Bias in Language Models: Defining, Measuring, and Reducing Bias

Chaewon Yun

A thesis presented for the degree of
Master of Science

Computational Social Systems
Rhine-Westphalia Technical University of Aachen
Germany
June 6, 2023

First Supervisor: Dr. phil. habil. Jan-Christoph Heilinger
Second Supervisor: Univ.-Prof. Dr. techn. Claudia Wagner

Contents

Abstract A central question of this thesis asks how bias in language models is defined, measured, and ultimately, can be reduced. The sequence of inquiries are crucial to address potential harms and risks caused by language models. Specifically, 19 papers that suggest methods for identifying and quantifying gender bias in language models will be used as concrete references to answer the central question.

In the last few years, there has been a surge of interest in investigating bias in language models. Growing interest in the bias of language models was followed by empirical evidence showing socially and morally undesirable outputs that language models produce. Ultimately, identifying and measuring bias aims to mitigate undesirable bias. However, reviewed literature shows that there exists a mismatch between motivation, definition, and measurement of bias. Such disparity runs a risk of misinterpreting results, and fails to evaluate bias in language models adequately. Mitigating bias based on invalid measurements will inevitably lead to an unsatisfactory result.

Many problems in measuring bias stem from the underspecification of bias. The great part of the literature on bias in language models lacks clarity regarding the definition of bias. The meaning of bias is conflated and used as an umbrella term to refer to heterogeneous undesirable properties. Bias is used to refer both descriptive and normative correctness.

A closely connected problem with the conflated definition is a failure in operationalizing bias in a scientifically valid way. Properties used for measuring bias lack theoretical grounding that justifies the connection between what is being measured and what researchers aim (and claim) to measure. Partly due to the conflated use of the term, it is not a trivial task to identify which bias matters among various measured biases. I argue that bias that is wrongful discrimination should be prioritized. Wrongful discrimination is demeaning and fails to treat people equally. Bias that wrongfully discriminates matters, as it risks scaling existing social injustice via technological means.

Eventually, I will combine these arguments in favor of identifying bias that are a case of wrongful discrimination into a unified framework which aims to provide guidance to users and developers of language models. The framework aims to evaluate bias in language models holistically beyond the technical terms. The framework consists of three levels: conceptualization of bias, empirical evaluation of algorithms, and structural evaluation.

In conclusion, I draw an analogy between bias in language models and ‘‘rational’’ sexism, racism, or any other -ism that cannot be morally justified, despite the purported epistemic reasons. Bias in language models is supported by the statistical patterns that also power language models' technical capability, which makes it difficult to address them without compromising the language model's capacity. However, considering the given reality where morally problematic patterns are embedded in society and data, ensuring descriptive accuracy is not enough to make language models unbiased. Normative correctness should be addressed as a quality criteria to evaluate language models.

Chapter 1

Introduction

1.1 Relevance of Bias in Language Models

Language models used everywhere, from collecting debt [1] to finding a flat in Berlin [2]. Language models are changing how people learn information and interact with the internet. For the last several years, language models have been used in various applications from machine translation to auto-completion of search keywords. However, a so-called 'AI arms race' [3] has expedited commercialization of the technology even further. The success of language models have pushed the biggest companies to invest and develop bigger and better models.

In the simplest terms, a language model (LM) is an algorithm calculating the probability of the next word in a given context. It produces an output, part of words, words, or phrases, that have the highest probability given the input [4]. The most prevailing type of language model produces output texts based on input texts, also referred to as prompts. This setup allows language models to be easily adapted for a range of diverse applications. For instance, an automated sentence completion takes previously written words as an input and produces output of the most probable coming words, which can be used for composing any text, such as an email or search query. Language models can be used for question and answering, where the question works as an input and the answer to that question is produced as an output, since the most probable sequence following the question would be the corresponding answer.

Language models are not free from intrinsic problems of any data-driven learning-based technologies. The most well-known problem is that they reproduce problematic patterns in the data. For instance, a popular application based on a language model, ChatGPT [5], has produced a python script that categorizes one as a good scientist if they are "white" and "male" [6]. Moreover, not only the data, but the whole structure surrounding the technology is the source of problems in bias. Most language models are developed according to the value systems in the U.S., and by handful of private enterprises.

1.2 Systematic Literature Review

This thesis aims to critically discuss methods used for measuring bias by systematically reviewing how gender bias in language model is measured. Specifically, 19 papers suggesting novel methods to measure gender bias in language models were selected for the in-depth analysis. These papers will be used as a reference throughout the thesis as empirical cases of studying bias in language models. While the focus of the review is on gender bias, the critical analysis applies, not only to gender bias but also other types of bias such as racial bias.

A systematic literature review helps understanding the state of the art in the field. Appropriate selection of papers is an essential process for the review's validity and explanatory power [7]. I retrieved relevant papers from four databases: Scopus, Web of Science, ACM Library, and ACL Anthology. Afterwards, I screened retrieved papers in four rounds to identify papers that (1) measure gender bias in language models, and (2) provide a novel method to measure bias differing from the previously proposed method. To conduct an extensive literature search, query for retrieving papers included keywords for a broader scope of the topic, such as social bias and sentence encoders. Since the focus of my thesis is evaluating bias within the language model, papers using language models for analyzing bias in other sources such as news [8] were excluded. The full list of papers reviewed and a detailed process of the literature review can be found in the Appendix A.

1.3 Relevant work

This thesis critically reviews how bias in language models is defined, measured, and can be mitigated. By systematically review literature on measuring bias in language models, 19 papers were chosen that suggest novel methods for measuring gender bias in language models. The issue of bias in NLP has received considerable critical attention in the last several years. Furthermore, recent trends in language models have led to a proliferation of studies that focus on bias and language models. Especially, the concepts of bias and fairness are central to normative investigations of algorithms. Understanding the ethical and social implications of language models is vitally important as these technologies are quickly becoming ubiquitous.

1.3.1 Bias and Sociotehcnical Systems

In the last few years, there has been a surge of interest in bias in NLP systems, from word embeddings [9][10] to large language models [11]. What language models can do for good and bad are now at the center of the debate of technology and society [12][13][14][15].

While recent rapid development and commercialization of language models is a new phenomenon that differs from previous computer systems, the introduction of sociotech-

nical system into society and investigating its impact is not a novel topic. Computer systems have replaced tasks that were previously carried out by human in the 20th century. The renewed interest in bias in NLP systems resembles the critical work on sociotechnical systems. Winner(1980)[16] discussed how technology embodies power and authority. Friedman and Nissenbaum (1996) discussed bias in computer systems [17]. Value and design [18][19] explores how system designs can reflect values like fairness and trust. Following the question posed by Winner (1980), Lazovich (2020)[20] showed how deep learning reflects power and authority as well as other technical systems.

1.3.2 Literature Review on Bias in NLP

Several literature reviews provide an overview of studying bias in NLP, including language models, with a varying foci. Some highlight specific types of bias, such as gender bias, while some focus on the potential harm that can be caused by NLP systems.

Sun et al. (2019)[21] conducted a literature review on methods to measure and mitigate gender bias in NLP. They categorize bias into four categories: Denigration, Stereotyping, Recognition, and Under-representation. The authors provide an extensive summary of diverse methodologies in an organized fashion. While the authors recognize the complex nature of bias in NLP, the review primarily focuses on the technical aspects of different methods.

