, or societal stereotypes. For instance, a crowdworker's cultural background might influence their interpretation of what constitutes relevance, potentially leading to the inclusion or exclusion of certain documents based on gender-related assumptions or stereotypes.

2. **Inherent biases in the content:** The material available for annotation might inherently contain biases due to the nature of its source or the domain it represents. This could encompass biases in language usage, representation of specific demographics, or the framing of topics. For instance, documents sourced from certain domains or publications might predominantly focus on specific gender-related issues, thus skewing the dataset towards particular gender perspectives.

In this section, we thoroughly investigate each of these potential sources of bias to understand their implications within the context of gender biases in information retrieval datasets.

4.3.1 Impact of Perceived Bias on Human Judgements

Information retrieval models are typically trained and evaluated using a collection of relevance judgments determined by human assessors. If these documents display biases toward a particular gender, such biases have the potential to manifest in the retrieved list of documents presented to users. It is plausible that biases ingrained within individuals' mental frameworks may influence their decisions regarding the relevance or irrelevance of information (Ellemers, 2018; Ellis, 2018; Swim *et al.*, 1989).

In this section, we explore the critical question of whether these perceived biases impact human decision-making during the document judgment process, based on the research work by Krieg *et al.* (2022). The experimental setup of this study employs the Grep-BiasIR dataset, introduced in Section 3.3.1, which includes bias-sensitive queries and documents with varying gender indications (male, female, and neutral). The experiments are conducted in two distinct settings: gender-specific (where the gender of the participants is known) and gender-agnostic (where the participants' gender is unknown).

The queries in this study span six categories: Appearance, Career, Domestic Work, Child Care, Cognitive Capabilities, and Physical Capabilities. Each query is paired with relevant and non-relevant documents presented in both male and female versions. Participants from the Amazon Mechanical Turk platform were tasked with rating the relevance of these query-document pairs on a scale from non-relevant to perfectly relevant. Relevance judgments were collected from 50 participants for the gender-agnostic setting and from 10 male and 10 female participants for the gender-specific setting.

The study aimed to investigate three primary hypotheses. First, it examined whether relevant documents aligned with expected gender stereotypes received higher relevance scores (H1). Second, it assessed whether non-relevant documents reflecting gender stereotypes also influenced relevance scores (H2). Lastly, the research explored whether participants' gender influenced their judgments of gender-biased content (H3). The design of the study allowed the researchers to explore these hypotheses comprehensively within the controlled experimental setup.

In the gender-agnostic experiments, findings revealed that participants generally rated documents aligning with expected gender stereotypes higher in relevance. For example, in the Domestic Work category, relevant documents with female-indicating content scored higher than those with male content. This trend aligns with societal stereotypes where women are often associated with domestic responsibilities. However, statistical significance was limited, indicating a subtle but observable influence of biases. For non-relevant documents, the results were mixed, with some categories showing stereotype-disconfirming content being rated as more relevant, potentially due to the surprising nature of such information. For instance, in the Appearance category, male-indicating non-relevant documents were rated higher, possibly reflecting a perceived novelty or unexpectedness in this context.

The gender-specific experiments aimed to assess whether participants' gender impacted their judgments of gender-biased documents. Results showed no significant interaction between participant gender and their relevance ratings for either relevant or non-relevant documents. For instance, male and female participants similarly rated queries such as "how to build muscles" and "what is considered plus size" with no significant deviations linked to their gender. This suggests that gender bias in perceived relevance judgments is not directly influenced by the annotator's gender under the study's experimental conditions.

The study's hypotheses regarding stereotype confirmation (H1), stereotype-confirming effects for non-relevant documents (H2), and the impact of participant gender (H3) were only partially supported. While relevance scores often aligned with stereotypes, the effect sizes were small, and statistical significance was generally not observed. An exception was observed in the Appearance category for non-relevant documents, where stereotype-disconfirming male content scored higher than female content, potentially influenced by how unexpected or surprising the content seemed.

The study is not without its limitations, which constrain the generalizability and interpretability of its findings. First, the sample size, particularly for the gender-specific experiments, was relatively small, reducing the statistical power to detect significant effects. Additionally, the study relied on binary gender classifications (male and female), which limits the scope of the findings and excludes insights from individuals who do not identify within these categories. Another limitation arises from the controlled nature of the experimental setup, where participants evaluated isolated query-document pairs without additional real-world contextual factors, such as document ranking, source credibility, or temporal relevance.

4.3.2 Bias in the Content of Gold Standard Datasets

The presence of gender biases within the content of relevance judgment collections is examined in this section. To achieve this, a three-staged methodological approach, as proposed by Bigdeli *et al.* (2021b), is employed. The first stage involves identifying and labeling queries based on their gender. To facilitate this, Rekabsaz and Schedl (2020), introduced a gender-annotated dataset (detailed in section 3.3.1), which included queries labeled as female, male, or neutral. Various models were trained to classify query gender, encompassing both dynamic embeddings (such as BERT, DistilBERT, RoBERTa, and XLNet) and static embeddings (including fastText and Word2Vec). The performance of these classifiers was rigorously evaluated using a 5-fold cross-validation strategy. The uncased fine-tuned BERT model demonstrated superior performance, with high accuracy and F1 scores across all gender classes. Subsequently, this model was employed to label the entire MS MARCO development query set for gender, resulting in a comprehensive dataset, as introduced in section 3.3.1.

In the second stage, the authors measured various psychological characteristics of the relevance judgment documents associated with gendered queries. Using the Linguistic Inquiry and Word Count (LIWC) toolkit (Pennebaker *et al.*, 2001a), they quantified affective processes, cognitive processes, drives, and personal concerns within gold standard documents associated with development set labeled queries from each of the male, female, and neutral categories. This stage was crucial for determining whether the psychological expressions within the documents aligned with known findings from psychological literature or exhibited stereotypical biases. This analysis helped to uncover implicit biases embedded in the content of the documents, providing a detailed understanding of how these biases manifest in relevance judgments.

The third stage involved reporting findings on gender stereotypical biases in the gold standard relevance judgments. The results demonstrated that documents related to female queries exhibited higher degrees of negative emotions such as anxiety and sadness, while male query-associated documents showed more anger. In terms of cognitive processes, documents associated with female queries demonstrated higher cognitive complexity. For drives, male query-related documents expressed more affiliation, achievement, and power, whereas female query-related documents emphasized reward and risk avoidance. Regarding personal concerns, documents linked to male queries had a higher focus on work and leisure. This stage revealed that the relevance judgment documents indeed reflected stereotypical gender biases, consistent with some psychological research findings while highlighting unexpected biases in the datasets.

These findings underscore the necessity of exploring de-biasing methods to mitigate prevalent gender biases in gold-standard datasets used for training retrieval methods.

5

Data-Driven De-Biasing Methods

In this chapter, we present data-driven techniques designed to enhance bias-aware model training, to address bias inherent in Machine Learning (ML), Natural Language Processing (NLP), and Information Retrieval (IR) models. These de-biasing strategies aim to mitigate bias while preserving the original model architecture, achieving this solely through adjustments to the training data to refine the model's learning approach. These straightforward yet powerful methods have proven highly effective in reducing biases inherent in both the algorithmic and representational aspects of neural models.

A substantial body of research focuses on de-biasing neural-based models by balancing the training data fed to the model during the training phase. In this section, we delve into existing balancing strategies employed for de-biasing:

- **Machine Learning Models:** Techniques such as re-weighting, re-sampling, and synthetic data generation are explored. Re-weighting adjusts the importance of different training samples based on their associated bias, while re-sampling involves either over-sampling underrepresented groups or under-sampling overrepresented groups. Synthetic data generation creates new samples to balance the representation of protected attributes.

- **Natural Language Processing Models:** Approaches like counterfactual data augmentation and bias-specific data editing are discussed. Counterfactual data augmentation involves generating alternate versions of text data to ensure diverse representation, whereas bias-specific data editing manually or automatically adjusts text data to remove biased language or representation.

- **Information Retrieval Models:** Methods such as balancing the gold standard training data and negative sampling strategies are introduced to train bias-aware neural ranking models that present a fairer ranking list of documents to the users.

For each of these categories, we explain how balancing the training data concerning protected attributes should be performed to effectively reduce bias in the model. In addition, we discuss the challenges and limitations of these methods, including potential trade-offs in model performance and the complexities involved in accurately identifying and addressing biases.

By examining these diverse techniques, we aim to provide a comprehensive overview of data-driven de-biasing methods that can be applied across different types of models to create more equitable and fair outcomes.

5.1 Machine Learning Models

In addressing bias in ML models, several data-driven techniques are employed to ensure a more balanced and fair training process. This section explores the primary methods: re-weighting, re-sampling, and synthetic data generation. Each technique offers a unique approach to mitigating bias by adjusting the composition and significance of training data. These methods have been applied effectively in various domains, including Computer Vision and recommender systems, to enhance fairness and reduce bias.

5.1.1 Re-weighting

Re-weighting is a technique that modifies the importance assigned to different training samples based on their associated bias. This approach ensures that samples from underrepresented or marginalized groups are given more weight during the training process, while samples from overrepresented groups are given less weight. By adjusting the weights, the model is encouraged to learn from a more balanced perspective, thereby reducing the influence of biased data distributions.

For instance, in the context of recommender systems, re-weighting can adjust the importance of user interactions to ensure that recommendations are not biased towards the preferences of the majority group. Steck (2011) introduced a re-weighting method that adjusts the contribution of user ratings based on item popularity, addressing selection bias caused by the over-representation of popular items. The goal of this method is to mitigate the bias in recommendation accuracy that stems from the skewed popularity distribution in historical data. To derive the weighting factor, the author proposes that the probability of observing a rating is proportional to the power of the item's popularity, which follows a power-law distribution. This leads to a stratification weight calculated as the inverse of the observed relevant ratings raised to a power derived from the power-law exponent. This approach ensures that items with fewer ratings, typically less popular, are given appropriate consideration in the evaluation metrics.

In the training phase, the method is extended through modifications to the matrix factorization approach where the weights for observed ratings are adjusted according to their stratified popularity-based probability. The weight for each observed rating is adjusted inversely to the item's popularity, maintaining the overall balance by keeping the cumulative weight of all observed ratings constant. This re-weighting helps in producing a more balanced training dataset, improving the model's ability to recommend less popular items accurately.

5.1.2 Re-Sampling

Re-sampling techniques focus on altering the frequency of samples from different groups within the training dataset. There are two main approaches to re-sampling: over-sampling and under-sampling.

Over-Sampling: This involves duplicating samples from underrepresented groups to increase their presence in the training dataset. By doing so, the model is exposed to these groups more frequently, helping it to learn their characteristics better and reduce bias.

Under-Sampling: This involves reducing the number of samples from overrepresented groups. By selectively removing samples from these groups, the training dataset becomes more balanced, preventing the model from becoming overly biased towards the overrepresented data.

For instance, Buda *et al.* (2018) address the class imbalance problem in convolutional neural networks (CNNs) through oversampling and undersampling techniques. The study systematically investigates the impact of class imbalance on CNN classification performance and evaluates various methods to mitigate this issue using three benchmark datasets: MNIST, CIFAR-10, and ImageNet. To address the class imbalance, the study examines seven methods. The primary technique, *random minority oversampling*, involves replicating randomly selected samples from minority classes to balance the dataset. This method aims to equalize the number of samples in minority and majority classes. Another approach, random majority undersampling, randomly removes samples from majority classes to match the number of samples in minority classes, thereby reducing the dataset size but balancing class distribution.

Two-phase training is explored as pre-training on an oversampled or undersampled dataset followed by fine-tuning on the imbalanced dataset, maintaining the same hyperparameters with a slightly reduced learning rate during fine-tuning. Thresholding adjusts the decision threshold of classifier outputs to compensate for prior class probabilities, ensuring balanced classification decisions. The study also combines oversampling and undersampling with thresholding to further adjust class probabilities.

The study finds that oversampling is the most effective method for handling class imbalance in CNN training, significantly improving classification performance without causing overfitting. Oversampling should be applied to eliminate imbalance for optimal results, while the extent of undersampling should be tailored to the specific imbalance ratio. Thresholding, when applied to adjust class probabilities, proves beneficial, particularly when combined with oversampling, ensuring better overall accuracy. Both oversampling and undersampling contribute to more stable training, with oversampling providing the best results without the risk of overfitting.

In information retrieval systems, similar techniques can be applied to ensure that retrieval methods do not disproportionately favor content from dominant gender groups. Oversampling techniques can help in making sure that diverse content is represented in search results, while undersampling can prevent the model from being overly influenced by the majority of data, thus promoting a fairer and more balanced retrieval process. Combining these techniques with thresholding further refines the balance between different classes, contributing to a more equitable and effective IR system.

5.1.3 Synthetic Data Generation

Synthetic data generation is a technique used to create new samples that mimic the characteristics of the original data, thereby balancing the representation of protected attributes. This method involves generating artificial data points that resemble the underrepresented groups in the training dataset.

Deep Learning Approaches: Generative Adversarial Networks (GANs) are a type of neural network that can generate new data samples by learning the distribution of the original data. They are particularly effective in creating realistic synthetic data that can help balance the dataset. In Computer Vision, GANs have been used to generate synthetic images, enhancing the diversity of the training data (Cubuk *et al.*, 2018; Gurumurthy *et al.*, 2017; Lemley *et al.*, 2017; Osokin *et al.*, 2017; Perez and Wang, 2017; Yi *et al.*, 2019). Alongside GANs, neural style transfer and adversarial training are also employed to generate synthetic data. Neural style transfer manipulates the style of images while preserving their content, allowing the creation of diverse and visually varied training samples (Perez and Wang, 2017).

For instance, Perez and Wang (2017) employs Generative Adversarial Networks (GANs) for data augmentation to address biases and enhance image classification performance. The approach is multifaceted, encompassing traditional transformations and GAN-based methods to create a diverse and balanced dataset. The study utilizes CycleGAN, a specific GAN framework, to perform style transfers, generating synthetic images that augment the training data. The traditional data augmentation techniques involve applying affine transformations, such as rotation, scaling, flipping, and color adjustments. These methods generate variations of the existing images, effectively doubling the dataset size. The GAN-based data augmentation employs CycleGAN to generate new images by transferring styles from a set of predetermined styles (e.g., Cezanne, Monet, Van Gogh, Winter) to the original images. This technique allows for the creation of images in different styles while preserving the original content, thereby enhancing the diversity of the training dataset.

The results from the experiments demonstrate the effectiveness of GAN-based data augmentation. Traditional transformations significantly improved validation accuracy for both the dog vs. goldfish and dog vs. cat classification tasks. For instance, traditional augmentation increased validation accuracy from 85.5% to 89.0% for the dog vs. goldfish task. GAN-based augmentation using CycleGAN also enhanced performance, achieving a validation accuracy of 86.5% for the same task, a notable improvement over the baseline without augmentation. The paper introduces a novel neural augmentation method, where an augmentation network is trained alongside the classification network. This approach yielded the highest performance, with validation accuracy reaching 91.5% for the dog vs. goldfish task and 77.0% for the dog vs. cat task. This method outperformed both traditional and GAN-based augmentations, demonstrating the potential of learning augmentations directly from the data. Control experiments indicated that simply increasing the network's complexity without augmentation did not improve performance, underscoring the importance of effective data augmentation strategies. The study concludes that while traditional augmentation techniques are effective, GAN-based and neural augmentation methods provide significant benefits by generating realistic and varied synthetic data. These advanced techniques help mitigate biases, ensuring the training of robust deep-learning models and improving overall classification accuracy.

Generative models can be leveraged to generate synthetic data that better represent underrepresented gender groups in the training data. In the context of IR, a generative model can create diverse query-document pairs reflecting the perspectives and information needs of genders that are typically underrepresented. This approach can mitigate gender bias by ensuring that the IR models do not disproportionately favor content associated with the majority gender group, thereby promoting more equitable search results.

Data Augmentation: This involves creating new samples by applying various transformations to existing data. For example, in image data, transformations might include rotation, scaling, or color adjustment. In textual data, it might involve paraphrasing or introducing slight variations in the text. Data augmentation has been extensively used in Computer Vision to artificially increase the diversity of image datasets, thereby reducing bias (Mikołajczyk and Grochowski, 2018; Perez and Wang, 2017; Shorten and Khoshgoftaar, 2019). Shorten and Khoshgoftaar (2019) provided a comprehensive survey on image data augmentation techniques, highlighting their effectiveness in improving model performance and fairness. These techniques generate new samples by applying various transformations to existing data, thereby increasing the diversity and balance of the training dataset. Geometric transformations are a primary method used in this approach. These include rotation, where images are rotated at various angles to simulate different viewpoints, helping the model become invariant to the orientation of objects within the images. Scaling involves resizing images to different scales, allowing the model to handle variations in object sizes within the images. Flipping images horizontally or vertically creates mirror images that enhance the dataset's diversity. Cropping involves extracting smaller patches from the original images, which helps the model focus on different parts of the image and improves its robustness to changes in object positions. Color space transformations manipulate the color properties of images. These include color jittering, which adjusts the brightness, contrast, saturation, and hue of images. This helps the model become invariant to different lighting conditions and color variations. Histogram equalization modifies the intensity values in the color histograms to enhance image contrast and detail, particularly useful in images with poor lighting.

The impact of these traditional data augmentation techniques is evaluated through experiments on various datasets. For example, rotating images by small angles can significantly improve classification performance by making the model more robust to rotational variations. Similarly, cropping and flipping images can help the model generalize better by exposing it to a wider range of possible object positions and orientations. The findings from the paper indicate that traditional data augmentation techniques are effective in enhancing the performance of deep learning models. By artificially inflating the size of the training dataset and introducing a diverse set of variations, these techniques help reduce the risk of overfitting and improve the model's generalization ability. This is particularly beneficial in scenarios with limited data, where creating a large and varied training dataset from a small number of samples is crucial for training robust models.

In addressing gender bias in IR, data augmentation techniques can be employed to increase the representation of gender-specific queries and content. By generating paraphrases or introducing variations in documents and queries that reflect different gender perspectives, the model can learn to recognize and fairly represent the interests and information needs of all genders. This helps reduce bias in search results, ensuring that the system is more inclusive and responsive to diverse user queries.

Synthetic Minority Over-sampling Technique (SMOTE): SMOTE is a popular technique that generates synthetic samples for underrepresented classes by interpolating between existing minority class samples. This method creates new samples that lie along the line segments joining any or all of the k-nearest neighbors of the minority class. For instance, Douzas *et al.* (2018) proposed an improved oversampling technique based on k-means clustering and SMOTE to address imbalances and enhance classification performance in datasets with skewed class distributions. Unlike other methods that rely solely on class labels, the k-means clustering algorithm is applied to the entire dataset without considering class labels, enabling the discovery of natural groupings within the data. This unsupervised clustering approach helps to identify regions of the input space that are “safe” for oversampling. Following clustering, the method filters the clusters to select those dominated by minority class instances. Clusters are considered suitable for oversampling if their proportion of minority samples exceeds a predefined imbalance ratio threshold. This filtering step avoids oversampling in regions with significant class overlap, thereby reducing the risk of noise generation.

In the final step, SMOTE is applied within each selected cluster to generate synthetic samples. The number of synthetic samples generated in each cluster is proportional to its minority sample sparsity, with sparse clusters receiving more samples than dense ones. This ensures that within-class imbalances are addressed effectively, with synthetic samples distributed in a manner that balances sparsely populated minority regions without redundant data generation. By targeting underrepresented areas, the method addresses both between-class and within-class imbalances, which are common challenges in imbalanced learning tasks.

5.2 Natural Language Processing models

In the field of natural language processing, many studies have employed Counterfactual Data Augmentation (CDA) and Counterfactual Data Substitution (CDS) to address gender bias in both algorithmic and representational aspects of pre-trained language models.

5.2.1 Counterfactual Data Techniques: CDA and CDS

Counterfactual data techniques are employed to address gender bias by manipulating the training data to balance gender representation. *Counterfactual Data augmentation* involves creating a duplicate corpus where gendered terms are swapped. For instance, the sentence ’the woman cleaned the kitchen’ is altered to ’the man cleaned the kitchen,’ and both versions are used for training word embeddings. This approach neutralizes gender biases by ensuring equal representation of gendered terms.

However, corpus duplication can lead to unnatural statistical properties. To address this, *Counterfactual Data Substitution* replaces gendered terms probabilistically in the original corpus rather than duplicating it. This method preserves the statistical properties of natural text while maintaining grammaticality and discourse coherence. For example, in the original corpus with the sentences 'The woman cleaned the kitchen. The woman is a nurse,' CDS might result in 'The man cleaning the kitchen. The woman is a nurse.' This probabilistic substitution replaces only one instance of 'woman' with 'man,' preserving the text's naturalness while addressing gender bias.