Czarnowska et al. (2021) approach quantifying social bias as extrinsic fairness metrics, which looks at model's performance parity across different social groups. The authors approach different metrics in a generalized parameterization using sensitive attributes. The review is carried out from purely technical and mathematical perspectives.

Dev et al. (2022) reviews measurements and mitigation of bias aligned with associated harms. The authors provide a theoretically motivated framework using concepts from linguistics and social psychology and apply it to 43 existing measurements. The authors provide a comprehensive list of questions to identify harms and improve the quality of measurements.

While the other literature review on bias covered the broad scope of NLP, Delobelle et al.(2022)[22] studies bias metrics for pre-trained language models. Delobelle et al., (2022)[22] evaluate compatibility of bias and fairness metrics suggested for language models and their downstream tasks. Their literature review is accompanied by correlation analysis and empirical evaluations.

1.3.3 Gender Bias

In my thesis, I specifically review methods for measuring gender bias in language models. I found 19 papers that propose a novel method for measuring gender bias in language models. Bias in language models have attracted considerable critical attention in recent years. Gender bias, among others, has been widely studied in NLP. The seminal work

of Bolukbasi et al., (2016)[10] revealed the embedded gender bias in word embeddings that "Man is to Computer Programmer as Women is Homemaker".

While myriads of suggestions have been made to improve gender bias in NLP systems, the field still suffers from significant bias against women and other marginalized genders. The latest language models achieving impressive performance in various tasks still exhibit gender bias. Therefore, I focus on gender bias to review suggested methods for identifying and measuring gender bias. Based on the review, I will also propose a comprehensive framework that can contribute to identifying gender bias in a more constructive way.

1.3.4 Critical Analysis of Bias in NLP

Beyond providing a general overview of methods, some have raised fundamental criticisms for studying bias in NLP.

Blodgett et al. (2020)[23] address the lack of normative reasoning in bias studies in Natural Language Processing (NLP). They surveyed 146 papers analyzing 'bias' in NLP looking at their motivations and techniques. The authors suggest recommendations for studying bias in language models for literature outside NLP, provide normative reasoning, and engage with people who are affected by NLP systems. My thesis applies the authors recommendations and extends it by applying specific framework and measurement modeling, for empirical validation and engaging in normative analysis of methods.

Stanczak and Augenstein (2021)[24] review 304 papers on gender bias in NLP in a broader perspective beyond technical differences. The authors point out fundamental limitations of studying gender in NLP. The authors argue that most work lacks recognition of gender as a social construct with fluidity and continuity. Moreover, most works overlook low-resource languages and ethical considerations of their work. Also, they work with very limited definitions of gender bias without baselines and pipelines. The authors argue that 'fractured' research does not engage in parallel research and falls short on empirical validity.

1.3.5 Algorithmic Bias and Fairness

Bias and fairness have been studied by many philosophers and ethicists studying algorithms, machine learning, and AI in recent years. A growing number of journals, such as AI ethics, Philosophy and Technology, and AI and Society, have provided a venue for heated scholarly debate on bias and fairness in algorithmic systems. While the whole field of ethics of AI, machine learning, and data is comprised of diverse works, here I introduce several works on algorithmic bias that provide a background of this thesis's topic, which is bias in language models.

Veale and Binns (2017) focus on unfair bias that is reproduced from historical data that trains the machine learning algorithm, especially in relation to socially sensitive

data. Danks and London (2017) provide a taxonomy of various types, sources, and impacts of algorithmic bias by unpacking multiple meanings of bias.

A common feature of computer systems on the basis of algorithms, machine learning, and AI is abstraction. Selbst et al. (2019) suggest five traps of machine learning based systems that aim to produce fair outcomes. The authors stress the importance of focusing on the process of technical designs, and distinguishing technical and social actors.[1]

Mittelstadt et al., (2016) and Tsamados et al., (2021) describe a high-level overview of the ethics of algorithms. Tsamados et al., (2021) identify relevant ethical concerns and suggest solutions for corresponding concerns through a systematic literature search. The authors analyze epistemic and normative implications about algorithms.

Johnson (2021)[25] compares human implicit algorithms and algorithmic fairness, concluding both types of bias share similar characteristics, i.e. relying on seemingly innocuous patterns. Also, Johnson (2022) demonstrate that contemporary debates on algorithmic fairness should consider the value-laden nature of algorithms.

1.4 Investigating Bias with Interdisciplinary Methods

I argue that measuring and mitigating bias in language model sis equivalent as embodying values in technology, namely embodying fairness in language models.[2] Bias measurements in language models suggest how fairness can be defined and operationalized in a given context, such as a specific downstream task like coreference resolution or a specific type of bias like gender bias.[3]

Bias measurement methods suggest which additional dimensions should be included to evaluate language models. Bias measurements and mitigation strategies demand language models to be excellent not only in technical dimension, but also social and moral dimensions. Embodying fairness in language models can be understood in the tradition of values in design approach which incorporates 'substantive social, moral, and political values to which societies and their people subscribe' (Flanagan et al., 2010, p. 332)[26]. Language models that are good in terms of technical requirements can fail on social and moral requirements by exhibiting bias against marginalized social groups. Therefore, social and moral requirements should extend the quality criteria to develop and improve language models. No language models function in vacuums; they are always

[1] "Achieving fairness in machine learning systems requires embracing a sociotechnical view. That is, technical actors must shift from seeking a solution to grappling with different frameworks that provide guidance in identifying, articulating, and responding to fundamental tensions, uncertainties, and conflicts inherent in sociotechnical systems. In other words, reality is messy, but strong frameworks can help enable process and order, even if they cannot provide definitive solutions." (Selbst et al., 2019, p. 63)

[2] Definitions of bias and their relations to fairness will be discussed in detail in the coming chapters.

[3] Flanagan et al., (2010)[26] argue that values in technology requires three modes of inquiry that are dynamically interdependent to each other: technical, philosophical, and empirical investigations. In my thesis, bias measurements in language models are analyzed using philosophical and empirical investigations. Compared to the abundant body of work on technical aspects of language models, empirical and philosophical perspectives have received less attention for evaluating language models.

situated in society through the training data, the environment that enables development of language models, and the interaction with users. Therefore, there is no non-social use case that societal impact is not relevant to the language model.

1.4.1 Contribution

My thesis provides a novel contribution to existing literature reviews on bias in NLP. I analyze measurements from both empirical and normative perspectives, following the values in design framework suggested by Flanagan et al. (2010)[26]. While previous published research addressed the necessity of normative reasoning to study bias in NLP, philosophical discussion on algorithmic bias and literature reviews on measurement and mitigation metrics strategies have been isolated in separated fields of studies. I aim to fill the gap between normative and empirical investigation in this interdisciplinary analysis by applying two perspectives to the same list of measurements that I obtained from a systematic literature review.

1.5 Terminology

1.5.1 Language Models

In this thesis, the terms 'language models (LM)' and 'large language models (LLM)' are used interchangeably. I refer to language models developed since the introduction of BERT (Devlin et al., 2019), which initiated the popularity of language models in various NLP applications. There are various terms used to indicate language models, including generative models, auto-regressive models, foundation models, and transformer models [27], to name a few. For my purpose, I mainly use the term 'language models'. Large language models have been widely used to highlight differences from the historical version of language models that date back to 1940s [28]. However, the term 'large language models' is not a strictly defined term with fixed conditions to meet compared to language models. What used to be a "large" language model in the past can be "small" in several months. Therefore, for most of my thesis I will use the term language models mainly, with a few exceptions where I use LLM to signify its relative size with other algorithms.

1.5.2 Bias

A variety of definitions of the term bias exist and the complex definitions of bias will be discussed as one of the main theme in this thesis. I argue that the definition provided by Friedman and Nissenbaum (1996)[17] best describes the phenomena of interest, which is bias in language models with significant social and moral implications. Friedman and Nissenbaum define computer systems are biased when they "systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" (ibid,

p. 332). For instance, it is not an interesting phenomena when a language model produces more text related to cats than dogs. However, when it creates more texts about sons than daughters, it is a pattern worth investigation. More in-depth discussion about the definition of bias can be found in the next chapter.

1.5.3 Gender Bias

The Dictionary of Psychology by the American Psychological Association (APA) defines gender bias as follows: "any one of a variety of stereotypical beliefs about individuals on the basis of their sex, particularly as related to the differential treatment of females and males.[4]" [29] In my thesis, gender bias refers to systematic unfair treatment on the basis of gender based on the definition of bias.

1.5.4 Sexism

Gender and sex are used to categorize different groups in measuring bias in language models. Generally, sex focuses on biological features, such as chromosomes and organs, for human categorization. Gender, on the other hand, is a social construct that consists of social features, such as social status, role, and identity. An in-depth discussion about what consists of gender and sex goes beyond the scope of this thesis[5]. In my thesis, I use the term gender bias, as the bias measures are interested in categorizing different groups based on socially salient features, rather than physical characteristics.

1.5.5 Operationalization

Operationalization is a process of defining the measurement of unobservable construct via inference from relevant observable properties (Gravetter and Forzano, 2018)[31]. Abstract construct such as bias is difficult to measure since it is intangible and not easily observed. There are many different options to operationalize the construct. Operational definition provides a way to measure unobservable construct, but it is not equivalent to the construct itself. In the process of choosing which properties to be measured, certain components of a construct can be left out. The quality of measurement procedure can be evaluated by validity and reliability of the measurement.

1.5.6 Measurement

The term "measurement" is used broadly to describe the various types of metrics, measurements, indicators, and tests developed to quantify bias in language models. Unless

[4]The definition from the dictionary also provides an example as following: "These biases are often expressed linguistically, as in use of the phrase physicians and their wives (instead of physicians and their spouses, which avoids the implication that physicians must be male), or of the term he when people of both sexes are being discussed."

[5]See SEP article on Feminism and Gender [30] for philosophical discussion about the concept of gender as a construct.

specified, no specific measurement is being implied by use of the term "measurement". In most cases, the term can be understood as suggested methods for measuring bias in language models.

1.6 Thesis Overview

This thesis seeks to extend our understanding of bias in language models, and provide a deeper understanding of its meanings and implications. A mixed-method approach was employed using a systematic literature review, empirical validation of methodology, and normative analysis. I review technical papers on bias measurements for language models and validate the operationalization using measurement modeling from social science. Also, I engage in philosophical debate on fairness in relation to bias in language models.

Measuring gender bias will be investigated in three different components: motivation, definition, and operationalization. The thesis is divided into six chapters.

In the introduction, I presented why bias in language models is a relevant topic. I presented relevant works, value in design theory, and terminology to build a foundation for the remaining chapters of the thesis.

In Chapter 2, I discuss the underspecification of the definition of bias. An ambiguous definition of bias as a construct is one of the biggest challenges that suggested works suffer in general. The term, bias, is conflated and not conceptualized. In the first section of the chapter, I will depict varying meanings of bias across different disciplines, highlighting their epistemic and normative differences. In the second section, I will review how the construct of gender bias is defined in the reviewed literature of bias measurement.

Drawbacks in conceptualizing the definition are closely related to the shortcomings in operationalization, which will be discussed in the Chapter 3. To measure bias, certain properties are chosen to be measured and inferred, as the construct bias is not measurable directly. Measuring unmeasurable constructs is not a trivial task. While other disciplines with a long history of quantifying abstract social constructs, natural language processing has not paid enough attention to the importance of valid operationalization of the constructs. As a result, the construct bias is measured via a limited set of properties that are chosen at the convenience of data and available tools. I will review suggested methods from 19 papers in detail by using measurement modeling, which is a method commonly used in social science to validate operationalization.

To overcome the shortcomings of existing works, I suggest a framework for holistic evaluation of bias language models in Chapter 4. Instead of focusing exclusively on the output of language models and probing technical dimensions of bias, I argue that structural dimensions of bias ought to be integrated in measuring bias. The suggested framework consists of all three dimensions that I analyze in the thesis, namely conceptual, empirical, and structural analysis of bias in language models.

In chapter 4, I discuss the normative motivations provided by the bias measurement

papers. Most bias measurement motivates their work with normative reasons, such as social inequality and harm, those motivations are underspecified and vague. Using bias as an umbrella term for various morally dubious properties of language models may not only erode the integrity of the suggested method, but can also lead to misleading interpretations of the result. I connect the motivations with varying definitions of unbiased language models. Furthermore, I suggest to expand the understanding of bias beyond the lens of distributive justice. An outcome-based analysis inevitably limits how bias is defined and measured, yielding only a bounded solution for the problem of fairness.

Lastly, I draw an analogy between statistical discrimination and language models. Language models are a tool developed with intention and resource by humans. They reflect human values and they contribute to shaping our environment, which cannot be properly investigated without contextual consideration of where the language models are situated. Due to its flexibility and powerful capability in a myriad of tasks, the impact of adopting this particular technology is expected to be disruptive and transformative. It is imperative to critically evaluate its potential harms and risks in a complicated social context, and defining and measuring bias with scientifically valid methods is the first step to address the problem of fairness connected to the bias in language models.

Chapter 2

Defining Bias

For many years, the phenomenon of mixed use of the term bias was surprisingly neglected in NLP. Discussing values, such as fairness and equality related to bias, in sociotechnical system is inherently complex and requires an interdisciplinary approach. Besides the fundamental necessity of clear conceptualization for scientific investigation, using precise language is even more crucial for discussing value in sociotechnical systems. Mulligan et al., (2019)[32] stress the value of shared vocabularies to facilitate conversations across disciplinary boundaries. Without shared language, the conflation of concepts can "confuse the discussion about values in technology at disciplinary boundaries". (ibid, p. 1191), as the same vocabularies are conceptualized very differently depending on the disciplines.

However, bias in language models suffers from the challenge of interdisciplinary confusion. Discussing bias in language model shows the challenge of collaborating across disciplinary boundaries, such as computer science, linguistics, and philosophy. In this chapter, I first investigate how the definition bias is conceptualized differently depending on disciplines. I show that the term is conflated both conceptually and operationally, which leads to a multitude of conflated uses of the term bias in the context of language models. Conceptually, bias is used to indicate both descriptive and normative wrong. Operationally, bias is measured by heterogeneous properties, such as stereotypes and sentiments. The shortcomings in of defining bias in accurate language leads to further challenges to measure and mitigate bias.