5.2.2 Applications and Evaluations of CDA and CDS

The application of CDA and CDS in various studies highlights their effectiveness in mitigating gender bias. Zhao *et al.* (2019) conducted a comprehensive analysis on the ELMo model, revealing a disproportionate representation of male entities compared to female entities. Their study demonstrated that ELMo embeddings exhibited gender biases, particularly in the coreference resolution task. To mitigate this bias, the researchers used data augmentation techniques with the OntoNotes 5.0 (Weischedel *et al.*, 2012) dataset. They swapped gender-revealing entities such as 'he,' 'him,' 'she,' and 'her' with their opposite counterparts while maintaining the original syntactic structure of the sentences. This process created a gender-swapped variant of the corpus, effectively doubling its size and balancing gender representation.

The augmented dataset was used to retrain the coreference resolution system, which initially exhibited significant gender bias. The system's performance was evaluated using the WinoBias probing corpus, designed to test coreference resolution capabilities in both stereotypical and anti-stereotypical contexts. Before applying CDA, the ELMo-based coreference system showed a notable disparity in accuracy between pro-stereotypical and anti-stereotypical predictions, indicating a strong gender bias. After retraining with the augmented dataset, the results showed a substantial reduction in this bias. The difference in performance between pro-stereotypical and anti-stereotypical contexts was reduced to insignificant levels, demonstrating that CDA effectively mitigated gender bias. This improvement was observed across both the *Semantics Only* and *w/ Syntactic Cues* subsets of the WinoBias dataset, highlighting the robustness of the CDA approach in addressing different types of coreference resolution challenges.

Dinan *et al.* (2019) explored the application of counterfactual data augmentation to mitigate gender bias in dialogue generation tasks. Their research initially demonstrated that gender biases not only existed but were amplified in dialogue generation across six different datasets. To quantify this bias, the researchers counted the number of gendered terms within the datasets and reported the percentage of male bias, finding that male bias reached 73% across the six datasets.

To address this bias, they applied counterfactual data augmentation, which involved duplicating every dialogue containing gendered words and swapping these words using a predefined list provided by Zhao *et al.* (2018c). This list included a comprehensive set of gendered word pairs, ensuring systematic and consistent replacements across the dataset. The primary goal of CDA was to balance gender references automatically without altering the contextual meaning of the dialogues.

The effectiveness of CDA was evaluated using several metrics to measure gender bias and the quality of the generated dialogues. These metrics included the percentage of gendered words in the generated utterances, the percentage of male bias among all gendered words generated, and the F1 score, which measures the overlap between the generated and the gold standard responses. The evaluation involved comparing the CDA-augmented model's performance with the baseline model trained on the original unmodified dataset. Results indicated that applying CDA led to a significant reduction in gender bias compared to the baseline model. Specifically, the CDA-augmented model generated dialogues with a lower percentage of gendered words and a more balanced distribution of male and female gendered words. This reduction in gender bias was observed across different evaluation bins, categorized based on the presence of gendered words. The CDA model showed improvements in controlling the genderedness of the language generated, although it did not completely eliminate bias.

In a comprehensive study by Maudslay *et al.* (2019), the authors targeted the mitigation of gender bias in word embeddings using both

CDA and CDS techniques by conducting a comprehensive empirical comparison using extensive corpora, namely the English Gigaword and Wikipedia, thus offering a detailed evaluation of the efficacy of these methods across different textual datasets. The study also introduces a novel Names Intervention technique designed to specifically address biases associated with first names. The Names Intervention technique specifically addresses biases inherent in first names by using a bipartite graph matching strategy based on name frequency and gender-specificity to create name pairs for substitution, ensuring balanced gender representation without affecting grammaticality. Methodological improvements include CDA with Grammar Intervention (gCDA), which enhances CDA by using coreference information to avoid swapping gendered terms when they refer to proper nouns, maintaining grammaticality. CDA with Names Intervention (nCDA) applies the name-pairing strategy to extend CDA, treating biases in first names. CDS with Grammar Intervention (gCDS) combines CDS with grammar intervention to avoid ungrammatical substitutions, while CDS with Names Intervention (nCDS) combines CDS with the name-pairing strategy for a more comprehensive bias treatment.

To evaluate the impact of the proposed de-biasing methods, the authors evaluate them on the English Gigaword and Wikipedia corpora, comparing their effectiveness in reducing direct and indirect gender biases, maintaining the quality of word embeddings, and improving non-biased gender analogies. Direct bias is measured using the Word Embedding Association Test (WEAT). Results show that the Names Intervention variants (nCDA and nCDS) outperform other methods in reducing direct bias, with nCDS showing significant improvements over traditional CDA and grammar-based interventions. For example, in the careers-family test, nCDS reduces bias more effectively than other CDA/S variants and even Word Embedding Debiasing (WED) methods. Indirect bias is evaluated through clustering and reclassification tests. Results indicate that nCDA and nCDS significantly lower the purity of biased word clusters, suggesting a more thorough mitigation of indirect bias. For instance, nCDS achieves a 58% reduction in cluster purity on the Gigaword corpus and a 39% reduction on Wikipedia. The reclassification test supports these findings, with nCDS showing the lowest reclassification accuracy of previously biased words, implying less residual bias information.

The quality of the embeddings is assessed using the SimLex-999 word similarity dataset and a sentiment classification task. Word similarity scores indicate minimal reduction in quality across all methods, with WED methods slightly outperforming unmitigated embeddings. In sentiment classification, nCDA and nCDS maintain competitive performance, demonstrating their effectiveness in preserving embedding utility. For example, WED70, nCDA, and nCDS achieve the best sentiment classification results with minimal error rates. The non-biased gender analogy task measures how well embeddings can draw appropriate gender analogies without bias. Results show that CDA and CDS methods, especially those with the Names Intervention, outperform WED methods significantly. For instance, nCDS shows a substantially lower error rate for non-biased gender analogies on Gigaword compared to WED variants, indicating better retention of meaningful gender information while mitigating bias.

5.3 Information Retrieval Models

In this section, we introduce methods designed to train bias-aware neural ranking models that present fairer ranking lists of documents to users. These methods focus on two primary strategies: (1) balancing the gold standard training data and (2) employing a bias-aware negative sampling technique. Together, these methods aim to address both algorithmic and representational biases in neural ranking models, ultimately enhancing the fairness of the search results presented to users. We will now explore each of these strategies, exploring their implementation and impact on model performance.

5.3.1 Balancing the Gold Standard Training Data

Balancing the gold standard training data involves adjusting the training dataset to ensure a fair representation of different gender groups. This can be achieved by re-weighting, re-sampling, or augmenting data, thereby providing a more equitable basis for model training. By addressing imbalances in the representation of data from different gender groups, this strategy helps to reduce the biases transferred to the neural ranking model.

In Section 4.3.2, it was discussed that the gold standard dataset used to train neural-ranking models contains prevalent psychological characteristic biases in documents associated with male and female queries. These biases can exacerbate psychological characteristic biases during neural ranking model training. To investigate how neural ranking models learn such biases and compare the original training dataset with a balanced version, Bigdeli *et al.* (2023) proposed a systematic approach for balancing training data used in neural-based retrieval models. Their methodology aims to mitigate gender biases in psychological characteristics present in ranked documents corresponding to different gender identities perceived by the neural ranking model. The authors' data balancing technique involves augmenting query-document pairs across diverse gender identities to ensure that documents associated with each identity exhibit comparable psychological characteristics. This methodology encompasses query generation, gender classification, data balancing, model training, and evaluation.

Initially, query-document pairs are generated using a fine-tuned text-to-text transformer model on the MS MARCO (Nguyen *et al.*, 2016) dataset to ensure diversity and relevance. For each document in the collection, a synthetic query resembling a user search query is generated. Subsequently, a BERT model fine-tuned on annotated data with gender labels classifies each query into male or female gender affiliations. This classification step is crucial for later stages where datasets are balanced based on gender affiliations.

To effectively balance the training data, the Linguistic Inquiry and Word Count (LIWC) toolkit (Pennebaker *et al.*, 2001a) is used to analyze documents for various psychological processes, including affective and cognitive aspects. Each document is represented as a vector of these psychological characteristics, ensuring a nuanced understanding and representation of textual content. The process of creating balanced datasets involves pairing queries from different gender affiliations based on the similarity of their associated document vectors. Cosine similarity is used as a distance metric between vectors to quantitatively measure similarity, guiding the construction of balanced training datasets. Consequently, for each query-document pair with a specific gender affiliation, a counterpart pair with the opposite gender affiliation is matched, ensuring differences in gender but similarity in psychological characteristics of associated documents. These balanced query-document pairs are then augmented with the original training dataset to construct a debiased dataset for training neural ranking models, which includes both the original MS MARCO dataset and the newly developed debiased datasets for comparative purposes. Finally, the BERT-base-uncased model is fine-tuned using the cross-encoder architecture on both the original and debiased datasets consisting of balanced query-document pairs sharing similar levels of psychological characteristics.

To evaluate gender biases in terms of psychological characteristics, neural ranking models trained on the original and debiased datasets rank documents for a set of male and female queries. The differences in psychological characteristics among the top-10 ranked documents for female and male queries are compared. The results indicate that models trained on debiased datasets show a slight decrease in retrieval effectiveness (e.g., MRR@10) at higher debiasing ratios; however, this trade-off is minimal compared to the benefits of bias reduction. Training on debiased datasets significantly reduces differences in psychological characteristics between documents retrieved for gendered queries and also decreases bias in retrieved documents for neutral queries. Therefore, the proposed debiasing method effectively reduces bias while maintaining retrieval effectiveness.

5.3.2 Bias-Aware Negative Sampling Technique

Another effective data-driven approach to mitigate bias in neural ranking models is employing a targeted negative sampling strategy rather than a random selection of irrelevant documents during training. This involves selectively including or excluding certain data points during training to counteract biases. By carefully choosing which negative examples (i.e., non-relevant documents) to include in the training process, models can learn to produce rankings that are less biased and more representative of diverse user needs. In neural ranking, particularly in cross-encoder architectures, the objective is to minimize the distance between relevant query-document pairs while maximizing the distance between queries and irrelevant documents. This approach is crucial for optimizing the model's performance, as highlighted in studies such as RocketQA (Qu *et al.*, 2021) and Dense Retrieval (Karpukhin *et al.*, 2020).

The method proposed by Bigdeli *et al.* (2022) emphasizes the importance of including not only irrelevant but also biased documents in the training dataset. By quantifying the level of bias within retrieved documents for each query, the strategy selects a subset of highly biased documents as negative samples. This selective negative sampling strategy aims to train the model to discern and avoid documents that exhibit significant bias towards specific gender identities during the ranking process.

This negative sampling approach ensures that the neural ranker learns not only to distinguish relevant from irrelevant documents but also to mitigate gender biases effectively. Exposing the model to biased documents as negatives encourages the neural network to develop a nuanced understanding of what constitutes bias in document retrieval tasks. This training process ultimately contributes to the creation of fairer and more equitable information retrieval systems.

The results obtained from training neural ranking models using different language model architectures such as BERT (Devlin *et al.*, 2018a), DistilRoBERTa (Sanh *et al.*, 2019), and Electra (Clark *et al.*, 2020a) demonstrate that when these models are trained on datasets enriched with bias-aware training examples—comprising not only irrelevant but also biased documents—they exhibit improved capability in generating fairer ranked lists of documents. To evaluate their performance, each model trained on both the original and bias-aware datasets was assessed using query sets designed to include neutral and socially problematic queries from $MSMARCO_{FAIR}$ (Rekabsaz *et al.*, 2021) and Gendered Queries (Rekabsaz and Schedl, 2020) (both described in Section 3.3.1). Subsequently, metrics such as $ARaB$ (Rekabsaz and Schedl, 2020) and NFaiRR (Rekabsaz *et al.*, 2021) were employed to quantify the level of bias and fairness within the ranked document lists generated by these models.

The findings indicate that models trained on bias-aware datasets, where biased documents are explicitly represented as negative samples during training, are capable of significantly enhancing the fairness of document rankings. Specifically, these models elevate the positions of fairer documents within the top-10 rankings presented to users. For example, consider the query 'How important is a governor?' Originally, a biased document depicting governors as predominantly male-oriented occupations might have been ranked second. However, through training on a bias-aware dataset, the same biased document now appears much lower in the rankings, specifically in the 88th position, showcasing the efficacy of the bias mitigation strategy employed by the model.

In conclusion, the integration of bias-aware training methodologies into neural ranking model training not only improves the model's ability to discern relevant from irrelevant documents but also enhances its capacity to reduce biased representations in search results. This approach represents a significant advancement towards achieving fairness and inclusivity in information retrieval systems, ensuring that users receive more equitable and unbiased access to information across diverse query scenarios.

6

De-biasing of Neural Embeddings

Neural embeddings have shown significant improvements in information retrieval systems by using embedded representations of queries and documents. However, these embeddings can mirror and amplify the gender biases captured in the content representations mainly from their training data (Prost *et al.*, 2019). The primary focus of this chapter is to show how these embedded representations can carry on biases that will implicitly or explicitly make the retrieved results biased (Bigdeli *et al.*, 2021a; Bordia and Bowman, 2019; Fabris *et al.*, 2020; Rekabsaz *et al.*, 2021). We discuss how biases in training data, such as those found in web content, can find their way into embeddings (Basta *et al.*, 2019), and place particular emphasis on how neural embeddings capture and reflect these biases (Bolukbasi *et al.*, 2016).

Acknowledging and recognizing these biases allow us to explore strategies to reduce bias in representations through both data-centric methods and algorithmic-centric approaches before and after training the embeddings. By employing these techniques, it is possible to counteract the biases inherent in the training data, thus producing fairer and more accurate embeddings. Later on, in Section 6.3, we explore the differences between the biases inherent in different kinds of embeddings, including static versus dynamic and contextualized ones.

6.1 Pre-training Debiasing Strategies

The process of debiasing neural embedding representations can be implemented either prior to or following the training of neural embeddings. In this context, we examine techniques that address debiasing before initiating the model's training, specifically focusing on pre-training debiasing strategies. Pre-training debiasing strategies involve training neural embeddings from scratch with the goal of bias reduction in the embedding representations. This category of strategies can be advantageous because they address biases at the source, potentially leading to more fundamentally fair embeddings. However, they also come with disadvantages, such as the significant computational resources required and the need for extensive data preprocessing and retraining.

In an interesting study (Caliskan *et al.*, 2022), the authors find that a significant majority of the most frequent words in the training data of the widely used embeddings are associated with men. For example, 77% of the top 1,000 most frequent words in the GloVe (Pennington *et al.*, 2014) embeddings are associated with men. This pattern holds true across different frequency ranges, with similar trends observed in fastText embeddings (Bojanowski *et al.*, 2017), albeit to a slightly lesser extent. In addition, their part-of-speech tagging analysis showed that male-associated words are more likely to be verbs, reflecting stereotypes of men as active and agentic. Female-associated words, on the other hand, are more likely to be adjectives and adverbs, suggesting a perception of women that requires additional description or explanation. Also, clustering analysis reveals that male-associated words often relate to domains such as big tech, engineering, sports, and violence. In contrast, female-associated words frequently pertain to appearance, sexual content, and kitchen-related terms. The unequal representation of both genders within the training data of neural embeddings can cause the model to develop biased representations of men and women.

One notable impact of pre-training on large pre-trained language models (PLMs) is how the training data influences gender biases. For example, the training corpus for ELMo (Peters *et al.*, 2018b), the One

Billion Word Benchmark, contains a significant gender skew: male pronouns (e.g., "he" "his" and "him") occur three times more frequently than female pronouns (e.g., "she" "her") (Zhao *et al.*, 2019). Specifically, the dataset shows approximately 5.3 million occurrences of male pronouns compared to 1.6 million occurrences of female pronouns. This imbalance not only reflects societal biases but also propagates these biases into the trained embeddings, leading to a higher likelihood of male-biased associations in downstream tasks. Moreover, male pronouns co-occur more frequently with occupation words, regardless of whether those occupations are stereotypically male or female. For instance, in the training corpus for ELMo, male pronouns co-occur with occupation words 170,000 times, whereas female pronouns co-occur with occupation words only 36,000 times (Zhao *et al.*, 2019). These statistics underscore the need for balanced training datasets and pre-training debiasing strategies to mitigate inherent gender biases and ensure fairer and more accurate embeddings.

More recently, (Kotek *et al.*, 2023) transitioned from studying pre-trained language models to investigating large language models and how they perpetuate gender stereotypes, particularly in the context of occupational roles. They found that LLMs are 3-6 times more likely to choose occupations that stereotypically align with a person's gender. These choices align more closely with societal perceptions than with official job statistics, indicating that LLMs amplify existing biases beyond what is reflected in reality. Additionally, the study revealed that the behaviour of LLMs correlates more closely with human judgments about gender stereotypes than with actual labour statistics. This suggests that the training data for these models reflect societal biases rather than objective realities.

In the following, we explain two common pre-training debiasing strategies namely through data augmentation and masking gender indicators.

6.1.1 Debiasing through Data Augmentation

This method (Zhao *et al.*, 2019) involves augmenting the training corpus with gender-swapped variants of sentences i.e., creating a parallel corpus where gender-specific words are swapped, ensuring an equal representation of male and female entities in the training data. For example, every instance of ‘he’ is swapped with the opposite gender version i.e., ‘she’ and vice versa. This ensures that the model is exposed to a balanced representation of gendered entities during training. Data augmentation has been shown to significantly reduce bias in downstream tasks such as coreference resolution.

By training on both the original and gender-swapped datasets, the model learns to treat gendered terms more equitably. This approach leverages the inherent diversity in the augmented data to mitigate gender bias that often exists due to the imbalance in the frequency of gendered terms in the original training data. Specifically, the augmented corpus helps the model understand that gendered terms are interchangeable in many contexts, which reduces the tendency to associate certain roles or actions exclusively with a particular gender. In addition to reducing bias, this method also contributes to more robust and generalizable models. By exposing the model to a wider variety of contexts in which gendered terms appear, it can better handle real-world scenarios where the gender distribution may not match the skewed distributions often found in training data. This leads to improvements not only in fairness but also in the overall quality and reliability of the model’s predictions.

6.1.2 Debiasing through Masking Gender indicators

Debiasing neural embeddings through masking gender indicators is a method (Prost *et al.*, 2019) aimed at reducing gender bias in text classification tasks by explicitly removing gender-specific terms from the training data. The first step involves identifying and listing gender-specific terms such as pronouns titles, and other gendered words that are likely to introduce bias into the embeddings. Once the gender-specific terms are identified, they are systematically removed or replaced in the training data. This process, known as ‘scrubbing’ ensures that these terms do not influence the embeddings. The objective is to prevent the model from learning gender associations that could lead to biased outcomes in downstream tasks. The embeddings are then trained on this modified corpus. By excluding gender-specific terms, the resulting embeddings are expected to be less biased. While the approach is straightforward and can be easily implemented without requiring complex algorithms or extensive computational resources, complete masking of gender indicators might lead to a loss of contextual information that is necessary for certain tasks. For example, gender-specific terms might carry important semantic information in contexts like biographies or medical records.

6.1.3 Debiasing Through Loss Regularization

In (Bordia and Bowman, 2019), the authors propose a debiasing approach through regularizing the loss function for training the embeddings. In particular, they proposed a regularization loss term for the language model that minimizes the projection of encoder-trained embeddings onto an embedding subspace that encodes gender. This method aims to reduce the influence of gender bias in the learned embeddings. The regularization method is effective in reducing gender bias up to an optimal weight assigned to the loss term.

The first step involves identifying a subspace in the word embedding space that captures gender information. This is typically done by defining a set of gendered word pairs (e.g., he-she, man-woman, king-queen). The principal components of their difference vectors are then calculated. The principal component that captures the most variance among these difference vectors is considered the gender direction. This is a common practice to find a sensitive attribute common space (Bolukbasi *et al.*, 2016; Gonen and Goldberg, 2019; Zhao *et al.*, 2018a; Zhao *et al.*, 2019).

The core of the debiasing approach happens in the second step where there is the addition of a regularization term to the loss function of the language model. This term is designed to minimize the projection of word embeddings onto the identified gender subspace. Mathematically, for an embedding w and the gender direction g, the projection of w onto g is $(w \cdot g)g$. The regularization loss term L_{reg} is then given by:

$$L_{reg} = \lambda \sum_{w \in V} ||(w \cdot g)g||^2 \tag{6.1}$$

, where λ is a hyperparameter that controls the strength of the regularization and V, is the vocabulary.