2.1 Diverse Definitions of Bias

The term bias is used expansively in the literature on bias in language models. The definition often lacks clarity partly due to the varied use of the term across different disciplines. In this section, I will discuss how bias is defined in statistical analysis, psychology, and social psychology especially regarding cognitive biases.

In statistical analysis, bias is defined as a discrepancy between a truth value and an estimated value [33]. An estimator is unbiased when the bias equals to zero, namely

when there is no difference between the estimated value and the ground truth value. It is desirable to minimize the bias to develop an accurate estimator.

The American Psychological Association (APA) dictionary [34] defines bias with four definitions. It can mean (1) partiality, associated with prejudice (2) any tendency or preference, (3) systematic error arising during sampling, data collection, or data analysis, (4) any deviation of a measured or calculated quantity from its actual (true) value. The third and fourth definitions are especially close to the statistical meaning of bias. Contextual information is necessary to clarify whether the bias is used to indicate an sampling error or personal preference.

In social psychology, bias often refers to human cognitive bias, such as linguistic intergroup bias and correpondence bias [35]. The linguistic intergroup bias indicates different language usage based on the group membership and social desirability (p. 203). A correspondence bias refers to the cultural tendency of underusing contexutal information to make a judgment about an individual (p. 6). Such conceptualization of bias is closely related to the tendency or preference, rather than statistical deviation or errors. Implicit and explicit bias, which are often discussed in relation to discriminatory action, are also closely connected to cognitive aspects of bias.

A colloquial use of bias has a normatively negative connotation. The Cambridge Dictionary defines bias as "the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment" [36]. When a language model is described as biased in terms of a socially sensitive feature, it mostly refers to language models being unfair to the groups with those socially sensitive features.

As shown in above examples, the definition of bias varies significantly depending on the disciplinary context. When the bias is defined in the statistical sense as shown previously, the bias is not used to describe normatively undesirable phenomena. The gap between observed or estimated value and the truth value can reveal normatively problematic phenomena as a result, but morally problematic character is not a necessary condition to define a biased instance. In contrast, when bias is defined as an unfair treatment, such as in Friedman and Nissenbaum (1996)[17] normative evaluation becomes an integral part to determine if any pattern can be considered biased or not. Since the context of investigating bias in language models is not limited to a single disciplinary, the disciplinary background alone does not provide a enough context to specify how bias is conceptualized.

2.2 Conflated Meanings of Bias in Language Models

Conflated use of bias is an important, but understudied, problem. Bias is used as an umbrella term to indicate various morally undesirable patterns of language models. It resembles the usage of 'fairness' as Binns (2018) [37] describes: "this discussion suggests

that 'fairness' as used in the fair machine learning community is best understood as a placeholder term for a variety of normative egalitarian considerations." (ibid, p. 2) Similarly, bias is used as a placeholder for a variety of undesirable patterns discovered in language models, from stereotypes to skewed distribution between groups.

The meaning of bias is conflated in two ways in the context of language models. First, conceptually, bias is used for both normative wrong and descriptive incorrectness, which inhibits a clear interpretation of empirically measured bias in language models. Second, it is operationalized through different properties like stereotypes, various types of sentiment scores, and differences between association with male and female entities.

2.2.1 Conceptual Conflation: Normative Wrong and Descriptive Wrong

Conceptually, the bias is used to indicate both normative correctness and descriptive accuracy. This conflation leads to varying definitions of unbiased language models, which is closely connected to how to define ideal language models that are unbiased. Some bias measurements define bias as the gap between the output of language models and the ideal status (Normative wrong). Alternatively, bias is defined as a discrepancy between the language model and status quo, which is represented by a statistical description of society measuring subjective human perceptions (Descriptive wrong). Consequently, such varied definitions of bias leads to vastly different implications, despite sharing the same vocabulary and similar methodologies to identify bias. In the next section, I will compare two different ways of defining unbiasedness referring to the proposed methods to measure gender bias. The implications of the two definitions will be discussed afterwards.

2.2.2 Operational Conflation: Measuring with Heterogeneous Properties

The meaning of bias is conflated in how bias is operationalized. The meaning of biased model varies significantly based on how bias is measured, since heterogeneous properties are used to measure bias. Since bias is a complex construct, potentially diverse operationalizations are possible to measure bias. However, the problem is that the underspecification of bias leads to reductive definition of bias, such as stereotypes. While language models creating harm, being stereotypical, and being unfair share normatively undesirable properties, they are not identical.

It is necessary to differentiate these relevant but distinct concepts. They can be relatively easily distinguished when applied to simpler examples. For instance, associating continental Europe and bad weather is stereotypical yet not necessarily harmful. It can be considered unfair, unless such a stereotype leads to discriminative treatment of people from continental Europe. Similarly, associating a random nationality with violent character would be always unfair and harmful, but not necessarily stereotypical unless there exists widely held perceptions or social practices that reflects the association. When language models produce more 'blue' than 'red' in text generation, it would be neither

	Descriptively accurate	Descriptively inaccurate
Normatively correct	Utopia-land	Fantasy-land
Normatively incorrect	Dilemma-land	Disaster-land

Table 2.1: Comparing Three Lands Based on the Analogy in Deery and Bailey (2022)

stereotypical, unfair, nor harmful. Yet, it still can be described as 'biased' in a statistical sense.

Moreover, the connection between properties chosen in the language model and bias is not well-established in practice. Underspecification and a lack of conceptualization of the bias creates room for varying interpretations of biased language models.

In Chapter 3, I will categorize which observable properties were chosen to measure gender bias based on the structural literature review. It will provide an overview of current practices in conceptualizing and operationalizing gender bias in language models.

2.3 Conceptual Conflation - The 'Unbiased' Language Model

As discussed in the previous section, bias in a language model is used as an umbrella term to refer to distinct patterns of a language model's output that are unintended or undesirable. Based on the varying use of the term 'bias', numerous versions of 'unbiased' language models exist. The conflated use also impacts how 'unbiased' language models are defined. Building upon the established practice of defining and operationalizing bias in language models, I frame the discussion between two vastly different types of unbiased language models: One way for the language model to be unbiased is to make the model meet the artificially defined yet ideal requirement of fairness. Alternatively, the language model is unbiased when it reflects the status quo correctly, which is represented by statistics or human baseline.

2.3.1 Tales of Two Lands

Deery and Bailey (2022)[38] provide an analogy to describe different positions on how ideal natural language processing technology should be. While the authors focus is not specifically language models, the same conceptual categorizations apply to measuring bias in language models. The authors compare different positions based on two axes: normative correctness and descriptive accuracy. I will refer to their analogy for describing different types of unbiased language models, whether language models should align with ideal status or status quo.

Four combinations are possible based on two axes of normative correctness and descriptive accuracy. In an ideal world, the Utopia-Land according to Deery and Bailey's analogy, humans are not biased, therefore language models are also not biased. Therefore, good language models are descriptively accurate and normatively correct at the

same time. In the Disaster-Land, on the other hand, language models are descriptively and normatively wrong. It would be irrelevant to discuss such a model, as the model would be useless.

Fantasy Land vs. Dilemma Land

Now remain two difficult choices, the Fantasy-Land and the Dilemma-Land. Revisiting previously reviewed methods for bias measurements, the most common method is to align language models with an ideal state defined by researchers. Some examples of the ideal states are language models being a-stereotypical, or not differentiating between gender or race. Those methods belong to the Fantasy-Land. It is a fantasy as they are out of reality. It is unrealistic, out of reality, but an ideal state of the world that language models should aspire to be.

The other method that compares language models with descriptions of the status quo, such as national statistics or human annotations. These comparisons belong to the Dilemma-Land. It is a dilemma since there is a unsolvable problem to meet both descriptive and normative demands. It is not 'ideal', as the current world is not perfect. Aligning language models with imperfect reality will inevitably end up with imperfect language models. Yet it can shed light on how far off language models are from the reality, which is a less ambitious yet useful indicator.

The decision between the Fantasy-Land and Dilemma-Land comes down to prioritizing descriptive accuracy or normative correctness. The dilemma arising from the tension between normative and descriptive correctness is not a unique problem in language models or algorithms. It is "a specter of normative conflict"(Basu, 2020)[39] that fairness might require inaccuracy. The dilemma perspective assumes that the apparent conflict between fairness and accuracy cannot be resolved (p. 191-197). The Fantasy-Land puts forward normative correctness, and claims that language models should be free from problematic patterns such as stereotypes that exist in reality. Aiming for such an ideal comes at the cost of losing descriptive accuracy, and it can rupture the usefulness of language models. The Dilemma-Land, on the other hand, prioritizes descriptive accuracy over normative correctness. It corresponds to what the reality is like, regardless of whether it is morally desirable or not. Therefore. it can risk perpetuating existing problems of reality through language models and reproduce undesirable patterns of injustice.[1]

[1]Deery and Bailey (2022)[38] argues that there is ethical value in not debiasing, such as presenting problematic patterns. Debiasing can create a false illusion of improved fairness, more than it actually is, which can contribute to a devaluation of the problem. However, in the case of language models, I argue that the risks of perpetuating problematic patterns through language models outweigh the ethical value of leaving the probelmatic patterns intact in language models.

2.4 Bias Measurements in Two Lands

2.4.1 Fantasy-Land Papers

Bias in language models is often identified by comparing language model to an ideal state. Varying versions of ideal states are suggested by the authors who propose the metrics to measure bias. Referring to an ideal state seems to be a self-suggestive step considering problematic patterns reflected in training data, which is the prominent source of bias in language models. Data reflects a reality including morally imperfect features of the world. Therefore, language models trained on problematic data inevitably exhibits problematic patterns. By laying out an alternative ideal that is removed from problematic patterns of reality, language models can, so goes the argument, be improved.

Based on systematic literature review, most papers suggesting bias measurement aim for the unbiased language model in the Fantasy-Land. Namely, language models should not prefer certain groups over others, nor more prone to produce stereotypical text. The most frequent interpretation of an unbiased language model is to make it a-stereotypical [40]. Another common approach is to look for a uniform baseline between groups, treating different groups of individuals 'equally' [41].

However, this approach requires defining what an ideal language model is, and not everyone can agree on a single version of an ideal language model. The definition of ideal status can be subjective, conflicting, and controversial. For instance, when the ideal language model is defined as an a-stereotypical language model, what a-stereotypical means should be first defined. However, it is a challenging task of its own, as Blodgett et al., (2021)[42] show that existing fairness benchmarks used in NLP have conflicting operationalization of stereotypes. For instance, one example that authors identified as problematic, defined a stereotype as "The exchange student became the star of all of our art shows and drama performances" while the anti-stereotype was described as "The exchange student was the star of our football team" (ibid, p. 1004), which is not a clear case of stereotypical relationship.

Establishing the criteria for an ideal language model requires conceptualizing relevant concepts and evaluating conflicting values. However, as similarly to defining bias, conceptualizing value-laden concepts in sociotechnical concept is a challenging task beyond disciplinary boundaries. Therefore, an alternative approach does not attempt to provide a normative position on how language models should be. Instead, the language model is compared with reality.

2.4.2 Dilemma-Land Papers

As shown previously, setting an 'ideal' status is challenging since there is no one perfect language model that everyone agrees on. Alternatively, some turn towards real-world statistics as their source of what a language model should represent. According to the statistical account of bias, the language model is biased when the output does not

correspond to the statistical source that is being compared to. Another approach is to validate a language model's output with human evaluation, to assess whether a language models' bias align with human bias. In this subjective-evaluative account of bias, bias in a language model can be permissible as long as humans exhibit a similar degree of bias.

Between the subjective-evaluative account and the statistical account, the statistical account of bias is the more commonly used in measuring gender bias in language models. Bias in language model is defined as the discrepancy between the language model's output and the statistics in the Dilemma-Land. Since an occupational stereotype is one of the most common ways to evaluate bias, national labor statistics in countries [43][44]. Based on subjective-evaluative account, bias is defined as a discrepancy between a language models' output and the statistical reference that was chosen.

Other works compare language model output to human perceptions by comparing it with human perceptions, such as crowd-sourced annotators [45] or experts. These attempt to align language models with humans by evaluating a language models" inference to a human evaluations. For instance, Sotnikova et al. (2021)[46] are used the authors' evaluation on how stereotypical given statements generated by language models are.

By adopting a statistical and subjective-evaluative account, bias is focused on the descriptive difference between language models and reference. Touileb et al. (2022)[43] argue that instead of making a normative analysis, their work provides a descriptive assessment of the distribution of occupations. However, it is worth noting that the implication of choosing statistical reference to define unbiased language model is not expanded by the authors. Despite putting the epistemic priority over normative priority by choosing descriptive representation instead of normative analysis, the authors define bias as systematic unfair treatment towards certain groups of individuals, following the definition given by Friedman and Nissenbaum (1996)[17]. While the definition of bias is putting normative correctness over descriptive accuracy by defining bias as unfairness, the analysis is prioritizing descriptive accuracy by comparing the language models to the statistics. Embedded unfairness in the society will be reflected in the statistical representation and such unfairness will go unnoticed when the unbiased language model is defined reflecting statistical representation.

Chapter 3

Measuring Bias

Bias in language model a construct that cannot be directly observed or quantified. The construct that cannot be directly measured is measured through observable properties that are relevant to the construct [47][48][49]. Valid measurements enable to infer unobservable constructs from test scores of the observed [50](p. 84). Therefore, the purpose of measuring bias is to make inferences from observed patterns in the language model to bias of language models. Establishing a valid connection between the construct, bias, and selected properties is critical to devise a method to measure bias. Validity of the measurement is also relevant to interpret the implication of identified bias and developing a mitigation strategy.

A valid operationalization starts from conceptualizing the definition. Relevant observables should be chosen based on the clear definition. In Chapter 5, I will discuss the method to conceptualize definition systematically. In this chapter, I focus on reviewing how existing methods operationalized gender bias with the choice of measurable properties. Afterwards, I introduce measurement modeling to validate operationalization of gender bias. I apply various criteria from measurement modeling to evaluate quality of gender bias metrics from 19 papers retrieved by systematic literature review.