During training, the standard language modelling loss (e.g., cross-entropy loss) is augmented with the regularization loss. The total loss is:

$$L = L_{lm} + L_{reg} \tag{6.2}$$

where L_{lm} is the language modeling loss. By minimizing this combined loss, the language model learns to produce embeddings that are less influenced by gender bias, as the regularization term penalizes embeddings that align strongly with the gender subspace.

One of the advantages of this method is that by adjusting the strength of the regularization term with the hyperparameter λ, the method can balance between reducing bias and maintaining the performance of the language model. However, finding the optimal λ is crucial. It should also be noted that although the method reduces bias, it may not eliminate it, as some gendered information may still be encoded in other dimensions of the embeddings.

6.1.4 Debiasing through Gender Neutralization

This debiasing approach (Zhao *et al.*, 2018b) is focused on training debiased word embeddings from scratch by modifying existing word vectors. The authors proposed Gender-Neutral Global Vectors by altering the loss function of the GloVe (Pennington *et al.*, 2014) model to concentrate most of the gender information in the last coordinate of each vector. The steps are as follows:

1. **Word Co-occurrence Matrix**: Following the GloVe methodology, construct a word-to-word co-occurrence matrix X, where $X_{i,j}$ denotes the frequency of the j-th word appearing in the context of the i-th word. The embeddings of a center word w and a context word $\tilde{w}$ are represented as $w, \tilde{w} \in \mathbb{R}^d$, where d is the dimension of the embeddings.

2. **Decomposition of Word Vectors**: Each word vector w is decomposed into two parts: a neutral component $w^{(a)}$ and a

gendered component $w^{(g)}$:

$$w = [w^{(a)}; w^{(g)}] \tag{6.3}$$

where $w^{(a)} \in \mathbb{R}^{d-k}$ and $w^{(g)} \in \mathbb{R}^{k}$, with k being the number of dimensions reserved for gender information.

3. **Objective Function**: The optimization objective J combines three components of word proximity J_G, gender definition alignment J_D and Gender Neutrality component J_E.

   J_G captures word proximity following the original GloVe objective:

$$J_G = \sum_{i,j=1}^{V} f(X_{i,j}) \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{i,j} \right)^2 \tag{6.4}$$

   Where $f(X_{i,j})$ is a weighting function to reduce the influence of high-frequency co-occurrences, and $b_i, \tilde{b}_j$ are bias terms.

   J_D aligns gendered dimensions for gender-definition words:

$$J_D = - \left\| \sum_{w \in \Omega_M} w^{(g)} - \sum_{w \in \Omega_F} w^{(g)} \right\|_1 \tag{6.5}$$

   This term ensures that male-definition words (Ω_M) and female-definition words (Ω_F) are aligned in the gendered component.

   Lastly, J_E ensures gender-neutral words $w \in \Omega_N$ are neutral in the embedding space.

$$J_E = \sum_{w \in \Omega_N} \left(v_g^T w^{(a)} \right)^2 \tag{6.6}$$

   v_g is the gender direction i.e., it is a vector that represents the gender subspace within the embedding space. It is estimated by averaging the differences between the embeddings of male-definition words and female-definition words. Specifically, v_g is computed as follows:

$$v_g = \frac{1}{|\Omega'|} \sum_{(w_m, w_f) \in \Omega'} (w_m^{(a)} - w_f^{(a)}) \tag{6.7}$$

Where Ω' is a set of predefined gender word pairs and $w_m^{(a)}$ and $w_f^{(a)}$ are the neutral components of the embeddings for the male and female words in the pair, respectively. This term ensures that the neutral component $w^{(a)}$ of gender-neutral words (Ω_N) remains in the null space of the gender direction v_g.

Finally, we combine the three loss components as

$$J = J_G + \lambda_d J_D + \lambda_e J_E \tag{6.8}$$

where λ_d and λ_e are hyperparameters controlling the importance of each component.

During the training of the GN-GloVe model, v_g is used to ensure that the neutral component $w^{(a)}$ of gender-neutral words remains orthogonal to this gender direction, thereby reducing gender bias.

This allows for the use of word representations that exclude the gender coordinate. They achieve this by using two groups of male/female seed words and encouraging words from different groups to differ in their last coordinate. Additionally, they ensure that the representation of gender-neutral words (excluding the last coordinate) is orthogonal to the gender direction. We further optimize the combined objective using SGD. The gender direction v_g is estimated by averaging the differences between male and female word pairs and is kept fixed during each epoch. The GN-GloVe embeddings are evaluated on their ability to isolate gender information while preserving word proximity. The debiased Glove embeddings (GN-GloVe embeddings) improve performance in downstream tasks, such as coreference resolution, by reducing gender bias. This approach attempts to eliminate bias during training rather than in post-processing, which is considered more effective. However, the bias is not entirely removed but hidden; the underlying associations in the embedding space remain biased.

6.2 Post-training Debiasing Strategies:

6.2.1 Hard debiasing

Hard debiasing (Bolukbasi *et al.*, 2016) is a technique aimed at ensuring that gender-neutral words do not exist in the gender subspace and remain at an equal distance from equality pairs like ‘he-she’ (Bolukbasi *et al.*, 2016). This process involves subtracting the projection of the embedding on the bias direction from the vector. The key steps in hard debiasing are:

1. **Identify Gender Subspace**: This involves defining a gender direction based on the principal component of gender pair differences e.g., ‘she-he’ (See Section 3.2.2). In other words, for each gender pair (u_i, v_i) we compute the difference vector $d_i = u_i - v_i$. Further, we stack these difference vectors and perform PCA on them to find the principal components that capture the most variance. The top principal component is typically considered as the gender direction g.

2. **Project and Neutralize**: The embedding of gender-neutral words is projected onto the gender direction. For each word embedding w, we project it onto the identified gender direction g to obtain $w_g = (w.g)$. This projection captures the gender-specific information in the embedding. This projection is then subtracted from the original vector to neutralize the bias. In other words, we subtract the gender projection from the original embedding to neutralize it and obtain $w' = w - w_g$. This ensures that the gender-neutral words do not have any component in the gender direction.

By only modifying the gender-neutral words, the method preserves the useful properties of the embeddings, such as semantic relationships and analogy-solving capabilities. However, the hard debiasing approach may remove certain distinctions that are valuable in specific applications which might cause the meaning of some phrases which rely on gender-specific meanings to be lost. In addition, hard debiasing might not eliminate all forms of bias. Some biases might persist in other dimensions of the embeddings.

6.2.2 Soft Debiasing

Soft debiasing (Bolukbasi *et al.*, 2016) recognizes that sometimes gender-specific terms contain more meaning that is required to be captured, and a complete neutralization might not be desirable. Instead, it aims to 'soften' the effect of gender bias on the embeddings based on a parameter λ which controls the extent of debiasing, allowing for a partial reduction of bias rather than complete neutralization. The steps include:

1. Identify Gender Subspace: Similar to hard debiasing (step 1 in Section 6.2.1), the gender direction is identified.

2. Adjust Based on Parameter: Instead of fully neutralizing, the embeddings are adjusted only to the extent defined by λ. If $\lambda = 0$, it is essentially the same as hard debiasing, while higher values of λ allow more of the original gender associations to remain (Bolukbasi *et al.*, 2016). In other words $w = w - \lambda(w.g)$.

Soft debiasing strikes a balance between reducing bias and maintaining some level of gender-specific information that might be contextually important. For instance, gendered nuances in certain professional titles or roles can be preserved to reflect realistic and meaningful distinctions. By only partially reducing bias, the method ensures that useful properties of the embeddings, such as semantic relationships, are largely preserved.

In general, while hard debiasing focuses on complete neutralization, soft debiasing allows for a more controlled adjustment.

6.3 Debiasing in Static vs Dynamic Embeddings

This section examines the differences between debiasing approaches for static and dynamic embeddings.

Static embeddings, such as Word2Vec and GloVe, represent words with fixed vectors regardless of their context (Mikolov *et al.*, 2013; Pennington *et al.*, 2014). These embeddings capture biases present in the training data, leading to associations that reflect societal stereotypes. For example, words like 'engineer' might be more closely associated with male pronouns, while 'nurse' might be more closely associated with female pronouns. On the other hand, dynamic embeddings (or contextualized embeddings), such as ELMo (Peters *et al.*, 2018a), BERT (Devlin *et al.*, 2018a), and GPT (Brown *et al.*, 2020), generate word vectors that change depending on the context in which the words appear. These embeddings capture more nuanced meanings but also encode more complex forms of bias because they consider a broader context.

The primary distinction between debiasing static and dynamic embeddings lies in their handling of context. Static Embeddings are easier to debias using straightforward projection techniques (soft or hard debiasing); well-suited for tasks where context is less critical. Static embeddings require simpler debiasing techniques that focus on the inherent associations within the word vectors. In contrast, dynamic embeddings necessitate more sophisticated approaches that consider the context-dependent nature of the word representations. They are also complex to debias and require advanced techniques like data augmentation and contextual neutralization.

Studies have shown that contextualized word embeddings are generally less biased than static ones, *even when the latter are debiased.* This is demonstrated through several measures. For example, it has been shown that for contextualized embeddings like ELMo, PCA reveals that the embeddings encode gender information in two principal components, whereas static embeddings like GloVe typically show only one principal component for gender. This indicates that contextualized embeddings capture more nuanced gender information (Zhao *et al.*, 2019). In addition, contextualized embeddings tend to show a lower direct bias (how closely certain words align with gender vectors) compared to static embeddings (Basta *et al.*, 2019). Contextualized embeddings also generalize gender bias less effectively than static embeddings. Classification tasks using contextualized embeddings achieve lower accuracy in predicting the gender of occupations compared to static embeddings, indicating less bias (Basta *et al.*, 2019).

7

Method-Level De-biasing

In this chapter, our attention is directed towards strategies for reducing gender biases in information retrieval systems during their training process. Three primary methods have been utilized for this purpose: Loss Function Regularization, Adversarial Training, and Query Reformulation. Loss Function Regularization entails introducing a bias-regularizer term into the neural ranker's loss function. Adversarial training is employed to eliminate gender-related attributes from the intermediate representation of query-document pairs through a mini-max game, and query reformulation focuses on modifying the query such that the retrieved documents for the reformulated query are less biases. In the following sections, we are going to explain each of these methods in more detail.

7.1 Preliminaries

To lay the groundwork, let's formally define an IR system. An IR system operates with a set of queries, denoted as $Q = \{q_1, q_2, ..., q_n\}$, and a pool of documents represented as $D = \{d_1, d_2, ..., d_m\}$. The goal is to retrieve the top-n most related documents for each of the queries in Q, from D.

Let us assume there is a function Φ, which is able to calculate a relevance score s between query q_i, and the documents in the corpus D, such that:

$$s_{ij} = \Phi(q_i, d_j) \tag{7.1}$$

For each query q_i, the documents are sorted based on their relevance score s_{ij} to the query. The same process is done for all of the queries in Q, and, the outcome is a ranked list R including the queries and their top-n most relevant documents.

It is important to note that the retrieval process may involve a re-ranking stage. This stage employs a more powerful, albeit computationally expensive, model to re-rank the top-n retrieved documents for each query. This additional step ensures a more accurate final ranked list. Given the computational cost of the re-ranker, a practical approach is adopted. Instead of using the computationally intensive model for the entire document pool, a two-step process is implemented.

- Initial Retrieval: A lighter retriever, possibly based on the exact matching of words in the query and documents, is employed to quickly retrieve the top-n relevant documents from the entire pool, which may consist of millions of documents. BM25 (Robertson, Zaragoza, *et al.*, 2009) is amongst the most widely employed first-stage retrievals for re-ranking tasks.

- Subsequent Re-ranking: The top-n retrieved documents are then subjected to the more computationally expensive re-ranker model– usually a neural ranker–, which fine-tunes the ranking, enhancing the precision of the final list by considering deeper contextual and semantic relationships.

This strategic two-step process strikes a balance between computational efficiency and result accuracy, making it feasible to handle large document pools.

Neural rankers employ deep neural networks as a function to predict the relevance score s_{ij} between the quey q_i, and the document d_j. There are two general architectures for the neural rankers, which are widely adopted; bi-encoder architecture, and cross-encoder architecture. We are going to explain the detailed architecture and training strategies in the following sections.

A bi-encoder architecture includes two separate networks (transformers), namely, T_1, and T_2, one responsible for encoding the query, and the other for encoding the document.

$$E_q = T_1(q_i), E_d = T_2(d_j), \tag{7.2}$$

where E_q is the vector representation of the quey q_i, and E_d is the vector representation of the document d_j. Then, to calculate the relevance score between the query, and the document, cosine similarity between their vector representations is employed.

$$s_{ij} = cos(E_q, E_d) \tag{7.3}$$

In cross-encoder architecture, query and document are concatenated, and the concatenated text is fed to the network (transformer), and a joint vector representation E is generated for the combination.

$$E = T(q_i \oplus d_j), \tag{7.4}$$

where $\oplus$ is a sign for concatenation. The vector representation E, is then fed to a Multi Layer Perceptron (MLP) so the relevance score s_{ij} between the query q_i, and the document d_j is calculated.

$$s_{ij} = \sigma(WE + B), \tag{7.5}$$

where σ is an activation function, and W, and B are weight and bias matrices of the linear layer.

Training Strategies. For training the neural rankers, contrastive training is employed. For each query, we have some relevant and some irrelevant documents. there are two common strategies to train the model to be able to discern relevant and irrelevant documents for the queries.

1. **Point-wise**: In this strategy, each document is processed individually, and is going to be classified as being either relevant or irrelevant to the query. the relevant documents are labeled as 1, and the irrelevant documents are labeled as 0. A binary cross

entropy loss is applied between the predicted score s_i, and the true label for each of the samples as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(s_i) + (1 - y_i)\log(1 - s_i)] \tag{7.6}$$

In this equation, N represents the total number of instances, y_i is the actual label for the ith instance, which can be 0 or 1, and s_i denotes the predicted relevance score of the ith instance being classified as class 1.

2. **Pair-wise**: In this scenario a contrastive learning strategy (Zou *et al.*, 2013) is employed. Within a marginal ranking loss, we have the relevance score of the query, and the relevant document which is going to be maximized during the optimization, and the relevance score of the query and the irrelevant documents, which is going to be minimized during the optimization. The loss function is formalized as follows:

$$L = \frac{1}{n}\sum_{i=1}^{N^+}\sum_{j=1}^{N^-}\max(0, m - T(q_i \oplus d_j^+) + T(q_i \oplus d_j^-)), \tag{7.7}$$

Where n is the total number of samples, N^+ is the total number of relevant documents, N^- is the total number of irrelevant documents, and m is the margin for the loss function.

3. **List-wise**: In the list-wise loss function the true and predicted distributions of the relevance scores of a list of documents for each query are being compared. let us say that the relevance score of a query q_i with a dpcument d_j in list of documents L is predicted as s_{ij}. The probability of the document d_j bering the top-1 document is calculated as:

$$P_s(j) = \frac{exp(s_{ij})}{\sum_{k=1}^{n} exp(s_{ik})} \tag{7.8}$$

Then the cross entropy between the true distribution of the probabilities, and the predicted probability for each query is calculated as:

$$L = -\sum_{j=1}^{n} P_y(j) \log(P_s(j)) \qquad (7.9)$$

Where, $P_y(j)$ is the true distribution of the relevance scores, and n is the total documents in the list L for the query q.

Li *et al.* (2022) introduce a method called In-Batch Balancing Regularization (IBBR) aimed at reducing ranking disparities within subgroups. Specifically, we create a differentiable normed Pairwise Ranking Fairness (nPRF) and apply T-statistics to nPRF across subgroups as a form of regularization, enhancing fairness in the process.

7.2 Loss Function Regularization

The loss function is a crucial component in training a neural ranker as it directly impacts parameter optimization and weight updates. Adjusting the loss function to align with the intended application can significantly influence the training process and achieve desired changes. Loss function regularization to reduce gender biases has been employed in NLP for debiasing the word embeddings (Qian *et al.*, 2019), language models (Barikeri *et al.*, 2021), and language generation (Garimella *et al.*, 2021).

The work by Qian *et al.* (2019) involves modifying the standard loss function used in training language models. The authors introduce a new term to the loss function that aims to equalize the probabilities of male and female words in the model's output. They use a pre-trained GloVe word embedding and an LSTM language model, tuning various hyperparameters to optimize performance. This new loss function encourages the model to treat gender pairs (like 'he' and 'she') equally, thereby reducing gender bias.

Similarly, Bordia and Bowman (2019) focus on identifying and mitigating gender bias in word-level language models. The authors propose a metric to measure gender bias in text corpora and the generated text from recurrent neural network language models trained on these corpora. They introduce a regularization loss term that minimizes the projection of embeddings onto a subspace encoding gender information.

Saunders and Byrne (2020) address the challenge of fine-tuning neural machine translation (NMT) models to reduce gender bias without losing general translation performance, which is often compromised due to catastrophic forgetting. This is achieved through loss regularization, specifically using Elastic Weight Consolidation (EWC). EWC helps retain knowledge from the original domain while adapting the model to a new, smaller, and more specific domain—in this case, a gender-balanced domain. During training, a regularization term is added to the original loss function, penalizing significant changes to parameters crucial for the original task.

The study presented in (Park *et al.*, 2023) addresses and reduces gender biases in pre-trained language models (PLMs) used for coreference resolution tasks. It identifies two primary types of gender biases: stereotype and skew. To mitigate these biases, two regularization techniques are introduced during the fine-tuning phase: Stereotype Neutralization (SN) and Elastic Weight Consolidation (EWC). SN neutralizes gender information in stereotypical words by distancing them from gender-inherent terms in the embedding space, while EWC preserves the model's essential linguistic capabilities by maintaining important model parameters, thus preventing performance loss. Additionally, a new metric called the Stereotype Quantification (SQ) score measures the consistency of gender pronoun predictions. The effectiveness of these methods and metrics is demonstrated on the WinoBias dataset, showing notable reductions in gender biases while preserving the model's linguistic performance.

A general formulation of loss function regularization can be written as:

$$L_{debias} = G(L(\theta, \lambda), \lambda', L'(\theta)) \tag{7.10}$$

Where L is the initial loss function which is going to optimize the parameters θ. In some cases, the regularization is done inside the loss function with a coefficient λ. In some other cases, the regularization is such that another loss function $L'(\theta)$ is interpolated with the initial loss with the coefficient λ', and as a result, the total loss is considered as a regularized version of the initial loss L. The function G acts as interpolating between the two losses L and L'.

7.2.1 Bias-aware Loss Function

Adjusting the ranking loss function to reduce gender bias in information retrieval systems is presented in a paper by Seyedsalehi *et al.* (2022a). The authors have proposed a loss function regularization method to mitigate gender biases in neural ranking systems while maintaining retrieval effectiveness. The core contribution of this work is a bias-aware neural ranker that explicitly considers and penalizes gender bias in documents during the ranking process. The proposed approach begins by defining the problem. The ranker is expected to balance two main objectives: maintaining retrieval effectiveness and reducing gender bias. These objectives are mathematically formalized as follows:

$$U(\hat{\Pi}, Q) \sim U(\Pi, Q) \tag{7.11}$$

$$\beta(\hat{\Pi}, Q) < \beta(\Pi, Q) \tag{7.12}$$

where Π is the state-of-the-art ranker, $\hat{\Pi}$ is the fair ranker, Q is the set of neutral queries, U represents the retrieval effectiveness, and β represents the quantifiable measure of bias.

To achieve these objectives, the ranking loss function used in neural rankers is adapted. The traditional ranking loss function is defined as:

$$L = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - \Phi(q, d_i^+) + \Phi(q, d_j^-)) \tag{7.13}$$

where q is the query, d_i^+ and d_j^- are relevant and irrelevant documents respectively, m is a margin, and $\Phi(q, d)$ is the relevance score of document d with respect to query q. This function focuses on maximizing the relevance of positive documents and minimizing the relevance of negative documents.

The bias-aware approach modifies this function to include a penalty for gender bias. The modified relevance scores are:

$$\Phi_B(q, d_i^+) = \Phi(q, d_i^+) - \Psi(d_i^+) \tag{7.14}$$

$$\Phi_B(q, d_j^-) = \Phi(q, d_j^-) + \Psi(d_j^-) \tag{7.15}$$

where $\Psi(d)$ is the bias measure of document d. This leads to the revised loss function:

$$L = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - \Phi_B(q, d_i^+) + \Phi_B(q, d_j^-)) \tag{7.16}$$

To ensure that the ranker's effectiveness is not overly compromised, the penalty is applied only to negative documents:

$$L = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - \Phi(q, d_i^+) + \Phi_B(q, d_j^-)) \tag{7.17}$$

This approach introduces a bias term as a regularizer, which adjusts the representation vectors in the embedding space, ensuring that biased irrelevant documents are pushed farther from the query vector.

Extensive experiments were conducted using the MS MARCO dataset and two sets of gender-neutral queries. The results demonstrated that the proposed method consistently reduced gender bias while maintaining or even improving retrieval effectiveness compared to existing methods. The method proved robust across different pre-trained contextual embeddings (ALBERT (Lan *et al.*, 2020), ELECTRA (Clark *et al.*, 2020b), and DistilRoBERTa (Sanh *et al.*, 2019)) and various datasets, confirming its effectiveness in balancing bias reduction with retrieval performance.

This approach addresses the critical need for fair and unbiased information retrieval systems, ensuring that users are presented with more equitable and representative search results without sacrificing the accuracy and relevance of the information retrieved.

7.2.2 COntextual Document Embedding Reranking

The methodology described in (Zerveas *et al.*, 2021) focuses on optimizing a retrieval model to score documents according to their relevance to a given query while simultaneously imposing a neutrality constraint on the top-ranked documents. This is done to ensure that the retrieved documents do not exhibit biases, particularly for gender.

The framework used for this purpose is the Contextual Document Embedding Reranking (CODER) framework. This approach involves several key steps:

1. *Query Embedding:* A pre-trained transformer encoder Z transforms a tokenized query q of length w into a sequence of d-dimensional embedding vectors:

$$Z = [z_1; \ldots ; z_w] = Z(q; \theta_Q) \in \mathbb{R}^{w \times d} \tag{7.18}$$

An aggregator function $g(Z)$ then extracts a single vector from these embeddings. In this work, the query encoder from TAS-B, based on DistilBERT, is chosen due to its effectiveness. The aggregation function selects the output embedding corresponding to the first query token, $g(Z) = z_1 \in \mathbb{R}^d$.

2. *Document Scoring:* A scoring function ϕ computes a scalar relevance score s_i for each document embedding $x_i \in \mathbb{R}^d$, based on their similarity to the query embedding $g(Z)$. The set of N documents includes the ground-truth relevant documents and the top candidates retrieved by an initial method such as BM25. Their embeddings are precomputed by the document encoder of a dual-encoder model. The scoring function uses a dot-product to evaluate similarity:

$$\hat{s} = \phi(g(Z), X) = X \cdot g(Z) \in \mathbb{R}^N \tag{7.19}$$

3. *ListNet Loss for Relevance:* The parameters of the query encoder are fine-tuned through the ListNet loss, which is equivalent to the KL-divergence between the distributions over target relevance labels y and the predicted scores $\hat{s}$:

$$L_u(y, \hat{s}) = D_{KL}(\sigma(y) \| \sigma(\hat{s})) = -\sum_{i=1}^{N} \sigma(y)_i \log \frac{\sigma(\hat{s})_i}{\sigma(y)_i} \tag{7.20}$$

where σ denotes the softmax function.

4. *Neutrality Loss:* To impose neutrality, a neutrality loss term is added, defined by the KL-divergence between the predicted scores and the neutrality scores y_n:

$$L_n(y_n, \hat{s}) = D_{KL}(\sigma(\hat{s}) \| \sigma(y_n)) = -\sum_{i=1}^{C} \sigma(\hat{s})_i \log \frac{\sigma(y_n)_i}{\sigma(\hat{s})_i} \tag{7.21}$$

Here, C is the cutoff rank for considering neutrality (set to 10), and y_n are the neutrality scores based on the frequency of bias-indicative terms.

5. *Total Loss:* The total loss function combines the relevance and neutrality losses with a regularization coefficient λ_r:

$$L_{tot} = L_u + \lambda_r L_n \tag{7.22}$$

The methodology includes using a large set of negative documents to effectively capture relevance and the use of the CODER framework for joint scoring of candidate documents, which together form a ranking context. The parameters of the encoder models are optimized to ensure both high relevance and reduced bias in the top-ranked documents.

The approach is compared with adversarial training methods, showing that the CODER framework achieves higher utility and fairness with stable optimization dynamics and a predictable intensity of bias mitigation.

7.2.3 In-Batch Balancing

Methods concentrating on in-batch balancing are designed to address sample distribution within a batch and introduce regularization to the batch loss.

The work by Li *et al.* (2022) focuses on mitigating ranking disparities among demographic groups through in-batch balancing. This approach incorporates a regularization term into the training loss function to directly optimize for fairness alongside accuracy. This approach utilizes the two-tower Dense Passage Retriever (DPR) model, which employs two dense encoders, E_P for passages and E_Q for queries. These encoders map the given text passage and input query to two d-dimensional vectors. The similarity score between a query q_i and a passage $p_{i,j}$ is defined as the dot product of their vectors:

$$\text{sim}(q_i, p_{i,j}) = z_{q_i}^\top z_{p_{i,j}} \tag{7.23}$$

where $z_{q_i} = E_Q(q_i)$ and $z_{p_{i_j}} = E_P(p_{i,j})$.

The total loss function comprises two components: the ranking loss and the fairness loss in which the ranking loss L_{Rank} measures the ranking performance for each data pair s_i:

$$L_{\text{Rank}} = -\log \frac{e^{\text{sim}(q_i, p_{i,1}^+)}}{e^{\text{sim}(q_i, p_{i,1}^+)} + \sum_{j=2}^{K} e^{\text{sim}(q_i, p_{i,j}^-)}} \tag{7.24}$$

where $p_{i,1}^+$ is the ground truth passage and $p_{i,j}^-$ are the non-clicked passages.

To mitigate bias, the paper introduces the concept of in-batch balancing, which includes two types of fairness loss functions:

1. *Pairwise Difference (PD) Loss.* The Pairwise Difference (PD) loss L_{Fair}^P measures the average ranking disparity between two groups (e.g., male and female) over a batch size B:

$$L_{\text{Fair}}^P = \frac{1}{n_m n_f} \sum_{c \in P_m^{[1:B]}} \sum_{d \in P_f^{[1:B]}} (\text{nPRF}_m(s_c) - \text{nPRF}_m(s_d))^2 \tag{7.25}$$

here, $P_m^{[1:B]}$ and $P_f^{[1:B]}$ are sets of clicked passages belonging to male and female groups over batch size B, and n_m and n_f are their respective counts. The normed Pairwise Ranking Fairness (nPRF) metric nPRF_m is defined as:

$$\text{nPRF}_m = \left\{ \frac{1}{n_{m1}(s_i) n_0(s_i)} \sum_{j \in g_{m1}(s_i)} \sum_{k \in g_0(s_i)} |R(p_{i,j})|^2 1[R(p_{i,j}) \geq R(p_{i,k})] \right\}^{\frac{1}{2}} \tag{7.26}$$

2. *T-statistics (TS) Loss.* The T-statistics (TS) loss L_{Fair}^T is based on the ranking disparity but includes second-order statistics (variance) of each group within the batch. It is defined as:

$$L_{\text{Fair}}^T = \left\{ \frac{(\hat{\mu}_m - \hat{\mu}_f)^2}{\sqrt{\hat{\text{var}}_m / n_m + \hat{\text{var}}_f / n_f}} \right\}^2 \tag{7.27}$$

where $\hat{\mu}_m$ and $\hat{\mu}_f$ are the means of the nPRF scores for male and female groups, respectively, and $\hat{\mathrm{var}}_m$ and $\hat{\mathrm{var}}_f$ are the variances.

The total loss function combines the ranking loss and the fairness loss:

$$L_{\text{total}}^{[1:B]} = L_{\text{Rank}}^{[1:B]} + \lambda L_{\text{Fair}}^{[1:B]} \tag{7.28}$$

here, λ is a hyperparameter controlling the balance between the ranking loss and the fairness loss. The fairness loss can be either the PD loss or the TS loss, depending on the chosen regularization approach.

In-batch balancing regularization (IBBR) directly incorporates fairness constraints into the training process of neural retrieval models. By penalizing ranking disparities within each batch of training data, the model learns to reduce biases while maintaining good retrieval performance. The proposed nPRF metric, along with the PD and TS loss functions, ensures that the model's fairness is optimized in a differentiable manner, making it suitable for integration into the training process.

7.3 Adversarial Training

Adversarial training, an essential technique in machine learning, originated from the groundbreaking work by Goodfellow *et al.* (2020). Their introduction of Generative Adversarial Networks (GANs) laid the foundation for this concept, where two neural networks—the generator and the discriminator—are trained in a competitive framework. The generator creates data, while the discriminator evaluates it, pushing the generator to produce increasingly realistic outputs.

Adversarial training for bias mitigation has been employed in many different areas, each requiring tailored adversarial attack designs based on specific applications. For example, in the paper by Yang *et al.* (2023), the authors introduce an adversarial training framework to mitigate algorithmic biases in clinical machine learning, specifically applied to predicting COVID-19 status. The methodology involves training a neural network-based classifier to predict outcomes while an adversary network attempts to predict sensitive features like ethnicity and hospital location. By optimizing the classifier to perform well while making the adversary's task difficult, the model reduces bias according to the equalized odds metric.

Similarly, Fleisig and Fellbaum (2022) present an adversarial learning framework aimed at reducing gender bias in sequence-to-sequence (seq2seq) machine translation models. The methodology involves fine-tuning large pre-trained language models using an adversarial objective that minimizes the gendered information in sentence embeddings. The protected variable, gender, is defined using two methods: a 'gender direction' derived from principal component analysis on sentence encodings and a simpler pronoun usage heuristic. The framework was tested on English-German and English-French translation tasks, demonstrating significant reductions in gender bias while maintaining or even slightly improving overall translation quality.

In the field of neural dialogue generation, an adversarial learning framework called Debiased-Chat, proposed by Liu *et al.* (2020), addresses gender bias in dialogue systems. The methodology involves using a disentanglement model to separate unbiased gender features from biased ones in dialogue utterances. The dialogue model is trained adversarially to generate responses containing only unbiased gender features while excluding biased ones.

Furthermore, Zhang, Ananiadou, *et al.* (2022) employ adversarial training to mitigate gender bias in text emotion detection by jointly training emotion detection and gender prediction models. This method uses CNN and Transformer-based emotion detection models, where the initial layers generate representations fed into a decoder for emotion prediction and an adversary for gender prediction. The adversarial framework aims to minimize the emotion detection loss while maximizing the adversary's loss, thus reducing gender-specific features in the emotion model. The training process includes a pretraining phase for individual model performance and an adversarial phase to balance emotion prediction accuracy and gender bias reduction, effectively reducing bias with minimal impact on overall performance.

Additionally, Chowdhury *et al.* (2021) introduce the Adversarial Scrubber (ADS) framework, designed to eliminate demographic information from textual data representations while preserving task performance. ADS employs an adversarial learning approach involving four modules:

Encoder, Scrubber, Bias Discriminator, and Target Classifier. The Encoder generates contextual representations, the Scrubber refines them to exclude demographic information, and the Bias Discriminator and Target Classifier aim to predict demographic attributes and task labels, respectively. The framework's efficacy was theoretically validated and empirically tested on eight datasets, demonstrating that ADS effectively minimizes demographic attribute information leakage while maintaining high task performance.

In the context of information retrieval, Rekabsaz *et al.* (2021) have proposed the method AdvBert. The AdvBert model builds upon the BERT architecture by introducing an adversarial training mechanism to mitigate biases in the relevance scoring process. The model consists of three main components:

1. BERT Encoder: Encodes the input query-document pair into an interaction embedding z.

2. Utility Network: Predicts the relevance score based on the interaction embedding z.

3. Adversarial Network: Predicts protected attributes (e.g., gender) from the interaction embedding z.

The BERT encoder processes a query q and a document d and produces an interaction embedding $z = f(q, d)$, where f is the encoding function. The utility network g then predicts the relevance score as $g(z)$. The adversarial network h aims to predict the protected attribute from z, using a two-layer feed-forward neural network with a tanh activation followed by a softmax layer, denoted as $h(z)$.

The objective of adversarial training is to make the interaction embedding z invariant to the protected attribute, thus reducing the model's bias. This is achieved through a min-max optimization problem, where the adversarial network h tries to predict the protected attribute, while the encoder f learns to make h fail in its prediction. The overall objective function L is defined as:

$$\arg\min_{f,g}\max_{h} L = L_{\text{util}}(q, X^{+}, X^{-}) - \lambda \left(L_{\text{adv}}(q, X^{+}) + L_{\text{adv}}(q, X^{-})\right) \tag{7.29}$$

here, L_{util} is the utility loss (typically a max-margin or cross-entropy loss) for predicting the relevance of positive (X^{+}) and negative (X^{-}) documents, while L_{adv} is the adversarial loss (cross-entropy loss) for predicting the protected attribute. The parameter λ controls the weight of the adversarial loss.

The adversarial network is trained to minimize the cross-entropy loss L_{adv}:

$$L_{\text{adv}}(q, X) = L_{\text{CE}}(h(f(q, d)), l) \tag{7.30}$$

where L_{CE} is the cross-entropy loss and l is the label indicating the presence of the protected attribute.

To optimize the encoder f and the utility network g, a gradient reversal layer (GRL) is introduced between the encoder and the adversarial network. The GRL acts as an identity function during the forward pass but multiplies the gradient by $-\lambda$ during backpropagation. This forces the encoder to learn representations that are informative for the utility task but uninformative for the adversarial task. The adversarial training procedure involves:

1. *Initialization:* The parameters of the encoder f and utility network g are initialized using a pre-trained BERT model fine-tuned on the original training data.

2. *Adversarial Training:* The adversarial network h is trained using a balanced dataset of gendered and non-gendered data points to ensure effective learning.

3. *Joint Optimization:* The entire model (encoder, utility network, and adversarial network) is jointly optimized using the adversarial training objective.

This adversarial training setup aims to balance the trade-off between fairness and utility by ensuring the relevance scores are less biased while maintaining high prediction performance.

7.4 Query Reformulation

Sparse representations, commonly known as bag-of-words models, have been a foundational technique in information retrieval for many years. Their simplicity and efficiency make them widely popular. However, like other representation techniques, they can be susceptible to gender biases. This section uncovers how gender bias can be subtly manifested in such representations and offers solutions to rectify this (Doughman *et al.*, 2021; Bigdeli *et al.*, 2021a; Bigdeli, 2021). More specifically, we focus on how query refinement methods such as weighted query expansion can help mitigate gender biases in retrieved documents (Bigdeli, 2021).

The foundational role of sparse representations, particularly the bag-of-words (BoW) model, cannot be overstated in the domain of information retrieval (IR). Originally developed as a simple yet effective means of text representation, the BoW model treats documents as a collection of words, disregarding grammar and word order but preserving multiplicity. This method has historically facilitated the implementation of various IR tasks due to its simplicity and computational efficiency. Moreover, its widespread adoption has catalyzed significant advancements in search technologies, making it a critical area of study within the field (Manning, 2008). Despite its limitations, such as the inability to capture semantic relationships between words, the BoW model remains a pivotal component in the evolution of text representation techniques.

As research progresses, it is imperative to address the gender biases inherent not only in traditional IR models like BoW but also in more sophisticated approaches such as neural embeddings. Neural embeddings, which represent text in continuous vector spaces, have been praised for their ability to capture deep semantic meanings that BoW models cannot. However, they also inherit and sometimes amplify biases present in the training data. Researchers such as Bolukbasi *et al.* (2016) have highlighted how word embeddings can exhibit stereotypical gender biases, with analogies like ‘man is to computer programmer as woman is to homemaker’ emerging from the models. This revelation underscores the necessity of examining gender biases across various representation models to develop more equitable IR systems. By studying these biases in both traditional and advanced models, researchers can gain insights into the mechanisms of bias propagation and intervention, setting the stage for comprehensive analyses and the development of debiasing techniques.

Thus, a thorough examination of representation techniques—from the basic bag-of-words to the complex neural embeddings—is crucial for advancing the field of IR towards more fair and unbiased systems. Investigating how gender biases manifest and are perpetuated by these models provides a clearer path to mitigating such biases. It is not only a technical challenge but also a societal imperative, as the outcomes of biased models can have profound implications on information access and equity in the digital age. By engaging with both historical and cutting-edge technologies, the research community can better understand and tackle the multifaceted nature of bias in information retrieval systems.

The study of sparse representation models, such as the bag-of-words (BoW), in the field of information retrieval (IR) reveals a critical but often overlooked aspect: the manifestation of gender bias within these frameworks. Sparse models, characterized by their straightforward approach to document representation, tend to encode and perpetuate biases present in the data they process. Since these models treat documents as mere collections of word occurrences, they lack the semantic depth to discern contextual nuances that could mitigate inherent biases. This can lead to skewed retrieval outcomes where documents reflect stereotypical or biased views, impacting the fairness and objectivity of search results and recommendations. The implications of such biases are profound, influencing how information is accessed and perceived by users, potentially reinforcing harmful stereotypes (Robertson, Zaragoza, *et al.*, 2009).

Despite their foundational role in both academic research and industry applications, sparse representation models have not received as much attention for bias mitigation as their neural counterparts. In recent years, the focus of bias studies and mitigation efforts in IR has shifted towards neural-based rankers and embeddings. These advanced models offer richer semantic representations and have been the subject of extensive scrutiny regarding how they encode, propagate, and can be adjusted to handle biases (Bolukbasi *et al.*, 2016). This shift in focus is partly due to the complex nature of biases in neural models and their increasing dominance in state-of-the-art IR systems. However, the less sophisticated sparse models are still widely used, especially in scenarios requiring computational efficiency and simplicity, which means that biases in these systems continue to affect a significant portion of users.

Highlighting this oversight is crucial for advancing the field of IR towards more equitable practices. There is a pressing need for more comprehensive research into how gender biases manifest in sparse representations and to develop effective strategies for their mitigation. This entails not only identifying the biases but also understanding their implications for the retrieval process and ultimately the end-users. By increasing awareness and research focus on sparse models, the IR community can ensure a more balanced approach to bias mitigation across different technological frameworks, fostering fairness and accuracy in retrieved documents. This effort is essential for building trust in information systems and for ensuring that advances in technology benefit all users equally.

Bias mitigation in IR systems, particularly those utilizing sparse representations like the bag-of-words (BoW) model, can significantly benefit from sophisticated query refinement techniques. One effective approach is weighted query expansion, which enhances the original query by adding relevant terms weighted by their significance to reduce bias and improve retrieval performance. This technique adjusts the query to capture a broader and more balanced set of documents that might otherwise be overshadowed by biased terms or associations in the original query formulation (Carpineto and Romano, 2012).

Weighted query expansion works by identifying terms that are semantically related to the original query but are less likely to carry gender biases. For instance, in a job-related search, terms like ‘engineer’ might traditionally retrieve documents that skew toward male-associated contexts. By expanding the query to include neutral or female-associated terms with calculated weights, the system can counteract this skew, leading to a more gender-balanced retrieval. This method not only broadens the search but also strategically alters the emphasis placed on certain terms, thus promoting fairness in the documents retrieved. Researchers like Diaz and Metzler (2006) have shown that such expansions can effectively diversify search results and enhance the relevance and fairness of the retrieval process.

The potential impact of applying weighted query expansion in sparse IR systems is substantial. By integrating these refinements, systems can move towards mitigating inherent biases, thus improving the accuracy and fairness of the search results. This approach aligns with the growing need to ensure that IR systems serve all users equitably, providing access to information that is not only relevant but also unbiased. Further research and implementation of these techniques can lead to more sophisticated and socially aware IR technologies, fostering a more inclusive digital information environment.

The methodology presented by Bigdeli *et al.* (2021a) focuses on investigating whether the tradeoff between bias and utility in IR can be mitigated. The authors propose a bias-aware pseudo-relevance feedback (PRF) method to achieve this goal. The primary hypothesis of the paper is that one can find a revised query q' that maintains the same level of utility as the original query q but significantly reduces bias in the retrieved documents. The authors suggest that by considering the degree of bias in the documents when selecting terms for query expansion, they can generate a revised query q' that is less biased.

The process begins by retrieving a set of documents D_q using an initial query q with a retrieval method M. The documents in D_q are then re-ranked based on both their relevance and their bias scores:

$$\mathrm{Rel}_{\mathrm{debiased}}(d) = (1-\lambda)\mathrm{Rel}(d) - \lambda\mathrm{Bias}(d) \tag{7.31}$$

where $d \in D$, $\mathrm{Rel}(d)$ is the relevance of document d to query q, $\mathrm{Bias}(d)$ is the bias score of document d, and λ is a linear interpolation coefficient that balances the importance of relevance and bias. Lower values of $\mathrm{Bias}(d)$ are desirable, hence they are subtracted from $\mathrm{Rel}(d)$.