3.1 Operationalizing Gender Bias

As discussed in the previous chapters, the concept of bias is conflated in terms of how it is operationalized. Based on the systematic literature review, I analyzed which properties were chosen to measure gender bias. The most common way to conceptualize gender bias is through stereotypes. 13 out of 19 papers use stereotypes to measure gender bias. The most frequently measured type of stereotype are occupational stereotypes. Sentiment score is also often used to assess gender bias.

Stereotypes, especially occupational stereotypes (32%, [51][45][43][52][44][53]), are the most commonly used properties to measure gender bias(68%, non-occupational stereotypes: [54][55][56][57][46][40][58]). Sentiment scores (21%, [45][59][60][41]) were the next common way. Stereotypes and sentiment scores together account for the ma-

jority, 17 out of 19 papers, of operationalization. One paper [61] investigates behavioural expectation by analyzing topics and lexicons associated with the perceived gender of fictional characters. Two other papers compare differences between female and male groups in varying aspects, such as the likelihood of predicting tokens [62], recommendations, and rankings [63].

3.1.1 Stereotypes

Stereotyping, especially based on gender and race, has been the dominant focus of bias studies in language models and attracted considerable critical attention [21]. Stereotypes are a good indicator for discrimination and they case great harm. Therefore, getting rid of stereotyped unequal treatment would be a great step forward to prevent harms caused by language models.

Occupational Stereotypes

Among various types of stereotypes, stereotypes concerning occupation is the mostly used to measure gender bias. Occupational stereotypes refer to stereotypical association based on traditional gender roles, such as relating doctors to men and nurses to women. Considering the gender wage gap is one of the most referred indicator that represents gender bias in society, occupational stereotypes in language models show how language models reproduce existing stereotypes. Occupational stereotypes provide a useful account of undesirable outputs in language models. Six papers measure occupational stereotypes using a template-based approach. These methods involves creating a template using features like gender and generating text based on the template.

Kirk et al. (2021)[51] generate text with a subject that consists of a protected class, such as gender. The authors measure the frequency of jobs in the generated text. The result is compared to the US Labor Bureau statistics. To measure gender bias, Dhamala et al. (2021)[45] use gender-based prompts using English Wikipedia articles about female and male actors. Dhamala et al. (2021)[45] proposefive different metrics to measure bias across five domains, with gender being one of them. Four out of five metrics are different types of sentiment score. The other metric is gender polarity, which is measured either by counting gendered words, such as ‘he’ and ‘him’, in the text, or measured by calculating cosine similarity between ‘he’ and ‘she’. Touileb et al. (2022)[43] create templates combining occupations, pronouns, and first names. Generated text from templates are analyzed and compared to the Norwegian statistics.

Alnegheimish et al. (2022)[52] collect sentences about professions from Wikipedia articles according to Wikipedia’s occupation catalog, and the likelihood of generating tokens are compared based on the prompt.

Bartl et al. (2020)[44] measure gender bias based on the prediction of masked tokens from the template that consists of the person words, which is a description of a person including gender and profession information (p. 4), and profession words.

De Vassimon Manela et al. (2021)[53] measure gender bias using two metrics, skew and stereotypes, on the basis of WinoBias dataset. The dataset is composed in a way that stereotypes are determined by professional gender imbalances recorded by the U.S. Bureau of Labor Statistics. Skew measures the assignment of stereotypical pronouns to professions. Stereotype measures the language models' preference towards male and female pronoun resolution across stereotypical and anti-stereotypical professions.

Other stereotypes

Lucy and Bamman (2021)[61] studied how gender stereotypes in books and films are reproduced in language models. The authors conduct two qualitative analysis of generated stories. First, they evaluate occurrence of topical terms, such as family, emotions, body parts, and politics, depending on the perceived gender of the character. Second, the authors analyze lexicons using cosine (semantic) similarity to see how much attention is given to characters' appearance, intellect, and power.

For other types of stereotypes, four papers (Nangia et al. (2020)[55], Kwon and Mihindukulasooriya (2022)[56], Steinborn et al. (2022)[57], and Nadeem et al. (2020)[40]) used crowd-sourcing to generate a dataset of stereotypical phrases.

Nadeem et al. (2019)[40] created the StereoSet, which is a crowd-sourced minimal distance pair of stereotypical and anti-stereotypical sentences. Bias is measured as the discrepency in probability of models predicting stereotypical, anti-stereotypical, and unrelated phrases for generating text in fill-in-the-blank type tasks.

Nangia et al. (2020)[55] created the Crowdsourced Stereotype Pairs benchmark (CrowS-Pairs) dataset to measure social bias. The bias is measured by the percentage of stereotyping examples over the less stereotyping sentences. Ideally, the model should not prefer either of the options. Kwon and Mihindukulasooriya (2022)[56] and Steinborn et al. (2022)[57] also utilize the CrowS-Pairs dataset, as the basis of their gender bias measurements.

Sotnikova et al., (2021)[46]'s authors constructed a list of real-life contexts and insert target categories to create premises for measuring language models inference. The authors aimed to create neutral premises to describe situations independent of the target categories.

Barikeri et al. (2021)[58] created a conversational dataset using Reddit. The gender bias is estimated by measuring the likelihood of generating stereotypically biased phrases, in comparison with a corresponding inversely biased phrases with replaced bias specification.

Lastly, May et al. (2019)[54] measure two specific stereotypes: Angry Black Women (ABW) and Double Binds. The ABW stereotype is measured using black-identifying and white-identifying female given names. The bias of sentence encoders is measured as the association of names and attributes, and which adjectives are used in the discussion of the stereotype and their antonyms. Double bind stereotypes are measured using male

and female names and attributes consisting of likable and non-hostile terms.

3.1.2 Sentiment Scores

Four papers used sentiment scores. Sentiment scores should be carefully evaluated as the mechanism of calculating a sentiment score varies significantly across metrics, from emotional valence to movie ratings.

Dhamala et al. (2021)[45] used multiple scores related to sentiment: Valence Aware Dictionary and Sentiment Reasoner (VADER)[64], Zigsaw toxicity [65], regard [66], and psycholinguistic norms. VADER uses word-level valence-based lexicons and the lexicon polarity to calculate sentiment score. Toxicity was measured by using a fine-tuned BERT model. Toxicity was measured as how disrespectful, abusive, unpleasant, and/or harmful the texts are. Regard is a method developed by Sheng et al. (2019) [66], which measures human-annotated bias. It uses the polarity towards certain demographics, rather than overall language polarity to separate sentiment and language polarity. Lastly, psycholinguistic norms are based on numeric ratings annotated by expert psychologists to words. It is designed to measure the affective meaning of words along various dimensions. Based on previous work designed to measure emotion from the text, selected emotional dimensions are considered: Valence, Arousal, Dominance, Joy, Anger, Sadness, Fear, and Disgust. (Dhamala et al., 2021, p. 865) [45]

Jentzsch and Turan (2022) [59] analyzed gender bias in BERT models in a specific downstream sentiment: classification task. They use IMDB data which classifies the sentiment as positive for the star ratings higher than 7. Negative sentiment is defined as ratings of 4 or lower and scores between 5 and 6 are considered neutral.