Once the documents are re-ranked to produce $D_{\mathrm{debiased},q}$, the revised query q' is developed using the RM3 strategy, a pseudo-relevance feedback framework for query expansion. The revised query is formed by selecting the top-n terms with the highest scores:

$$\mathrm{Score}_t = \sum_{d \in D_{\mathrm{debiased},q}} \left(P(t|d) \log \frac{P(t|d)}{P(t|C)} \right) \tag{7.32}$$

here, $P(t|d)$ is the probability of term t given document d, and $P(t|C)$ is the probability of term t in the entire collection C.

The terms in the revised query are weighted as follows:

$$W_{\text{debiased}}(w, q) = \alpha P(w|q) + (1 - \alpha) P(w|D_{\text{debiased},q}) \tag{7.33}$$

where $\alpha \in [0, 1]$ and

$$P(w|D_{\text{debiased},q}) = \sum_{d \in D_{\text{debiased},q}} P(w|d) \prod_{t \in q} P(t|d) \tag{7.34}$$

This ensures that the weights assigned to the terms in the revised query reflect their likelihood in the less biased set of documents.

The findings demonstrate that it is possible to revise the initial query to maintain utility while significantly reducing bias. The paper concludes that the tradeoff between bias and utility is not absolute, and it is feasible to achieve both reduced bias and maintained utility through a bias-aware pseudo-relevance feedback approach. This methodology lays the foundation for considering fairness and utility as complementary rather than competing aspects in information retrieval systems. By revising the query using a bias-aware approach, the likelihood of including biased terms in the revised query is reduced, leading to a less biased ranked list of documents while maintaining retrieval effectiveness.

8

Challenges, Limitations, and Future Directions

In this section, we focus on highlighting the challenges and limitations of studying gender bias in information retrieval systems. We illuminate the difficulties associated with current metrics, datasets, and the limited definitions of gender available for research purposes. More specifically, we examine the properties of existing gender bias metrics from several perspectives, identifying which aspects have been thoroughly explored in previous work and which remain underexplored to enhance confidence in the bias measurements and studies. Additionally, later on in this chapter, we will discuss potential future research directions in this domain. We conclude the chapter by urging researchers to consider a broader range of psychological characteristics when quantifying biases and curating datasets and to adopt more comprehensive and theoretically grounded approaches for measuring and addressing gender biases.

8.1 Fairness Metrics Properties

A major challenge in quantifying gender bias within IR systems is the lack of a solid foundational understanding of the issue. Historically, this area has been relatively understudied, resulting in an incomplete grasp of both the nature of gender biases and the methodologies for accurately measuring them. This gap in foundational knowledge has significant implications for research and practice.

Researchers, confronted with the complexity of gender bias, have often had to develop their own measurement methods, frequently relying on assumptions and interpretations. As a result, the field is characterized by a diverse array of metrics, each with its own strengths and limitations (Qiu *et al.*, 2023). While these metrics may be individually valid, they often yield inconsistent findings, complicating the interpretation of results and obstructing the development of universally applicable strategies to mitigate gender bias (Sheng *et al.*, 2021). These inconsistencies raise critical questions about the effectiveness of current bias mitigation efforts and the accuracy of bias quantification. Ultimately, these issues in measurement may hinder the adoption of bias-free systems in real-world applications, as doubts about their effectiveness could emerge.

Developing reliable evaluation metrics requires not only a critical assessment of existing methods but also the proactive creation of new, more robust measures. Researchers should prioritize designing metrics that are resilient to variations in datasets, algorithms, and other environmental factors. These metrics would establish a stable foundation for assessing gender bias, ensuring that the results are not significantly affected by the peculiarities of the evaluation process.

Metrics for assessing biases in information access systems are crucial for ensuring fair and effective operation (Ekstrand *et al.*, 2021; Sundararaman and Subramanian, 2022). In the context of gender bias measurements, the *robustness* across different contexts to maintain integrity is crucial, as the nature and disclosure of gender bias can vary significantly across different datasets, cultural contexts, and application scenarios. *Trustworthy* is another pillar of reliable evaluation which is also a key to building trust in the metrics and, by extension, in the strategies developed to mitigate gender bias. For example, if two metrics yield divergent results for the same set of data, it would be challenging to determine the effectiveness of bias mitigation strategies. This investigation ensures that the metrics are not just theoretically sound but also practically applicable in real-world scenarios. As such, in evaluating a metric for assessing bias related to sensitive attributes, several critical dimensions must be considered. These include *validity*, *reliability*, *generalizability*, *sensitivity*, as well as *scalability and efficiency* of the fairness metric. The metric should align with clearly defined notions of fairness, whether counterfactual, individual, or group fairness. This includes ensuring that similar cases are treated similarly and that the metric does not inadvertently introduce or ignore discrepancies that could affect fairness evaluations (Sun *et al.*, 2022). Below, we will discuss each of these dimensions, exploring existing research that addresses these characteristics of fairness metrics in the context of gender bias in IR. We will also identify areas where potential growth exists for future research.

8.1.1 Validity

For metrics assessing gender bias in information retrieval systems, ensuring the validity of the assessment is very important. Validity refers to the metric's ability to accurately measure what it is intended to measure. The challenge in validating gender bias metrics lies in the notion of gender bias being quite abstractive and non-measurable. A common approach to validation of not-well-defined metrics on abstract concepts would be to validate against human data annotation. However, in this case, the human annotation could potentially introduce additional biases and intensify the bias influence on the results.

Without thorough validation studies, it remains uncertain whether these metrics align with actual societal and human perceptions of bias. Furthermore, the validation process must account for different scenarios in which the metrics are applied. For instance, the criteria for validating a metric might differ when assessing results for a neutral query versus a gendered query. Each scenario presents unique challenges and may require distinct validation approaches to ensure the metric's applicability across various contexts. As such, there exists no universal validation framework to ensure metrics validity.

Given these complexities, traditional methods of validation, such as absolute-level assessments, may not be entirely effective for gender bias metrics. Collecting human-level annotations on the degree of bias in individual documents is challenging to calibrate accurately. A more dynamic approach could involve pairwise preference-based assessments, where evaluators compare two sets of results for the same query to determine which set exhibits a relatively more intense attribute (e.g., higher level of bias or fairness) (Clarke *et al.*, 2020; Clarke *et al.*, 2021; Clarke *et al.*, 2023). This method focuses on relative comparisons rather than absolute values, which can be more informative for understanding the nuances of bias in information retrieval systems.

Overall, the field not only lacks extensive validation of gender bias metrics but also lacks sufficient research on how to effectively validate them. This gap leaves us uncertain about what these metrics truly measure and how accurately they reflect real-world biases. Future research should explore validation techniques that can accommodate the subjective and multifaceted nature of gender bias.

8.1.2 Reliability

Evaluation metrics need to be reliable; meaning they should consistently produce similar results under similar conditions and accurately measure what they are supposed to measure. Accurate measurement is vital, but equally crucial is the consistency of these measurements across different contexts. Without consistency, it becomes challenging to draw meaningful and actionable conclusions.

Uniform behaviour among metrics would indicate whether the methods are effective regardless of the evaluation metric that is being used. In the context of gender bias measurements, in (Klasnja *et al.*, 2022), the authors showed that while some metrics reliably detect highly biased and unbiased queries, ambiguities arise for the data points in between these extremes. The authors studied the level of agreement in commonly used metrics such as ARaB and NFaiRR in different scenarios (Fabris *et al.*, 2020; Rekabsaz and Schedl, 2020; Rekabsaz *et al.*, 2021). Their studies imply that gender bias quantifiers tend to converge more consistently at the extreme ends (highly biased or unbiased situations) but show lower agreement for queries that fall in the intermediate range. This suggests that current metrics might be less reliable for detecting subtle forms of gender bias. The divergence between different gender bias quantifiers, notably in the intermediate range of biases, implies that a limited selection of metrics may not offer a comprehensive view of gender bias. Such variances might jeopardize the credibility of research outcomes.

The inconsistency among existing metrics underscores the need for more thorough and integrative research in this field. A stronger backbone for these metrics can be developed through comprehensive studies that combine theoretical underpinnings with empirical validations. This entails a comprehensive exploration of the psychological characteristics underpinning gender bias, a deeper understanding of the socio-cultural factors at play, and a concerted effort to unify the disparate metrics into a cohesive and validated set of tools.

8.1.3 Generalizability

A metric should accurately reflect bias in varying contexts and should adapt to different types of data. A bias quantifier should be generalizable enough to be applicable across various demographic groups and not be confined to a specific attribute or scenario. An example of this could be the appropriate handling of language usage that may differ by region or group. One significant limitation of current studies is their reliance on gendered language and the focus predominantly on English language datasets. In general, the current gender bias metrics highly rely on gendered pronouns and terms and therefore mainly not applicable to non-gendered languages. This reliance raises questions about the applicability of these metrics to non-gendered languages and low-resource languages, where different linguistic structures and less data availability might slow down the direct application of existing methods.

In addition, the ability of the metric to effectively detect bias across multiple facets—including gender, race, ethnicity, and others—without requiring extensive customization for each new context would be highly beneficial. This universality would enhance the utility and relevance of the metric, making it a valuable tool in diverse settings and applications. Furthermore, while there is potential to adapt the concepts used in gender bias metrics to other types of bias measurements, such as those involving other sensitive attributes, the extent to which these adaptations would be effective remains largely unexplored. For instance, adapting gender bias quantifiers to assess biases related to age, ethnicity, etc could open new avenues for research. However, how effectively these adapted metrics would measure other sensitive biases and in what contexts they could be reliably applied have not been thoroughly investigated.

8.1.4 Sensitivity

Sensitivity is another crucial aspect in the quantification of any attributes, including gender bias in information retrieval systems. The metric should be resilient to slight changes in both data and algorithms, ensuring consistent and reliable measurements. An ideal bias quantifier metric should exhibit minimal sensitivity to environmental settings and hyperparameters. High sensitivity in these metrics can lead to significant challenges in accurately assessing gender bias. If a metric's outcomes vary dramatically with minor changes in its parameters, it becomes difficult to trust the consistency and objectivity of its results. Such variability can cast doubt on the metric's effectiveness, leading to potentially misleading conclusions about the presence or severity of gender bias.

Given that gender bias metrics rely heavily on predefined gender-related term lists, Klasnja et al. investigate how altering these lists can significantly impact bias measurements. The authors studied the sensitivity of gender bias metrics for the gendered term lists they rely on. They studied the effects of subsampling terms from comprehensive gender-related terms which most of the gender bias quantifiers such as ARaB (Rekabsaz and Schedl, 2020) and NFaiRR (Rekabsaz *et al.*, 2021) rely on. This methodological variation aimed to observe how the behaviour of each gender bias metric alters as the composition of gendered terms changes. The findings were revelatory – slight modifications in the list of gendered terms resulted in noticeable differences in the measured biases across all metrics. This indicates that the metrics' sensitivity to the specific terms used can lead to substantial variability in gender bias assessment. As a consequence of term list alterations the efficacy of these metrics in differentiating between different degrees of gender bias is questioned. As a result of this observation, we conclude that metric selection in evaluating gender bias is not trivial as it profoundly affects the conclusions drawn (Stanovsky *et al.*, 2019). Therefore, developing gender bias quantifiers that maintain their stability and accuracy, irrespective of minor alterations in hyperparameters like gendered term lists, is essential.

The dependency of gender bias metrics on psychological characteristics has also been explored in (Klasnja *et al.*, 2022). The authors examine the correlation between gender bias metrics and psychological attributes using tools like LIWC (Pennebaker *et al.*, 2001a). Their findings revealed that all examined metrics were highly correlated with male references, indicating an inclination towards the male gender in the metrics themselves. This brings two critical insights: first, the metrics themselves may be biased toward male inclinations, resulting in biased outcomes even when attempting to measure and mitigate bias. Second, it highlights the intrinsic link between the psychological characteristics of gender bias metrics and the choice of gendered terms.

8.1.5 Scalability and Efficiency

Efficiency and scalability are critical attributes for any metric designed to measure bias in increasingly complex information retrieval systems. As data volumes grow and models become more complex, the ability of a metric to adapt without requiring extensive computational resources becomes vital. Especially if they are being used to make decisions ad hoc and at the inference time.

Many existing metrics for assessing gender bias are primarily based on term frequencies, making them relatively inexpensive and fast to compute (Rekabsaz and Schedl, 2020; Rekabsaz *et al.*, 2021). This approach leverages simple statistical methods that do not require heavy computational overhead, ensuring that the metrics can be applied quickly even as datasets expand. Other groups of metrics utilize embedding representations of terms, which also contribute to computational efficiency (Basta *et al.*, 2019; Basta *et al.*, 2021; Bolukbasi *et al.*, 2016; Caliskan *et al.*, 2022; Chaloner and Maldonado, 2019; Kaneko *et al.*, 2022; Zhao *et al.*, 2020). These metrics typically require only the embedded representation of the text under assessment, operating primarily on projections onto gendered axes. This method, while slightly more complex than simple term frequency analyses, remains computationally manageable. The use of embeddings allows these metrics to capture deeper linguistic and semantic aspects without significant computational costs.

To the best of our knowledge, there has not been extensive research focused on studying the computational aspects of these metrics. However, the general consensus within the community is that these metrics are relatively cost-effective. This is supported by the absence of widespread complaints regarding their computational demands, suggesting that their efficiency has so far been adequate for current needs.

8.1.6 Positionality

As the field of responsible AI continues to grow, integrating concepts like positionality becomes essential for understanding and mitigating biases that affect model outcomes. Positionality in AI refers to the acknowledgment of how the identities, experiences, and perspectives of developers, data annotators, and end-users shape the creation and performance of AI systems. IR systems play a key role in distributing information across various contexts, from search engines to recommendation systems and academic databases. Positional biases embedded in these systems can influence how different groups are represented and served, which can perpetuate stereotypes or marginalize voices that are already underrepresented. In the context of IR, the understanding of positionality in AI-based IR systems is vital because IR systems are designed to mediate access to information based on user queries, relevance models, and the content of indexed data.

There could be several sources of positionality in IR development, including data and annotation biases (Posada, 2023). The content used to train IR models is often annotated based on subjective standards of relevance or quality. Positionality affects these annotations (Miceli *et al.*, 2022), since the interpretations of the annotators are influenced by their worldviews (Kapania *et al.*, 2023). This can result in data that disproportionately represents certain groups or reflects prevailing societal norms rather than a more inclusive understanding of user needs. Their system might also suffer from systemic Bias Reinforcement. IR systems often learn from user interaction data to improve relevance ranking. If these interactions are themselves biased—favouring majority user behaviours or perspectives—the system's learning process reinforces these patterns. This can result in search outputs that marginalize minority viewpoints or perpetuate biases in content ranking and retrieval.

Integrating positionality into IR systems is not without its challenges. One key difficulty is balancing the need for representative data with the technical and practical constraints of system development. In addition, there may be resistance within the industry to adopt reflexive practices that challenge established norms or workflows.

8.2 Gender Definition

In this section, we highlight the critical gap in research on gender definitions that transcend the binary paradigm, advocating for a more encompassing understanding of the notion of gender (Niousha *et al.*, 2023). Historically, gender has predominantly been perceived within a binary framework, often culminating in a constrained interpretation across various fields, including but not limited to information-seeking systems (Smith, 2021). Such an oversimplified perspective disregards the complicated shades of gender identities, encompassing non-binary, transgender, and other non-traditional categorizations. Several instances where IR systems displayed inadequacies due to this limited gender comprehension emphasize the pressing need for evolution (Krieg *et al.*, 2023; Wang *et al.*, 2021). In this section, we acknowledge the reasons for this limited exploration beyond the binary understanding of gender are multifaceted.

Societal norms and cultural biases have historically upheld binary gender definitions. Historically, the concept of gender has predominantly been constrained within a binary framework, wherein individuals are classified strictly as male or female. This binary perception of gender has deeply permeated various fields, including IR and NLP, influencing how data is structured, how algorithms are designed, and ultimately, how users interact with these systems. In other words, technological models were initially built reflecting the societal constructs of their time, therefore, inadvertently reinforcing this binary notion. Such a simplified understanding leads to oversimplifications in the treatment of gender, often ignoring other aspects of realities experienced by individuals who do not conform to this binary classification. This historical bias not only reflects but also reinforces societal norms and prejudices, thereby perpetuating a cycle of exclusion and misunderstanding within technological applications. As the modern world becomes increasingly aware of diverse gender identities, the collective call for a more holistic and inclusive approach intensifies. Embracing and integrating this extensive spectrum of gender identities is paramount to developing information retrieval systems that truly serve all users equitably.

The limitations imposed by a binary view of gender have concrete repercussions on the functionality and fairness of IR systems. For instance, systems that categorize users strictly as male or female fail to acknowledge and serve non-binary, transgender, and other non-traditionally gendered individuals (see Chapter 2). This oversight can lead to a lack of personalized and relevant search results and recommendations, affecting user satisfaction and system efficacy. In more critical applications, such as healthcare information retrieval, the failure to recognize and appropriately address diverse gender identities can result in significant disparities in the quality of information provided, potentially affecting the health and well-being of underserved populations. These examples highlight the pressing need for IR systems to move beyond binary classifications and towards more inclusive, nuanced understandings of gender that reflect the diversity of users' identities and experiences.

The imperative to transcend the binary definition of gender in IR research and application is driven by both ethical considerations and practical necessity. As societal awareness and acceptance of diverse gender identities increase, there is a corresponding expectation for systems to evolve to accommodate this diversity. Recognizing non-binary, genderqueer, transgender, and other identities is essential not only for the inclusivity and fairness of IR systems but also for their accuracy and effectiveness. The modern world's acknowledgment of these identities increases the demand for IR systems that can understand and cater to a broader spectrum of user needs and experiences. Moreover, as the data on gender diversity grows, the opportunity to innovate and improve these systems by incorporating a more comprehensive understanding of gender becomes possible and increasingly important.

Advocating for an inclusive approach in the development of IR systems involves acknowledging the full spectrum of gender identities and ensuring these perspectives are integrated into the technology we create. This advocacy is crucial in fostering equitable systems that serve all users effectively. Inclusive systems are more likely to be fair and unbiased, which is increasingly important in a globalized society where technology impacts a wide array of individuals with varied identities and needs. Ultimately, the integration of diverse gender identities into IR systems is not just a technical challenge but a moral imperative that reflects the principles of equity and justice in technological development.

8.3 Datasets Limitations

In this section, we examine the ongoing challenge of creating datasets that are genuinely diverse and representative, a priority that the IR community must urgently address (De-Arteaga *et al.*, 2019). Datasets are foundational to the development, benchmarking, and evaluation of IR systems (Krieg *et al.*, 2023). However, many commonly used datasets come with their set of biases and limitations. Skewed representations, particularly those that underrepresent certain gender identities, can have a domino effect, resulting in biased IR systems (Kopeinik *et al.*, 2023). Ethical dilemmas further complicate the matter, especially when it comes to annotating gender in datasets (Fekih *et al.*, 2022). Manual annotations, while time-consuming and costly to collect, can perpetuate existing biases if not performed with careful consideration. Furthermore, as we recognize the importance of representing the diversity of human experiences, there is an increased emphasis on ensuring datasets span a range of cultures, languages, and identities.

Datasets are fundamental to the development, benchmarking, and evaluation of IR systems, highlighting the crucial importance of proper dataset curation. Effective dataset curation not only supports robust system development but also ensures the reliability and validity of benchmarks used across the research community (Krieg *et al.*, 2023). Well-curated datasets enable researchers and developers to generate meaningful insights and drive innovations, making dataset curation a critical focus in IR research. However, this task presents significant challenges, given the complexity of capturing a comprehensive snapshot of human knowledge and interaction, especially as digital data continues to grow in volume and diversity. As the IR community advances, it is imperative to prioritize the creation of rich, diverse, and well-structured datasets to support the next generation of retrieval technologies and applications, ensuring they are both effective and relevant to a wide range of user needs.

Curating datasets that involve sensitive topics, such as gender identity, presents unique challenges beyond those encountered in traditional dataset construction. Sensitive topics require a nuanced approach to data collection and annotation, where ethical considerations and the potential for harm must be carefully balanced against research objectives (Fekih *et al.*, 2022). The complexity is further compounded by the need to accurately represent diverse perspectives and experiences without imposing a dominant cultural or societal bias. These datasets must be handled with a heightened awareness of privacy, consent, and the potential for re-identification. Researchers must navigate these ethical waters while striving to construct datasets that genuinely reflect the multifaceted nature of human identity, which is essential for developing IR systems that are truly inclusive and fair.

Inherent biases and limitations in commonly used datasets can significantly skew the development and performance of IR systems (Kopeinik *et al.*, 2023). These biases often stem from non-representative sample populations, where certain demographics, especially marginalized groups, are underrepresented. This lack of diversity can lead to systems that perform well for the majority but fail to address the needs of all users equitably. Recognizing and correcting these biases is crucial for developing IR systems that offer equal access and quality of information across diverse user bases. To combat these limitations, there is a growing emphasis on creating datasets that are diverse and representative of various cultures, languages, and identities. Such comprehensive datasets are fundamental to building equitable and inclusive systems that can serve the global community effectively and sensitively.

The annotation of gender in datasets presents significant ethical challenges, especially with manual annotations, which are time-consuming and potentially costly. If not approached with a critical and informed perspective, there is a considerable risk of perpetuating existing biases (Fekih *et al.*, 2022). Ethical annotation practices must prioritize transparency, informed consent, and the avoidance of reinforcing stereotypes. Additionally, researchers must recognize the dynamic and subjective nature of gender identity, which does not always fit into binary or static categories. Developing guidelines and methodologies that respect participant identities while ensuring data accuracy and utility is a delicate balance that requires ongoing attention and refinement.

The intersectionality of biases, where gender bias intersects with other demographic attributes such as ethnicity, adds significant complexity to dataset curation and system development. These overlapping biases can exacerbate the challenges of creating fair and balanced IR systems, as the interaction between different identity facets can influence user experiences and system accuracy in nuanced ways. Addressing these intersecting biases is essential for developing systems that are not only gender-aware but also broadly inclusive. Achieving this requires a holistic approach to dataset creation and algorithm design, one that accounts for the full spectrum of human diversity and the various ways biases can manifest and interact.

8.4 Future Directions

In this section, we study potential paths for future research on gender bias in information retrieval systems. We note that the quest for fair and unbiased IR systems is a continuous journey. As our understanding of gender evolves, IR systems must remain agile and adaptable. Beyond just gender, there is a growing acknowledgment of the need to study fairness in the broader sensitive attribute context, encompassing biases like race, age, or socio-economic status. Intersectionality, the interplay of multiple biases, adds another layer of complexity to this endeavour. As the field progresses, there's a growing emphasis on making IR systems transparent in their operations and methodologies, ensuring users have a clear understanding of how information is retrieved. Several key areas warrant attention for future research and development. These directions aim to build upon existing foundational work while advancing the creation of more inclusive, equitable, and effective IR systems.

8.4.1 Intersectionality and Multidimensional Biases

While substantial progress has been made in understanding and mitigating gender bias, future research must explore the intersectionality of biases. Intersectionality refers to the complex, cumulative manner in which different forms of discrimination—such as those based on gender, race, socioeconomic status, and ethnicity—intersect and interact. Addressing gender bias in isolation may lead to incomplete solutions, as individuals' experiences of bias are often multifaceted. Future work should focus on developing methodologies that consider the interplay of multiple biases, ensuring that IR systems provide equitable access to information for all users, regardless of their intersecting identities.

8.4.2 Beyond Binary Gender Classifications

Current IR systems largely operate within a binary framework of gender, which does not account for the diverse spectrum of gender identities present in modern societies. Future research should explore how IR systems can move beyond binary classifications to better serve users who identify as non-binary, genderqueer, or with other non-traditional gender identities. This includes developing datasets that accurately reflect this diversity and creating algorithms capable of processing and responding to gender in a more nuanced manner. Such advancements would not only enhance the inclusivity of IR systems but also improve their accuracy and relevance to a broader user base.

8.4.3 Addressing the Limitations of Annotated Datasets

The limited size of annotated datasets poses a significant challenge to the reliable detection and analysis of biases in IR systems. Small datasets can lead to overfitting, reduced generalizability, and unreliable bias metrics. Future research should focus on developing methods for efficiently expanding annotated datasets, including semi-supervised learning, active learning, and crowdsourcing techniques. Additionally, creating benchmarks and standards for dataset annotation processes will help ensure the quality and consistency of the data, enabling more accurate assessments of bias and fairness.

8.4.4 Development of Robust and Reliable Fairness Metrics

The development of robust fairness metrics is critical for assessing and mitigating bias in IR systems. However, current metrics often suffer from reliability issues, including inconsistencies across different datasets and contexts. Future research should focus on creating metrics that are not only resilient to variations in data and algorithms but also consistently reliable across diverse scenarios. These metrics should be validated through extensive testing and benchmarking, ensuring that they can detect biases accurately and consistently. Enhancing the reliability of fairness metrics will be key to making meaningful progress in bias mitigation.

8.4.5 Handling Multilingual Data and Cross-Cultural Biases

As IR systems are increasingly deployed in multilingual and multicultural contexts, addressing the unique challenges of multilingual data is essential. Biases can manifest differently across languages due to cultural nuances, differences in language structure, and varying levels of data availability. Future research should focus on developing methods for detecting and mitigating biases in multilingual datasets, including cross-lingual transfer learning and the creation of language-agnostic fairness metrics. Additionally, ensuring that IR systems are culturally sensitive and capable of fairly serving users from diverse linguistic backgrounds is critical for achieving global inclusivity.

8.4.6 Real-Time Bias Mitigation

As IR systems increasingly operate in real-time, the ability to detect and mitigate bias on-the-fly becomes crucial. Future research should explore the development of real-time bias detection and correction mechanisms that can be integrated into IR systems without compromising their performance. This includes advancements in machine learning models that can adapt to new data and evolving biases, ensuring that IR systems remain fair and equitable as they process information in dynamic environments.

8.4.7 Definition and Scope of Sensitive Attributes

A clear and consistent definition of sensitive attributes is essential for developing effective bias mitigation strategies in IR systems. Future research should focus on establishing standardized criteria for what constitutes a sensitive attribute, considering the socio-cultural context and the potential impact on different user groups. This includes not only traditional attributes like gender and race but also context-specific attributes that may emerge as relevant in certain applications. Defining sensitive attributes rigorously will guide the development of fairness metrics and the creation of datasets that accurately reflect the diversity of user identities and experiences.

References

Abdollahpouri, H., M. Mansoury, R. Burke, B. Mobasher, and E. Malthouse. (2021). “User-centered evaluation of popularity bias in recommender systems”. In: *Proceedings of the 29th ACM conference on user modeling, adaptation and personalization*. 119–129.

Abolghasemi, A., L. Azzopardi, A. Askari, M. de Rijke, and S. Verberne. (2024). “Measuring Bias in a Ranked List Using Term-Based Representations”. In: *European Conference on Information Retrieval*. Springer. 3–19.

Adam, A. (1998). “Feminist resources”. In: *Artificial Knowing: Gender and the Thinking Machine*. New York: Routledge. 11–34.

Adomavicius, G. and Y. Kwon. (2011). “Improving aggregate recommendation diversity using ranking-based techniques”. *IEEE Transactions on Knowledge and Data Engineering*. 24(5): 896–911.

Ainsworth, C. (2015). “Sex redefined”. *Nature*. 518(7539): 288–291. DOI: 10.1038/518288a. URL: https://doi.org/10.1038/518288a.

Albert, K. and M. Delano. (2022). “Sex trouble: Sex/gender slippage, sex confusion, and sex obsession in machine learning using electronic health records”. *Patterns*. 3(8): 1–11.

Alesina, A., P. Giuliano, and N. Nunn. (2013). “On the origins of gender roles: Women and the plough”. *The Quarterly Journal of Economics*. 128(2): 469–530. DOI: 10.1093/qje/qjt005. URL: https://doi.org/10.1093/qje/qjt005.

Anderson, S. M. (2020). "Gender matters: The perceived role of gender expression in discrimination against cisgender and transgender LGBQ individuals". *Psychology of Women Quarterly*. 44(3): 323–341. DOI: 10.1177/0361684320929354. URL: https://doi.org/10.1177/0361684320929354.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner. (2016). "Machine Bias". *ProPublica*. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arboleda, V. A., D. E. Sandberg, and E. Vilain. (2014). "DSDs: Genetics, underlying pathologies and psychosexual differentiation". *Nature Reviews Endocrinology*. 10(10): 603–615. DOI: 10.1038/nrendo.2014.130. URL: https://doi.org/10.1038/nrendo.2014.130.

De-Arteaga, M., A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. T. Kalai. (2019). "Bias in bios: A case study of semantic representation bias in a high-stakes setting". In: *proceedings of the Conference on Fairness, Accountability, and Transparency*. 120–128.

Ashley, F. (2023). "What Is It like to Have a Gender Identity?" *Mind*: fzac071.

Atagi, N., N. Sethuraman, and L. B. Smith. (2009). "Conceptualizations of gender in language". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 31. URL: https://escholarship.org/uc/item/69s2385t.

Baeza-Yates, R. (2018). "Bias on the web". *Communications of the ACM*. 61(6): 54–61.

Bannister, J. J., H. Juszczak, J. D. Aponte, D. C. Katz, P. D. Knott, S. M. Weinberg, B. Hallgrímsson, N. D. Forkert, and R. Seth. (2022). "Sex differences in adult facial three-dimensional morphology: application to gender-affirming facial surgery". *Facial Plastic Surgery & Aesthetic Medicine*. 24(S2): S–24. DOI: 10.1089/fpsam.2021.0301. URL: https://www.liebertpub.com/doi/full/10.1089/fpsam.2021.0301.

Barikeri, S., A. Lauscher, I. Vulić, and G. Glavaš. (2021). "RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models". *arXiv preprint arXiv:2106.03521*.

Barocas, S., M. Hardt, and A. Narayanan. (2020). *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. URL: https://fairmlbook.org.

Barocas, S. and A. D. Selbst. (2016). “Big Data’s Disparate Impact”. *California Law Review*. 104(3): 671–732. DOI: 10.15779/Z38BG31. URL: https://doi.org/10.15779/Z38BG31.

Basta, C., M. R. Costa-Jussa, and N. Casas. (2021). “Extensive study on the underlying gender bias in contextualized word embeddings”. *Neural Computing and Applications*. 33(8): 3371–3384.

Basta, C., M. R. Costa-Jussà, and N. Casas. (2019). “Evaluating the underlying gender bias in contextualized word embeddings”. *arXiv preprint arXiv:1904.08783*.

Beattie, L., D. Taber, and H. Cramer. (2022). “Challenges in translating research to practice for evaluating fairness and bias in recommendation systems”. In: *Proceedings of the 16th ACM Conference on Recommender Systems*. 528–530.

Beltz, A. M., A. M. Loviska, and A. Weigard. (2021). “Daily gender expression is associated with psychological adjustment for some people, but mainly men”. *Scientific Reports*. 11(1). DOI: 10.1038/s41598-021-88279-4. URL: https://doi.org/10.1038/s41598-021-88279-4.

Bernstein, E. S. and S. Turban. (2018). “The impact of the ‘Open’ workspace on human collaboration”. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 373(1753): 20170239. DOI: 10.1098/rstb.2017.0239. URL: https://doi.org/10.1098/rstb.2017.0239.

Biega, A. J., K. P. Gummadi, and G. Weikum. (2018). “Equity of attention: Amortizing individual fairness in rankings”. In: *The 41st international acm sigir conference on research & development in information retrieval*. 405–414.

Bigdeli, A. (2021). “Exploration and Mitigation of Stereotypical Gender Biases in Information Retrieval Systems”.

Bigdeli, A., N. Arabzadeh, S. Seyedsalehi, B. Mitra, M. Zihayat, and E. Bagheri. (2023). “De-biasing Relevance Judgements for Fair Ranking”. In: *Advances in Information Retrieval.* Ed. by J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, and A. Caputo. Cham: Springer Nature Switzerland. 350–358. ISBN: 978-3-031-28238-6.

Bigdeli, A., N. Arabzadeh, S. Seyedsalehi, M. Zihayat, and E. Bagheri. (2022). “A Light-Weight Strategy for Restraining Gender Biases in Neural Rankers”. In: *Advances in Information Retrieval.* Ed. by M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, and V. Setty. Cham: Springer International Publishing. 47–55. ISBN: 978-3-030-99739-7.

Bigdeli, A., N. Arabzadeh, S. Seyersalehi, M. Zihayat, and E. Bagheri. (2021a). “On the Orthogonality of Bias and Utility in Ad hoc Retrieval”. In: *Proceedings of the 44rd International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Bigdeli, A., N. Arabzadeh, M. Zihayat, and E. Bagheri. (2021b). “Exploring Gender Biases in Information Retrieval Relevance Judgement Datasets”. In: *European Conference on Information Retrieval.* Springer. 216–224.

Bingley, W. J., C. Curtis, S. Lockey, A. Bialkowski, N. Gillespie, S. A. Haslam, R. K. L. Ko, N. Steffens, J. Wiles, and P. Worthy. (2023). “Where is the human in human-centered AI? Insights from developer priorities and user experiences”. *Computers in Human Behavior.* 141: 107617. DOI: 10.1016/j.chb.2022.107617. URL: https://doi.org/10.1016/j.chb.2022.107617.

Binns, R. (2018). “Fairness in Machine Learning: Lessons from Political Philosophy”. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*: 149–159. URL: https://dl.acm.org/doi/10.1145/3287560.3287583.

Binns, R., M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt. (2018). “’It’s Reducing a Human Being to a Percentage’ Perceptions of Justice in Algorithmic Decisions”. In: *Proceedings of the 2018 Chi conference on human factors in computing systems.* 1–14.

Blackstone, A. (2003). “Gender roles and society”. In: *Human Ecology: An Encyclopedia of Children, Families, Communities, and Environments*. Ed. by J. R. Miller, R. M. Lerner, and L. B. Schiamberg. Santa Barbara, CA: ABC-CLIO. 335–338. ISBN: 1-57607-852-3.

Bogen, M. and A. Rieke. (2018). “Help wanted: An examination of hiring algorithms, equity, and bias”. *Upturn, December*. 7.

Boinodiris, P. (2024). “The importance of diversity in AI isn’t opinion, it’s math”. URL: https://www.ibm.com/blog/why-we-need-diverse-multidisciplinary-coes-for-model-risk/.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. (2017). “Enriching word vectors with subword information”. *Transactions of the association for computational linguistics*. 5: 135–146.

Bolte, G., K. Jacke, K. Groth, U. Kraus, L. Dandolo, L. Fiedel, M. Debiak, M. Kolossa-Gehring, A. Schneider, and K. Palm. (2021). “Integrating sex/gender into environmental health research: Development of a conceptual framework”. *International Journal of Environmental Research and Public Health*. 18(22): 12118. DOI: 10.3390/ijerph182212118. URL: https://doi.org/10.3390/ijerph182212118.

Bolukbasi, T., K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. (2016). “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”. *Advances in neural information processing systems*. 29.

Borau, S., T. Otterbring, S. Laporte, and S. Fosso Wamba. (2021). “The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI”. *Psychology & Marketing*. 38(7): 1052–1068. DOI: 10.1002/mar.21480. URL: https://doi.org/10.1002/mar.21480.

Bordia, S. and S. R. Bowman. (2019). “Identifying and reducing gender bias in word-level language models”. *arXiv preprint arXiv:1904.03035*.

Bösch, F., M. K. Angele, and I. H. Chaudry. (2018). “Gender differences in trauma, shock and sepsis”. *Military Medical Research*. 5(1): 35. DOI: 10.1186/s40779-018-0182-5. URL: https://doi.org/10.1186/s40779-018-0182-5.

Bowman-Smart, H., J. Savulescu, M. O'Connell, and A. Sinclair. (2024). "World Athletics regulations unfairly affect female athletes with differences in sex development". *Journal of the Philosophy of Sport*. 51(1): 29–53. DOI: 10.1080/00948705.2024.2316294. URL: https://doi.org/10.1080/00948705.2024.2316294.

Bredella, M. A. (2017). "Sex Differences in Body Composition". In: *Sex and Gender Factors Affecting Metabolic Homeostasis, Diabetes and Obesity*. Springer International Publishing. 9–27.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.* (2020). "Language models are few-shot learners". *Advances in neural information processing systems*. 33: 1877–1901.

Bruce, V., A. M. Burton, E. Hanna, P. Healey, O. Mason, A. Coombes, R. Fright, and A. Linney. (1993). "Sex discrimination: how do we tell the difference between male and female faces?" *perception*. 22(2): 131–152. DOI: 10.1068/p220131. URL: https://doi.org/10.1068/p220131.

Buda, M., A. Maki, and M. A. Mazurowski. (2018). "A systematic study of the class imbalance problem in convolutional neural networks". *Neural networks*. 106: 249–259.

Buolamwini, J. (2024). *Unmasking AI: My mission to protect what is human in a world of machines*. Random House.

Buslón, N., A. Cortés, S. Catuara-Solarz, D. Cirillo, and M. J. Rementeria. (2023). "Raising awareness of sex and gender bias in artificial intelligence and health". *Frontiers in Global Women's Health*. 4. DOI: 10.3389/fgwh.2023.970312. URL: https://doi.org/10.3389/fgwh.2023.970312.

Caira, C., L. Russo, and L. Aranda. (2023). "Artificially inequitable? AI and closing the gender gap". URL: https://oecd.ai/en/wonk/closing-the-gender-gap.

Caliskan, A., P. P. Ajay, T. Charlesworth, R. Wolfe, and M. R. Banaji. (2022). "Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 156–170.

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

Callan, J., M. Hoy, C. Yoo, and L. Zhao. (2009). “The ClueWeb09 dataset”. In: *Proc. of the 32nd annual international ACM SIGIR conference on Research and development in information retrieval.* 523–524.

Cao, Y. T. and H. Daumé. (2021). “Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle”. *Computational Linguistics.* 47(3): 615–661. DOI: 10.1162/coli_a_00413. URL: https://doi.org/10.1162/coli_a_00413.

Carpineto, C. and G. Romano. (2012). “A survey of automatic query expansion in information retrieval”. *Acm Computing Surveys (CSUR).* 44(1): 1–50.

Castets-Renard, C. and C. Lequesne. (2023). “Abortion in the age of AI: A need for safeguarding reproductive rights in the United States and the European Union”. *McGill Law Journal.* 69: 1–17.

Chaloner, K. and A. Maldonado. (2019). “Measuring gender bias in word embeddings across domains and discovering new gender bias word categories”. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing.* 25–32.

Chang, A. R. and S. M. Wildman. (2017). “Gender in/sight: Examining culture and constructions of gender”. *Georgetown Journal of Gender and the Law.* 18(1): 43–80.

Chelba, C., T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson. (2013). “One billion word benchmark for measuring progress in statistical language modeling”. *arXiv preprint arXiv:1312.3005.*

Chen, J., H. Dong, X. Wang, F. Feng, M. Wang, and X. He. (2023a). “Bias and debias in recommender system: A survey and future directions”. *ACM Transactions on Information Systems.* 41(3): 1–39.

Chen, J., X. Wang, F. Feng, and X. He. (2021). “Bias issues and solutions in recommender system: Tutorial on the recsys 2021”. In: *Proceedings of the 15th ACM Conference on Recommender Systems.* 825–827.

Chen, P., L. Wu, and L. Wang. (2023b). “AI fairness in data management and analytics: A review on challenges, methodologies and applications”. *Applied Sciences*. 13(18): 10258. DOI: 10.3390/app131810258. URL: https://doi.org/10.3390/app131810258.

Chen, Z. (2023). “Ethics and discrimination in artificial intelligence-enabled recruitment practices”. *Humanities and Social Sciences Communications*. 10(1). DOI: 10.1057/s41599-023-02079-x. URL: https://doi.org/10.1057/s41599-023-02079-x.

Chmielinski, K., S. Newman, M. Taylor, J. Joseph, K. Thomas, J. Yurkofsky, and Y. Qiu. (2022). “The dataset nutrition label (2nd gen): Leveraging context to mitigate harms in artificial intelligence”. *ArXiv*. abs/2201.03954.

Chouldechova, A. (2017). “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments”. *Big Data*. 5(2): 153–163. DOI: 10.1089/big.2016.0047. URL: https://doi.org/10.1089/big.2016.0047.

Chowdhury, A. G., R. Sawhney, R. Shah, and D. Mahata. (2019). “# YouToo? detection of personal recollections of sexual harassment on social media”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2527–2537.

Chowdhury, S. B. R., S. Ghosh, Y. Li, J. B. Oliva, S. Srivastava, and S. Chaturvedi. (2021). “Adversarial scrubbing of demographic information for text classification”. *arXiv preprint arXiv:2109.08613*.

Clark, K., M.-T. Luong, Q. V. Le, and C. D. Manning. (2020a). “Electra: Pre-training text encoders as discriminators rather than generators”. *arXiv preprint arXiv:2003.10555*.

Clark, K., M.-T. Luong, Q. V. Le, and C. D. Manning. (2020b). “Electra: Pre-training text encoders as discriminators rather than generators”. *arXiv preprint arXiv:2003.10555*.