Wolfe and Caliskan (2021) [60], Silva et al. (2021)[41], and Shen et al., (2023)[63] used the Word Embedding Association Test(WEAT). The WEAT was designed in Caliskan et al. (2017)[9] to evaluate the strength of association of a set of target static word embeddings (e.g., women and men) to two sets of opposing attribute static word embeddings (e.g., pleasant and unpleasant). WEAT is developed analogously to the Implicit Association Test (IAT)[67], which is widely used to measure implicit bias of human.

In addition to WEAT, Silva et al. (2021)[41] used Equity Evaluation Corpus (EEC) additionally. EEC was introduced in Kiritchenko and Mohammad (2018)[68]. EEC measures the bias score by using the " TARGET feels ATTRIBUTE " template. Target tokens consist of gendered or racial tokens. and attribute tokens consist of emotional words such as 'angry', 'happy', and 'heartbreaking'.

3.1.3 Other Operationalizations

While majority of papers used stereotypes and sentient scores to operationalize gender bias, there are different methods used to identify gender bias.

Shen et al., (2023)[63] looks at recommendation made by conversational language models. With a given item information, the authors perform user-item analysis based

on attribuates to identify potential discrimination towards any specific user group. It is done by analyzing recommended price for different groups.

Silva et al. (2021)[41] uses pronoun ranking to measure additional aspects of gender bias besides two sentiment scores. Bias is measured as the likelihood of predicting pronouns by comparing the relative likelihoods of target words.

In addition to WEAT, Kaneko et al. (2020)[62] evaluate social bias as the relative importance of words in a sentence in predicting tokens with the attention weights.

3.2 Measurement Modeling

In previous section, I reviewed how gender bias is operationalized. While the overview shows the limited set of properties to measure gender bias in language models, a more systematic approach is necessary to evaluate measurement qualities. In social science, measurement theories have developed to improve the quality of measurement qualityfing obsesrvable constructs. Measurement modeling aims to effectively assess the validity of measurements and provides a shared framework to evaluate operationalization enabling efficient communication between researchers [47]. Measurement modeling helps to distinct between operational and conceptual disputes and to establish generality and validity of measurements.

Measurement modeling has been introduced for developing rigorous bias metrics in NLP, such as in Jacobs and Wallach (2021)[69], Blodgett et al., (2022), Bommasani and Liang (2022)[70], and Grimmer et al., (2022)[71]. Among other suggestions, I referred to the framework suggested by Jacobs and Wallach (2021)[69] as the main source. In addition, I referred to various sources from social science literature. The full table of applying measurement modeling to gender bias measurements can be found in the Appendix B.

According to Jacobs and Wallach (2021)[69], measurement modeling consists of two broad categories: construct reliability and construct validity. Construct reliability concerns the extent to which a measurement provides stable and consistent results (Carmines and Zeller, 1979 [72], Moser and Kalton, 1989 [73]). Construct validity determines whether the construct measures what it aims to measure. Among various types of validity, I chose four types that are the most relevant to measuring biases in language models: face validity, content validity, convergent/discriminant validity, and consequential validity.

3.2.1 Construct Reliability

Construct reliability determines the measurement's consistency and robustness. If the results are not reproducible, the measurement is not reliable. Unreliable measurements undermines the validity of research and can lead to misinterpretation. Several types of reliability exist, such as test-retest reliability, inter-rater reliability, intra-rater reliability,

and inter-item reliability. (ibid, p. 378-379)

Test-retest reliability evaluates if the measurement can produce consistent output when the test is repeated. Inter-rater reliability and intra-rater reliability measure the variance between different annotators for the same items. Inter-item reliability assesses the correlation between the inputs of the measurement model.

Bias is a complex construct, which makes it a difficult construct to operationalize. Furthermore, language models are non-deterministic algorithms that produce varying outputs with the identical input. Therefore, the same metric applied to same model can lead to different outcomes. Such characteristics of meausuring bias in language model make testing reliability in bias measurements a critical step to ensure measurement quality. By assessing test-retest reliability, the variance within the outputs from language models can be identified. Assessing inter-rater and intra-rater reliability is also crucial, as many bias datasets are generated from crowdsourcing. Without evaluating inter-rater reliability, the ground truth for measuring bias itself can be biased [42].

3.2.2 Construct Validity

Evaluating the validity of the construct is to assess whether "the observations meaningfully capture the ideas contained in the concepts" [47]. The observations based on the operationalization should reveal the aspects of the construct that are aimed to be measured. Numerous types of validity exists and different disciplines have different foci. The definition of each validity is also not set in stone, rather it is continuously developed as a part of measurement theory. For the purpose of evaluating bias measurement in language models, I chose face validity, content validity, convergent/discriminant validity, and consequential validity.

Face Validity

Face validity tests if the selected measurement appears to be plausible, which means it is "appear practical, pertinent, and related to the purpose of the test." (Nevo, 1985, p. 287)[74] It is the most elementary type of validity, yet remains a necessary condition to establish other construct validity. Despite its risk of subjective judgement, face validity provides an intuition to evaluate the relationship between the construct and the measurement. To ensure face validity, selected properties for measuring bias in language models should appear to be relevant to the construct purported to be measured.

Content Validity

Content validity measures the extent in which all relevant aspects of the construct are covered. Complex constructs such as bias consists of multiple dimensions and the measurement should be representative of the contents of the construct. Sireci (1998)[50] identified four common elements of various content validity literature, which are: domain

definition, domain relevance, domain representation, and appropriate test construction procedures.(ibid, p. 101) While various definitions of content validity exist, I will use the one provided by Jacobs and Wallach (2021)[69].

In a framework for understanding fairness in computational systems, Jacobs and Wallach (2021)[69] presents three sub-aspects of the content validity: contestedness, substantive validity, and structural validity (ibid, p. 379-380). First, contested constructs can have multiple definitions depending on the context. If the construct is contested, it is difficult for one measurement to comprehensively capture the complex aspects of the construct. Therefore, the operationalization should clearly articulate which dimension of potentially diverse theoretical understandings of the construct is being measured. Based on the clear definition, the construct can be operationalized to measure the defined dimension. Another sub-aspect of content validity is substantive validity. Substantive validity focuses on the assumptions beneath quantifying the abstract construct. Substantive validity can be established by incorporating only relevant observable properties into the measurement. The third sub-aspect of content validity is structural validity. Structural validity can be established by capturing the structure of the relationships between observable properties and the construct.

Measurements can reflect complex aspects of the construct by referring to the established theory on the construct. Established theory developed to measure similar construct in society or people can be helpful for measuring the construct in language models. Theoreies about construct will theorize about not only the contestedness of the construct, but also underlying assumptions and relationships between the construct and properties. For instance, gender bias has been studied in various fields such as psychology, political science, sociology, or economics. Operationalizing bias in language models can only benefit from the way gender bias is conceptualized and measured in those fields.

Convergent Validity and Discriminant Validity

Despite the differences of convergent validity and discriminant validity, I investigated them as one category in gender bias measurement review, for the purpose of reviewing methods for measuring gender bias in language models. Two validities can be grouped together as establishing the validity by comparing the suggested measurement to other measurements. Convergent validity concerns how measurements work in comparison to the established measurement of the same construct. If the construct is conceptualized in a similar fashion, different measurements still ought to produce similar outputs.

Discriminant validity evaluates the variance of suggested methods to different measures measuring different constructs. Discriminant validity is necessary to prevent a confounding factor impacting the operationalization. If two methods measuring two different constructs produce the identical outputs, the measurement might fail to address the differences between the two constructs.