Clarke, C. L., N. Craswell, I. Soboroff, A. Ashkan, E. Agichtein, and F. Díaz. (2004). “Overview of the TREC 2004 Terabyte Track”. In: *TREC*. Vol. 2004. 74–85.

Clarke, C. L., N. Craswell, I. Soboroff, A. Ashkan, E. Agichtein, and F. Díaz. (2012). “The TREC 2012 Web Track”. In: *TREC*. Vol. 2012. 1–12.

Clarke, C. L., F. Diaz, and N. Arabzadeh. (2023). “Preference-based offline evaluation”. In: *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 1248–1251.

Clarke, C. L., A. Vtyurina, and M. D. Smucker. (2020). “Offline evaluation without gain”. In: *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 185–192.

Clarke, C. L., A. Vtyurina, and M. D. Smucker. (2021). “Assessing top-preferences”. *ACM Transactions on Information Systems (TOIS)*. 39(3): 1–21.

Craswell, N., B. Mitra, E. Yilmaz, D. Campos, and E. M. Voorhees. (2020). “Overview of the TREC 2019 deep learning track”. *arXiv preprint arXiv:2003.07820*.

Crenshaw, K. (1991). “Mapping the margins: Intersectionality, identity politics, and violence against women of color”. *Stanford Law Review*. 43(6): 1241. DOI: 10.2307/1229039. URL: https://doi.org/10.2307/1229039.

Crenshaw, K. W. (2017). *On Intersectionality: Essential Writings*. No. 255. Faculty Books. URL: https://scholarship.law.columbia.edu/books/255.

Cubuk, E. D., B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. (2018). “Autoaugment: Learning augmentation policies from data”. *arXiv preprint arXiv:1805.09501*.

D’Ignazio, C. and L. Klein. (2020). “What Gets Counted Counts”. In: *Data Feminism*. MIT Press. 9–27.

Dai, Z., C. Xiong, J. Callan, and Z. Liu. (2018). “Convolutional neural networks for soft-matching n-grams in ad-hoc search”. In: *Proceedings of the eleventh ACM international conference on web search and data mining*. 126–134.

Dastin, J. (2018). “Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women”. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Dellinger, J. and S. Pell. (2024). “Bodies of evidence: The criminalization of abortion and surveillance of women in a post-Dobbs world”. *Duke Journal of Constitutional Law Public Policy*. 19(1): 1–108.

Deng, W. H., M. S. Lam, Á. A. Cabrera, D. Metaxa, M. Eslami, and K. Holstein. (2023). "Supporting user engagement in testing, auditing, and contesting AI". In: *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 556–559.

Dev, S., T. Li, J. M. Phillips, and V. Srikumar. (2020). "On measuring and mitigating biased inferences of word embeddings". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 05. 7659–7666.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. (2018a). "Bert: Pre-training of deep bidirectional transformers for language understanding". *arXiv preprint arXiv:1810.04805*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. (2018b). "Bert: Pre-training of deep bidirectional transformers for language understanding". *arXiv preprint arXiv:1810.04805*.

Diaz, F. and D. Metzler. (2006). "Improving the estimation of relevance models using large external corpora". In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 154–161.

Diaz, F., B. Mitra, M. D. Ekstrand, A. J. Biega, and B. Carterette. (2020). "Evaluating stochastic rankings with expected exposure". In: *Proceedings of the 29th ACM international conference on information & knowledge management*. 275–284.

DiMarco, M., H. Zhao, M. Boulicault, and S. S. Richardson. (2022). "Why "sex as a biological variable" conflicts with precision medicine initiatives". *Cell Reports Medicine*. 3(4). DOI: 10.1016/j.xcrm.2022.100550. URL: https://doi.org/10.1016/j.xcrm.2022.100550.

Dinan, E., A. Fan, A. Williams, J. Urbanek, D. Kiela, and J. Weston. (2019). "Queens are powerful too: Mitigating gender bias in dialogue generation". *arXiv preprint arXiv:1911.03842*.

Doughman, J., W. Khreich, M. El Gharib, M. Wiss, and Z. Berjawi. (2021). "Gender bias in text: Origin, taxonomy, and implications". In: *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. 34–44.

Douzas, G., F. Bacao, and F. Last. (2018). “Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE”. *Information sciences*. 465: 1–20.

Dubenko, E. (2022). “Across-language masculinity of oceans and femininity of guitars: Exploring grammatical gender universalities”. *Frontiers in Psychology*. 13. DOI: 10.3389/fpsyg.2022.1009966. URL: https://doi.org/10.3389/fpsyg.2022.1009966.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. (2012). “Fairness through awareness”. In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

Ekstrand, M. D., A. Das, R. Burke, and F. Diaz. (2021). “Fairness and Discrimination in Information Access Systems”. *CoRR*. abs/2105.05779. arXiv: 2105.05779. URL: https://arxiv.org/abs/2105.05779.

Ellemers, N. (2018). “Gender stereotypes”. *Annual review of psychology*. 69: 275–298.

Ellis, G. (2018). “So, what are cognitive biases?” *Cognitive biases in visualizations*: 1–10.

Eskandanian, F. and B. Mobasher. (2020). “Using stable matching to optimize the balance between accuracy and diversity in recommendation”. In: *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 71–79.

Fabris, A., A. Purpura, G. Silvello, and G. A. Susto. (2020). “Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms”. *Information Processing & Management*. 57(6): 102377.

Feast, J. (2020). “4 ways to address gender bias in AI”. URL: https://hbr.org/2019/11/4-ways-to-address-gender-bias-in-ai.

Fekih, S., N. Tamagnone, B. Minixhofer, R. Shrestha, X. Contla, E. Oglethorpe, and N. Rekabsaz. (2022). “Humset: Dataset of multilingual information extraction and classification for humanitarian crisis response”. *arXiv preprint arXiv:2210.04573*.

Felkner, V. K., H.-C. H. Chang, E. Jang, and J. May. (2023). “Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models”. *arXiv preprint arXiv:2306.15087*.

Fersini, E., D. Nozza, P. Rosso, *et al.* (2020). "AMI@ EVALITA2020: Automatic misogyny identification". In: *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020).* (seleziona...)

Fersini, E., P. Rosso, M. Anzovino, *et al.* (2018). "Overview of the task on automatic misogyny identification at IberEval 2018." *Ibereval@ sepln.* 2150: 214–228.

Fleisig, E. and C. Fellbaum. (2022). "Mitigating Gender Bias in Machine Translation through Adversarial Learning". *arXiv preprint arXiv:2203.10675.*

Frable, D. E. (1997). "Gender, racial, ethnic, sexual, and class identities". *Annual review of psychology.* 48(1): 139–162. DOI: 10.1146/annurev.psych.48.1.139. URL: https://doi.org/10.1146/annurev.psych.48.1.139.

Friedman, B. and H. Nissenbaum. (1996). "Bias in Computer Systems". *ACM Transactions on Information Systems.* 14(3): 330–347. DOI: 10.1145/230538.230561. URL: https://doi.org/10.1145/230538.230561.

Frost, D. M. (2011). "Social stigma and its consequences for the socially stigmatized". *Social and Personality Psychology Compass.* 5(11): 824–839. DOI: 10.1111/j.1751-9004.2011.00394.x. URL: https://doi.org/10.1111/j.1751-9004.2011.00394.x.

Garimella, A., A. Amarnath, K. Kumar, A. P. Yalla, N. Anandhavelu, N. Chhaya, and B. V. Srinivasan. (2021). "He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.* 4534–4545.

Gebru, T., J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, and K. Crawford. (2021). "Datasheets for datasets". *Communications of the ACM.* 64(12): 86–92. DOI: 10.1145/3458723. URL: https://doi.org/10.1145/3458723.

Ghabrial, M. A. (2019). ""We can shapeshift and build bridges": Bisexual women and gender diverse people of color on invisibility and embracing the borderlands". *Journal of Bisexuality.* 19(2): 169–197. DOI: 10.1080/15299716.2019.1617526. URL: https://doi.org/10.1080/15299716.2019.1617526.

Ghosh, B. (2018). “A diachronic perspective of Hijra identity in India”. In: *Sociology of Motherhood and Beyond*. 107–119.

Gill-Peterson, J. (2024). *A Short History of Trans Misogyny*. Verso Books.

Gonen, H. and Y. Goldberg. (2019). “Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them”. *arXiv preprint arXiv:1903.03862*.

González-Álvarez, J. and R. Sos-Peña. (2022). “Sex perception from facial structure: Categorization with and without skin texture and color”. *Vision Research*. 201: 108127. DOI: 10.1016/j.visres.2022.108127. URL: https://doi.org/10.1016/j.visres.2022.108127.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. (2020). “Generative adversarial networks”. *Communications of the ACM*. 63(11): 139–144.

Google. (2024). “Google Scholar”. URL: https://scholar.google.com.

Guo, J., Y. Fan, Q. Ai, and W. B. Croft. (2016). “A deep relevance matching model for ad-hoc retrieval”. In: *Proceedings of the 25th ACM international on conference on information and knowledge management*. 55–64.

Gurumurthy, S., R. Kiran Sarvadevabhatla, and R. Venkatesh Babu. (2017). “Deligan: Generative adversarial networks for diverse and limited data”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 166–174.

Hamidi, F., M. K. Scheuerman, and S. M. Branham. (2018). “Gender recognition or gender reductionism? The social implications of embedded gender recognition systems”. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

Holstein, K., J. W. Vaughan, H. Wallach, H. D. III, and M. Dudik. (2019). “Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?” In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16. URL: https://dl.acm.org/doi/10.1145/3290605.3300830.

Hong, L. and S. E. Page. (2004). “Groups of diverse problem solvers can outperform groups of high-ability problem solvers”. *Proceedings of the National Academy of Sciences*. 101(46): 16385–16389. DOI: 10.1073/pnas.0403723101. URL: https://doi.org/10.1073/pnas.0403723101.

Hyde, J. S., M. Krajnik, and K. Skuldt-Niederberger. (1991). “Androgyny across the life span: A replication and longitudinal followup”. *Developmental Psychology*. 27(3): 516–519. DOI: 10.1037/0012-1649.27.3.516. URL: https://doi.org/10.1037/0012-1649.27.3.516.

InterACT. (2018). “InterACT statement on Intersex Terminology”. URL: https://interactadvocates.org/interact-statement-on-intersex-terminology/.

Internet Live Stats. (2024). “Google Search Statistics”. URL: https://www.internetlivestats.com/google-search-statistics/.

ITHAKA. (2024). “JSTOR”. URL: https://www.jstor.org.

Jakiela, P. and O. Ozier. (2020). “Gendered language”. *Tech. rep.* Bonn: IZA Discussion Papers, No. 13126, Institute of Labor Economics (IZA). URL: https://www.econstor.eu/bitstream/10419/216438/1/dp13126.pdf.

Jha, A. and R. Mamidi. (2017). “When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data”. In: *Proceedings of the second workshop on NLP and computational social science*. 7–16.

Johnson, J. L. and R. Repta. (2012). “Sex and gender: Beyond the binaries”. In: *Designing and conducting gender, sex and health research*. Ed. by J. L. Oliffe and L. Greaves. Thousand Oaks, CA: Sage. 17–37.

Jones, S. (1975). “Report on the need for and provision of an" ideal" information retrieval test collection”.

Kamiran, F. and T. Calders. (2012). “Data preprocessing techniques for classification without discrimination”. *Knowledge and information systems*. 33(1): 1–33. DOI: 10.1007/s10115-011-0463-8. URL: https://doi.org/10.1007/s10115-011-0463-8.

Kaneko, M., D. Bollegala, and N. Okazaki. (2022). “Gender bias in meta-embeddings”. *arXiv preprint arXiv:2205.09867*.

Kapania, S., A. S. Taylor, and D. Wang. (2023). “A hunt for the snark: Annotator diversity in data practices”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

Karpukhin, V., B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. (2020). “Dense passage retrieval for open-domain question answering”. *arXiv preprint arXiv:2004.04906*.

Karray, F., M. Alemzadeh, J. A. Saleh, and M. N. Arab. (2008). “Human-computer interaction: Overview on state of the art”. *International Journal on Smart Sensing and Intelligent Systems*. 1(1): 137–159. DOI: 10.21307/ijssis-2017-283. URL: https://doi.org/10.21307/ijssis-2017-283.

Katyal, S. K. and J. Y. Jung. (2021). “The gender panopticon: AI, gender, and design justice”. *UCLA Law Review*. 68: 692–785.

Keyes, O. (2018). “The misgendering machines: Trans/HCI implications of automatic gender recognition”. In: *Proceedings of the ACM on human-computer interaction*. 1–22.

Khan, S. I., M. I. Hussain, S. Parveen, M. I. Bhuiyan, G. Gourab, G. F. Sarker, S. M. Arafat, and J. Sikder. (2009). “Living on the extreme margin: Social exclusion of the transgender population (hijra) in Bangladesh”. *Journal of Health, Population and Nutrition*. 27(4). DOI: 10.3329/jhpn.v27i4.3388. URL: https://doi.org/10.3329/jhpn.v27i4.3388.

Klasnja, A., N. Arabzadeh, M. Mehrvarz, and E. Bagheri. (2022). “On the characteristics of ranking-based gender bias measures”. In: *Proceedings of the 14th ACM Web Science Conference 2022*. 245–249.

Kleisner, K., P. Tureček, S. C. Roberts, J. Havlíček, J. V. Valentova, R. M. Akoko, J. D. Leongómez, S. Apostol, M. A. Varella, and S. A. Saribay. (2021). “How and why patterns of sexual dimorphism in human faces vary across the world”. *Scientific reports*. 11(1): 5978. DOI: 10.1038/s41598-021-85402-3. URL: https://doi.org/10.1038/s41598-021-85402-3.

Kopeinik, S., M. Mara, L. Ratz, K. Krieg, M. Schedl, and N. Rekabsaz. (2023). “Show me a" Male Nurse"! How Gender Bias is Reflected in the Query Formulation of Search Engine Users”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

Kotek, H., R. Dockum, and D. Sun. (2023). "Gender bias and stereotypes in large language models". In: *Proceedings of The ACM Collective Intelligence Conference*. 12–24.

Krieg, K., E. Parada-Cabaleiro, G. Medicus, O. Lesota, M. Schedl, and N. Rekabsaz. (2023). "Grep-BiasIR: A Dataset for Investigating Gender Representation Bias in Information Retrieval Results". In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 444–448.

Krieg, K., E. Parada-Cabaleiro, M. Schedl, and N. Rekabsaz. (2022). "Do Perceived Gender Biases in Retrieval Results Affect Relevance Judgements?" arXiv: 2203.01731 [cs.IR].

Krieger, N. (2020). "Measures of racism, sexism, heterosexism, and gender binarism for health equity research: from structural injustice to embodied harm—an ecosocial analysis". *Annual Review of Public Health*. 41(1): 37–62. DOI: 10.1146/annurev-publhealth-040119-094017. URL: https://doi.org/10.1146/annurev-publhealth-040119-094017.

Kruks, S. (1992). "Gender and subjectivity: Simone de Beauvoir and contemporary feminism". *Signs: Journal of Women in Culture and Society*. 18(1): 89–110. DOI: 10.1086/494780. URL: https://doi.org/10.1086/494780.

L.P., B. (2024). "Bloomberg". URL: https://www.bloomberg.com.

Lam, M. S., M. L. Gordon, D. Metaxa, J. T. Hancock, J. A. Landay, and M. S. Bernstein. (2022). "End-user audits: A system empowering communities to lead large-scale investigations of harmful algorithmic behavior". In: *Proceedings of the ACM on Human-Computer Interaction*. 1–34.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. (2020). "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=H1eA7AEtvS.

Laqueur, T. W. (1992). *Making sex: Body and gender from the Greeks to Freud*. Harvard University Press.

Lee, P. A., C. P. Houk, S. F. Ahmed, and I. A. Hughes. (2006). "Consensus statement on management of intersex disorders". *Pediatrics*. 118(2). DOI: 10.1542/peds.2006-0738. URL: https://doi.org/10.1542/peds.2006-0738.

Lemley, J., S. Bazrafkan, and P. Corcoran. (2017). "Smart augmentation learning an optimal data augmentation strategy". *Ieee Access*. 5: 5858–5869.

LexisNexis. (2024). "LexisNexis". URL: https://www.lexisnexis.com.

Li, Y., X. Wei, Z. Wang, S. Wang, P. Bhatia, X. Ma, and A. Arnold. (2022). "Debiasing Neural Retrieval via In-batch Balancing Regularization". In: *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Seattle, Washington: Association for Computational Linguistics. 58–66. DOI: 10.18653/v1/2022.gebnlp-1.5. URL: https://aclanthology.org/2022.gebnlp-1.5.

Liu, H., W. Wang, Y. Wang, H. Liu, Z. Liu, and J. Tang. (2020). "Mitigating gender bias for neural dialogue generation with adversarial learning". *arXiv preprint arXiv:2009.13028*.

Lohia, P. K., K. Natesan Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri. (2019). "Bias mitigation post-processing for individual and group fairness". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DOI: 10.1109/icassp.2019.8682620. URL: https://doi.org/10.1109/icassp.2019.8682620.

Loideain, N. N. and R. Adams. (2020). "From Alexa to Siri and the GDPR: The gendering of virtual personal assistants and the role of Data Protection Impact Assessments". *Computer Law & Security Review*. 36: 105366. DOI: 10.1016/j.clsr.2019.105366. URL: https://doi.org/10.1016/j.clsr.2019.105366.

Manning, C. D. (2008). *Introduction to information retrieval*. Syngress Publishing,

Matthew, E. (2018). "Peters, mark neumann, mohit iyyer, matt gardner, christopher clark, kenton lee, luke zettlemoyer. deep contextualized word representations". In: *Proc. of NAACL*. Vol. 5.

Maudslay, R. H., H. Gonen, R. Cotterell, and S. Teufel. (2019). "It's all in the name: Mitigating gender bias with name-based counterfactual data substitution". *arXiv preprint arXiv:1909.00871*.

May, C., A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger. (2019). "On measuring social biases in sentence encoders". *arXiv preprint arXiv:1903.10561*.

Mazzuca, C., A. M. Borghi, S. van Putten, L. Lugli, R. Nicoletti, and A. Majid. (2023). "Gender is conceptualized in different ways across cultures". *Language and Cognition*. 16(2): 353–379. DOI: 10.1017/langcog.2023.40. URL: https://doi.org/10.1017/langcog.2023.40.

McLemore, K. A. (2014). "Experiences with misgendering: Identity misclassification of transgender spectrum individuals". *Self and Identity*. 14(1): 51–74. DOI: 10.1080/15298868.2014.950691. URL: https://doi.org/10.1080/15298868.2014.950691.

Medicine, U. N. L. of. (2024). "PubMed". URL: https://pubmed.ncbi.nlm.nih.gov.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. (2021). "A Survey on Bias and Fairness in Machine Learning". *ACM Computing Surveys*. 54(6): 1–35. DOI: 10.1145/3457607. URL: https://doi.org/10.1145/3457607.

Meinhardt-Injac, B., M. Persike, and G. Meinhardt. (2013). "Holistic face processing is induced by shape and texture". *Perception*. 42(7): 716–732. DOI: 10.1068/p7462. URL: https://doi.org/10.1068/p7462.

Memarian, B. and T. Doleck. (2023). "Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review". *Computers and Education: Artificial Intelligence*. 5: 100152. DOI: 10.1016/j.caeai.2023.100152. URL: https://doi.org/10.1016/j.caeai.2023.100152.

Miceli, M., T. Yang, A. A. Garcia, J. Posada, S. M. Wang, M. Pohl, and A. Hanna. (2022). "Documenting Data Production Processes: A Participatory Approach for Data Work". arXiv: 2207.04958 `[cs.HC]`. URL: https://arxiv.org/abs/2207.04958.

Mikołajczyk, A. and M. Grochowski. (2018). "Data augmentation for improving deep learning in image classification problem". In: *2018 international interdisciplinary PhD workshop (IIPhDW)*. IEEE. 117–122.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. (2013). "Efficient estimation of word representations in vector space". *arXiv preprint arXiv:1301.3781*.

Mitra, B., F. Diaz, and N. Craswell. (2017). “Learning to match using local and distributed representations of text for web search”. In: *Proceedings of the 26th international conference on world wide web*. 1291–1299.

Mort, J. A. (2023). “Fighting information termination”. *Index on Censorship*. 52(1): 27–29.

Namaste, V. (2000). *Invisible lives: The erasure of transsexual and transgendered people*. University of Chicago Press.

Nass, C., J. Steuer, and E. R. Tauber. (1994). “Computers are social actors”. In: *Conference Companion on Human Factors in Computing Systems - CHI ’94*. DOI: 10.1145/259963.260288. URL: https://doi.org/10.1145/259963.260288.

National Public Radio, N. P. R. (2024). “NPR’s embedded and CBC tackle sex testing in elite sports with “tested” podcast”. URL: https://www.npr.org/2024/07/12/g-s1-8943/npr-embedded-cbc-testing-in-elite-sports-tested-podcast.

Nguyen, T., M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. (2016). “Ms marco: A human-generated machine reading comprehension dataset”.

Nicas, J. (2019). “Apple Card Investigated After Gender Discrimination Complaints”. URL: https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html.

Niousha, R., D. Saito, H. Washizaki, and Y. Fukazawa. (2023). “Investigating the Effect of Binary Gender Preferences on Computational Thinking Skills”. *Education Sciences*. 13(5): 433.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan. (2019). “Dissecting racial bias in an algorithm used to manage the health of populations”. *Science*. 366(6464): 447–453. DOI: 10.1126/science.aax2342. URL: https://science.sciencemag.org/content/366/6464/447.

Osokin, A., A. Chessel, R. E. Carazo Salas, and F. Vaggi. (2017). “GANs for biological image synthesis”. In: *Proceedings of the IEEE international conference on computer vision*. 2233–2242.

Ovalle, A., A. Subramonian, A. Singh, C. Voelcker, D. J. Sutherland, D. Locatelli, E. Breznik, F. Klubicka, H. Yuan, J. Hetvi, H. Zhang, J. Shriram, K. Lehman, L. Soldaini, M. Sap, M. P. Deisenroth, M. L. Pacheco, M. Ryskina, M. Mundt, M. Agarwal, N. Mclean, P. Xu, A. Pranav, R. Korpan, R. Ray, S. Mathew, S. Arora, S. John, T. Anand, V. Agrawal, W. Agnew, Y. Long, Z. J. Wang, Z. Talat, A. Ghosh, N. Dennler, M. Noseworthy, S. Jha, E. Baylor, A. Joshi, N. Y. Bilenko, A. Mcnamara, R. Gontijo-Lopes, A. Markham, E. Dong, J. Kay, M. Saraswat, N. Vytla, and L. Stark. (2023). "Queer in AI: A case study in community-led participatory AI". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1882–1895.

Pang, L., Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng. (2016). "Text matching as image recognition". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 30. No. 1.

Pape, M., M. Miyagi, S. A. Ritz, M. Boulicault, S. S. Richardson, and D. L. Maney. (2024). "Sex contextualism in laboratory research: enhancing rigor and precision in the study of sex-related variables". *Cell*. 187(6): 1316–1326. DOI: 10.1016/j.cell.2024.02.008. URL: https://doi.org/10.1016/j.cell.2024.02.008.

Park, S., K. Choi, H. Yu, and Y. Ko. (2023). "Never too late to learn: Regularizing gender bias in coreference resolution". In: *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 15–23.

Pennebaker, J. W., M. E. Francis, and R. J. Booth. (2001a). "Linguistic inquiry and word count: LIWC 2001". *Mahway: Lawrence Erlbaum Associates*. 71(2001): 2001.

Pennebaker, J. W., M. E. Francis, and R. J. Booth. (2001b). *Linguistic Inquiry and Word Count*. Mahwah, NJ: Lawerence Erlbaum Associates.

Pennington, J., R. Socher, and C. D. Manning. (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

Perez, L. and J. Wang. (2017). “The effectiveness of data augmentation in image classification using deep learning”. *arXiv preprint arXiv:1712.04621*.

Perugia, G. and D. Lisy. (2023). “Robot’s gendering trouble: A scoping review of gendering humanoid robots and its effects on HRI”. *International Journal of Social Robotics*. 15(11): 1725–1753. DOI: 10.1007/s12369-023-01061-6. URL: https://doi.org/10.1007/s12369-023-01061-6.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. (2018a). “Deep Contextualized Word Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by M. Walker, H. Ji, and A. Stent. New Orleans, Louisiana: Association for Computational Linguistics. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. (2018b). “Deep contextualized word representations”. *CoRR*. abs/1802.05365. arXiv: 1802.05365. URL: http://arxiv.org/abs/1802.05365.

Petreski, D. and I. C. Hashim. (2022). “Word embeddings are biased. But whose bias are they reflecting?” *AI & Society*. 38(2): 975–982. DOI: 10.1007/s00146-022-01443-w. URL: https://doi.org/10.1007/s00146-022-01443-w.

Pinney, C., A. Raj, A. Hanna, and M. D. Ekstrand. (2023). “Much Ado About Gender: Current Practices and Future Recommendations for Appropriate Gender-Aware Information Access”. In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 269–279.

Pleiss, G., M. Raghavan, F. Wu, J. M. Kleinberg, and K. Q. Weinberger. (2017). “On fairness and calibration”. In: *The Societal Impacts of Algorithmic Decision-Making*.

Posada, J. (2023). “Platform Authority and Data Quality: Who Decides What Counts in Data Production for Artificial Intelligence”. *Tech. rep.* Technical Report. Berggruen Institute and Global Affairs Canada.

Prost, F., N. Thain, and T. Bolukbasi. (2019). "Debiasing embeddings for reduced gender bias in text classification". *arXiv preprint arXiv:1908.02810*.

Qian, Y., U. Muaz, B. Zhang, and J. W. Hyun. (2019). "Reducing gender bias in word-level language models with a gender-equalizing loss function". *arXiv preprint arXiv:1905.12801*.

Qiu, H., Z.-Y. Dou, T. Wang, A. Celikyilmaz, and N. Peng. (2023). "Gender Biases in Automatic Evaluation Metrics for Image Captioning". In: *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Qu, Y., Y. Ding, J. Liu, K. Liu, R. Ren, W. X. Zhao, D. Dong, H. Wu, and H. Wang. (2021). "RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5835–5847.

Raj, A., B. Mitra, N. Craswell, and M. Ekstrand. (2023). "Patterns of gender-specializing query reformulation". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2241–2245.

Rekabsaz, N., S. Kopeinik, and M. Schedl. (2021). "Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation for BERT Rankers". *arXiv preprint arXiv:2104.13640*.

Rekabsaz, N. and M. Schedl. (2020). "Do Neural Ranking Models Intensify Gender Bias?" In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2065–2068.

Reuters, T. (2024). "Reuters". URL: https://www.reuters.com.

Richards, J. and R. Hawley. (2011). "Sex determination: How genes determine a developmental choice". *The human genome*: 273–98. DOI: 10.1016/b978-0-08-091865-5.00008-4. URL: https://doi.org/10.1016/b978-0-08-091865-5.00008-4.

Richardson, S. S. (2022). "Sex contextualism". *Philosophy, Theory, and Practice in Biology*. 14(0). DOI: 10.3998/ptpbio.2096. URL: https://doi.org/10.3998/ptpbio.2096.

Roberts, T. K. and C. R. Fantz. (2014). “Barriers to quality health care for the transgender population”. *Clinical Biochemistry*. 47(10–11): 983–987. DOI: 10.1016/j.clinbiochem.2014.02.009. URL: https://doi.org/10.1016/j.clinbiochem.2014.02.009.

Robertson, S., H. Zaragoza, *et al.* (2009). “The probabilistic relevance framework: BM25 and beyond”. *Foundations and Trends® in Information Retrieval*. 3(4): 333–389.

Rodríguez-Sánchez, F., J. Carrillo-de-Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso. (2021). “Overview of exist 2021: sexism identification in social networks”. *Procesamiento del Lenguaje Natural*. 67: 195–207.

Rodríguez-Sánchez, F., J. Carrillo-de-Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, and P. Rosso. (2022). “Overview of exist 2022: sexism identification in social networks”. *Procesamiento del Lenguaje Natural*. 69: 229–240.

Rule, N. O. (2017). “Perceptions of sexual orientation from minimal cues”. *Archives of Sexual Behavior*. 46(1): 129–139. DOI: 10.1007/s10508-016-0779-2. URL: https://doi.org/10.1007/s10508-016-0779-2.

Russell, R. (2009). “A sex difference in facial contrast and its exaggeration by cosmetics”. *Perception*. 38(8): 1211–1219. DOI: 10.1068/p6331. URL: https://doi.org/10.1068/p6331.

Russell, R., P. Sinha, I. Biederman, and M. Nederhouser. (2006). “Is pigmentation important for face recognition? Evidence from contrast negation”. *Perception*. 35(6): 749–759. DOI: 10.1068/p5490. URL: https://doi.org/10.1068/p5490.

Samory, M., I. Sen, J. Kohne, F. Flöck, and C. Wagner. (2021). ““Call me sexist, but...”: Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples”. In: *Proceedings of the international AAAI conference on web and social media*. Vol. 15. 573–584.

Sandfort, T. G., H. M. Bos, T.-C. Fu, D. Herbenick, and B. Dodge. (2021). “Gender expression and its correlates in a nationally representative sample of the US adult population: Findings from the National Survey of Sexual Health and Behavior”. *The Journal of Sex Research*. 58(1): 51–63. DOI: 10.1080/00224499.2020.1818178. URL: https://doi.org/10.1080/00224499.2020.1818178.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. (2019). “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. *arXiv preprint arXiv:1910.01108*.

Sapir, E. (2023). *Selected writings of Edward Sapir in language, culture and personality*. Univ of California Press.

Saunders, D. and B. Byrne. (2020). “Reducing gender bias in neural machine translation as a domain adaptation problem”. *arXiv preprint arXiv:2004.04498*.

Saxena, S. and S. Jain. (2024). “Exploring and mitigating gender bias in book recommender systems with explicit feedback”. *Journal of Intelligent Information Systems*: 1–22.

Scheuerman, M. K., M. Pape, and A. Hanna. (2021). “Auto-essentialization: Gender in automated facial analysis as extended colonial project”. *Big Data Society*. 8(2): 1–15.

Scheuerman, M. K., J. M. Paul, and J. R. Brubaker. (2019). “How computers see gender”. *Proceedings of the ACM on Human-Computer Interaction*. 3(CSCW): 1–33. DOI: 10.1145/3359246. URL: https://doi.org/10.1145/3359246.

Seyedsalehi, S., A. Bigdeli, N. Arabzadeh, B. Mitra, M. Zihayat, and E. Bagheri. (2022a). “Bias-aware Fair Neural Ranking for Addressing Stereotypical Gender Biases.” In: *EDBT*. 2–435.

Seyedsalehi, S., A. Bigdeli, N. Arabzadeh, M. Zihayat, and E. Bagheri. (2022b). “Addressing gender-related performance disparities in neural rankers”. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2484–2488.

Shea, M. (2023). “The sinister side to female robots?” URL: https://www.bbc.com/future/article/20230804-is-there-a-sinister-side-to-the-rise-of-female-robots.

Sheng, E., K.-W. Chang, P. Natarajan, and N. Peng. (2021). "Societal biases in language generation: Progress and challenges". *arXiv preprint arXiv:2105.04054*.

Shin, D., A. Rasul, and A. Fotiadis. (2021). "Why am I seeing this? Deconstructing algorithm literacy through the lens of users". *Internet Research*. 32(4): 1214–1234. DOI: 10.1108/intr-02-2021-0087. URL: https://doi.org/10.1108/intr-02-2021-0087.

Shneiderman, B. (2020). "Human-centered artificial intelligence: Reliable, safe & trustworthy". *International Journal of Human–Computer Interaction*. 36(6): 495–504. DOI: 10.1080/10447318.2020.1741118. URL: https://doi.org/10.1080/10447318.2020.1741118.

Shorten, C. and T. M. Khoshgoftaar. (2019). "A survey on image data augmentation for deep learning". *Journal of big data*. 6(1): 1–48.

Shrestha, S. and S. Das. (2022). "Exploring gender biases in ML and AI academic research through systematic literature review". *Frontiers in artificial intelligence*. 5: 976838.

Smith, J. J. and L. Beattie. (2022). "RecSys Fairness Metrics: Many to Use But Which One To Choose?" *arXiv preprint arXiv:2209.04011*.

Smith, J. M. (2021). "Beyond the Gender Binary in Computing Education Research". In: *Proceedings of the 17th ACM Conference on International Computing Education Research*. 444–445.

Stanovsky, G., N. A. Smith, and L. Zettlemoyer. (2019). "Evaluating gender bias in machine translation". *arXiv preprint arXiv:1906.00591*.

Steck, H. (2011). "Item popularity and recommendation accuracy". In: *Proceedings of the fifth ACM conference on Recommender systems*. 125–132.

Sun, T., J. He, X. Qiu, and X. Huang. (2022). "BERTScore is Unfair: On Social Bias in Language Model-Based Metrics for Text Generation". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Goldberg, Z. Kozareva, and Y. Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. 3726–3739. DOI: 10.18653/v1/2022.emnlp-main.245. URL: https://aclanthology.org/2022.emnlp-main.245.

Sundararaman, D. and V. Subramanian. (2022). "Do Information Retrieval Models Exhibit Gender Bias?"

Surveillance Technology Oversight Project, I. (2022). “STOP: Surveillance Technology Oversight Project”. URL: https://www.stopspying.org.

Svechnikov, K. and O. Söder. (2008). “Ontogeny of gonadal sex steroids”. *Best Practice & Research Clinical Endocrinology & Metabolism*. 22(1): 95–106. DOI: 10.1016/j.beem.2007.09.002. URL: https://doi.org/10.1016/j.beem.2007.09.002.

Swim, J., E. Borgida, G. Maruyama, and D. G. Myers. (1989). “Joan McKay versus John McKay: Do gender stereotypes bias evaluations?” *Psychological Bulletin*. 105(3): 409.

Tadiri, C. P., V. Raparelli, M. Abrahamowicz, A. Kautzy-Willer, K. Kublickiene, M.-T. Herrero, C. M. Norris, L. Pilote, and G. Consortium. (2021). “Methods for prospectively incorporating gender into health sciences research”. *Journal of Clinical Epidemiology*. 129: 191–197.

Taksa, I. and J. M. Flomenbaum. (2009). “An integrated framework for research on cross-cultural information retrieval”. In: *2009 Sixth International Conference on Information Technology: New Generations*. IEEE. 1367–1372.

Tang, K., W. Zhou, J. Zhang, A. Liu, G. Deng, S. Li, P. Qi, W. Zhang, T. Zhang, and N. Yu. (2024). “GenderCARE: A Comprehensive Framework for Assessing and Reducing Gender Bias in Large Language Models”. *arXiv preprint arXiv:2408.12494*.

Tannenbaum, C., R. P. Ellis, F. Eyssel, J. Zou, and L. Schiebinger. (2019). “Sex and gender analysis improves science and engineering”. *Nature*. 575(7781): 137–146. DOI: 10.1038/s41586-019-1657-6. URL: https://doi.org/10.1038/s41586-019-1657-6.

Te’eni, D., J. Carey, and P. Zhang. (2007). *Human-Computer Interaction: Developing Effective Organizational Information Systems*. John Wiley & Sons, Inc.

Tolmeijer, S., N. Zierau, A. Janson, J. S. Wahdatehagh, J. M. Leimeister, and A. Bernstein. (2021). “Female by default? – Exploring the effect of voice assistant gender and pitch on trait and trust attribution”. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. DOI: 10.1145/3411763.3451623. URL: https://doi.org/10.1145/3411763.3451623.

UNESCO, I. (2024). “Challenging systematic prejudices: An investigation into gender bias in large language models”. URL: https://unesdoc.unesco.org/ark:/48223/pf0000388971.

University, U. N. (2019). “Taking Stock: Data Evidence on Gender Equality in Digital Access, Skills, and Leadership”. URL: https://collections.unu.edu/eserv/UNU:7350/EQUALS-Research-Report-2019.pdf.

Urbanek, J., A. Fan, S. Karamcheti, S. Jain, S. Humeau, E. Dinan, T. Rocktäschel, D. Kiela, A. Szlam, and J. Weston. (2019). “Learning to speak and act in a fantasy text adventure game”. *arXiv preprint arXiv:1903.03094*.

Voorhees, E. M. (2004). “Overview of the TREC 2004 robust retrieval track”. In: *TREC*. Vol. 2004. No. 5. 3.

Wang, J., Y. Liu, and X. E. Wang. (2021). “Are gender-neutral queries really gender-neutral? mitigating gender bias in image search”. *arXiv preprint arXiv:2109.05433*.

Wang, Y. and M. Kosinski. (2018). “Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.” *Journal of personality and social psychology*. 114(2): 246.

Waseem, Z. and D. Hovy. (2016). “Hateful symbols or hateful people? predictive features for hate speech detection on twitter”. In: *Proceedings of the NAACL student research workshop*. 88–93.

Weischedel, R., S. Pradhan, L. Ramshaw, J. Kaufman, M. Franchini, M. ElBachouti, N. Xue, M. Palmer, J. D. Hwang, C. Bonial, *et al.* (2012). “OntoNotes Release 5.0”. URL: https://catalog.ldc.upenn.edu/LDC2013T19.

Westlaw. (2024). “Westlaw”. URL: https://legal.thomsonreuters.com/en/westlaw.

Whorf, B. L. (2012). *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. MIT Press.

Wu, C., F. Wu, X. Wang, Y. Huang, and X. Xie. (2021). “Fairness-aware news recommendation with decomposed adversarial learning”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. No. 5. 4462–4469.

Xiong, C., Z. Dai, J. Callan, Z. Liu, and R. Power. (2017). "End-to-end neural ad-hoc ranking with kernel pooling". In: *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*. 55–64.

Yang, J., A. A. Soltan, D. W. Eyre, Y. Yang, and D. A. Clifton. (2023). "An adversarial training framework for mitigating algorithmic biases in clinical machine learning". *NPJ Digital Medicine*. 6(1): 55.

Yi, X., E. Walia, and P. Babyn. (2019). "Generative adversarial network in medical imaging: A review". *Medical image analysis*. 58: 101552.

Yin, H., B. Cui, J. Li, J. Yao, and C. Chen. (2012). "Challenging the long tail recommendation". *arXiv preprint arXiv:1205.6700*.

Young, E., J. Wajcman, and L. Sprejer. (2023). "Mind the gender gap: Inequalities in the emergent professions of artificial intelligence (AI) and data science". *New Technology, Work and Employment*. 38(3): 391–414. DOI: 10.1111/ntwe.12278. URL: https://doi.org/10.1111/ntwe.12278.

Zehlike, M., K. Yang, and J. Stoyanovich. (2022). "Fairness in ranking, part i: Score-based ranking". *ACM Computing Surveys*. 55(6): 1–36.

Zemel, R. S., L. Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. (2013). "Learning fair representations". In: *International Conference on Machine Learning*.

Zerveas, G., N. Rekabsaz, D. Cohen, and C. Eickhoff. (2021). "CODER: An efficient framework for improving retrieval through COntextual Document Embedding Reranking". *arXiv preprint arXiv:2112.08766*.

Zhang, D. T., S. Mishra, E. Brynjolfsson, J. Etchemendy, D. Ganguli, B. Grosz, T. Lyons, J. Manyika, J. Niebles, M. Sellitto, Y. Shoham, J. Clark, and R. Perrault. (2021). "The AI Index 2021 Annual Report". *ArXiv*. abs/2103.06312.

Zhang, G., S. Ananiadou, *et al.* (2022). "Examining and mitigating gender bias in text emotion detection task". *Neurocomputing*. 493: 422–434.

Zhao, J., S. Mukherjee, S. Hosseini, K.-W. Chang, and A. H. Awadallah. (2020). "Gender bias in multilingual embeddings and cross-lingual transfer". *arXiv preprint arXiv:2005.00699*.

Zhao, J., T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang. (2019). “Gender bias in contextualized word embeddings”. *arXiv preprint arXiv:1904.03310*.

Zhao, J., T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. (2017). “Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2979–2989. URL: https://aclanthology.org/D17-1323/.

Zhao, J., T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. (2018a). “Gender bias in coreference resolution: Evaluation and debiasing methods”. *arXiv preprint arXiv:1804.06876*.

Zhao, J., Y. Zhou, Z. Li, W. Wang, and K.-W. Chang. (2018b). “Learning Gender-Neutral Word Embeddings”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii. Brussels, Belgium: Association for Computational Linguistics. 4847–4853. DOI: 10.18653/v1/D18-1521. URL: https://aclanthology.org/D18-1521.

Zhao, J., Y. Zhou, Z. Li, W. Wang, and K.-W. Chang. (2018c). “Learning gender-neutral word embeddings”. *arXiv preprint arXiv:1809.01496*.

Zimman, L. (2014). “The discursive construction of sex”. In: *Queer Excursions: Retheorizing Binaries in Language, Gender, and Sexuality*. Oxford University Press. 13–34.

Zou, J. Y., D. J. Hsu, D. C. Parkes, and R. P. Adams. (2013). “Contrastive learning using spectral methods”. *Advances in Neural Information Processing Systems*. 26.