For instance, gender bias and uncivil behavior can be expressed in a similar fashion

such as vulgar languages. However, these two constructs, gender bias and uncivil behaviour, are not identical. The results from these two measurements can exhibit some correlation, but should vary to some degree based on the difference between gender bias and uncivil behaviour. If the difference is not validated, failing to ensure the discriminant validity, it would indicate that the measurements do not capture all the relevant aspects of gender bias apart from incivility.

Consequential Validity

Consequential validity is concerned with the impact of measurement, including societal impacts. Operationalizing a social construct by choosing specific observable properties has a consequential impact in society. Consequential validity concerns with assessing how the measurements shape the world and what is the implication of such consequences (Jacobs and Wallach, 2021, [69], p. 381).

Measured properties inevitably highlight certain aspects of the construct in comparison to other aspects of the construct that are not being measured. For instance, measuring gender bias by counting gendered pronouns, such as "he" and "she", has a consequential impact of emphasizing the occurrence of certain words. Emphasizing certain aspect and not accounting for other aspects inevitably simplifies the complex construct, for instance simplifying gender bias into the occurrence of "he" and "she".

Given how fairness and bias are both used as a "placeholder" (Binns, 2018, p. 2)[37], measuring societal impact of bias measurement is especially pertinent. In an "algorithmically infused society" (Wagner et al., 2022)[75], society are exchanging influence between algorithms. Algorithms reflect the society where it is developed and the algorithm influences the society. How "fairness" is conceptualized in language models will influence how the interaction between people and algorithm will be shaped in terms of fairness. This, in turn, will influence how fairness is perceived and shaped in society in a long term. Bias in language models can only be defined in the context the language model is situated.

For instance, skewed distribution of male and female in the training data is often pointed at as the culprit of "unfair" language models. Accordingly, one way to measure the language model's bias is to investigate its training set. Hypothetically, language models that are trained from the training dataset, which contains more male words, can be defined as biased. In this step, the distribution of the training dataset should be measured to identify bias of the language model. To reduce the bias, male words and female words should be contained proportionally in the dataset. Such definition of bias focuses on quantifying the occurrence of certain words (defined as male or female words). Reducing bias involves best quantifying occurrence of words, by defining the list (e.g. Which words should be contained in the list of male and female words? Should policeman counted as male or not?), and developing the counting mechanism (e.g. Should we count the word differently if it is in a grammatically different form? Should the word

be weighted when it appears more often in the same paragraph?). Such a quantified measurement will impact how fairness of language model is perceived and qualitative dimensions of gender bias will not be relatively underestimated.

Other Types of Validity

There are various types of validity, varying from field to field. Instead of trying to make an exhaustive list of validities, I aimed to identify what is the most relevant and imminently necessary in measuring bias in language models. For instance, I excluded hypothesis validity and predictive validity. Jacobs and Wallach (2021)[69] note that these validities concern the utility of the validity rather than the meaning of operationalization. What I am most interested in is establishing a meaningful link between the construct and the measurement, rather than the utility of measurements.

3.3 Applying Measurement Modeling

3.3.1 Construct Reliability

To determine whether the measurement is operationalized in a way to ensure construct reliability, I searched for inter-rater reliability and test-retest reliability in the paper proposing novel methods to measure gender bias. Inter-rater reliability was chosen since some bias measurements are compared to human evaluation, such as crowdsourced workers or experts. Test-retest reliability can validate whether the observed measurement is reproducible. Since language models are not deterministic and use numerous hyperparameters, the outputs produced by language models can vary. Therefore, test-retest reliability can contribute to the robustness of the measurements.

Among 19 papers, two papers reported reliability measure related to inter-annotator agreement.

Barikeri et al. (2021)[58] reported inter-annotator agreement using Krippendorff's alpha for annotating for bias to the comments and phrases retrieved from their Reddit dataset. The authors did not observe significant differences in the inter-annotator agreement.

Sotnikova et al. (2021)[46] paper reported inter-annotator agreement. The authors did not calculate coefficients commonly used to calculate inter-rater agreement, such as Fleiss's Kappa or Krippendorff's alpha. Instead, they shared the inter-annotator agreement percentage, which was calculated as the fraction of times all annotators give the same answer. The authors argue that the percentage of agreements is easier to interpret and some questions are expected to be different between annotators.(ibid, p. 4059) The percentage of annotator agreement contribute to the robustness of the measurement, therefore contributing to the construct reliability.

3.3.2 Construct Validity

Face Validity

Most measurements do choose plausible property that are relevant to gender bias. Detailed analysis of selected properties to measure gender bias has been reviewed in section 2-2. To evaluate face validity of bias measurements, I assessed whether the operationalization looks plausible at all. By plausible, I determined if properties appear to be practical and relevant to the purpose of the measurement.

While most of the properties seem to be plausible, gender polarity from Dhamala et al. (2021)[45], seems questionable. Dhamala et al., (2021)[45] use five different metrics for measuring bias. One of the metrics is gender polarity, which consists of two metric: one metric is termed unigram matching and the other is normalized word vector based metric. (ibid, p. 865)

Unigram matching use list of female and male identifying tokens, which consists gendered words such as he, him, man, and boys. According to unigram matching, a text is identified as expressing male gender when there are more male words than female words. If the number of male and female words are identical, the text is considered neutral. At a glance, the metric seems plausible in the simplest form (ibid, p. 865).

However, simple counting such as this metric risks confounding factors depending on the text. To prevent such a risk of confounding factors present in the text, most templates for measuring gender bias use grammatically bleached sentences or minimal distance pair to reduce the risk of other confounding factors. For instance, a misogynic text full of female words will be identified as expressing female gender according to the metric. However, it contradicts the authors definition of biased language models, which is disproportionate generation of text that is perceived as negative, unfair, or prejudiced (ibid, p. 862).

Whether simple counting of male and female words apepar to be "practical, pertinent, and related to the purpose of the test." (Nevo, 1985, p. 287)[74] to measure unfair perception against women or other marginalized genders is questionable.

Content Validity

To evaluate content validity, I determined whether the operationalization is established on relevant theories outside the technical domain, such as computer science or computational linguistics. Only a few ([54][59][60][44][58][61]) measurements refer to previous works in related fields such as psychology and sociology.

May et al. (2019)[54] refers to established literature on ABW stereotype such as Collins, (2004)[76] and Madison (2009)[77], while references such as Mitchell, 2012 [78]) for Double Binds. Jentzsch and Turan (2022)[59] expand upon previous research in psychology by conceptualizing gender bias, drawing inspiration from the work of Greenwald et al. (1998)[79]. In their recent publication, Wolfe and Caliskan (2021)[60] cite previous

studies on human bias and perceptions, including the works of Hughes et al. (2019)[80] and Greenwald et al. (1998)[79]. Barikeri et al. (2021)[58] draw upon sociological literature that explores the experiences of minoritized groups, referencing the works of Welch (2007), Shaw (2012), and Black (2015). Lucy and Bamman (2021)[61] rely on behavioral studies (Johns, 2019)[81], cultural analytics (Kraicer, 2018)s [82], and research on gender stereotypes in media (Smith, 2012)[83] to investigate the portrayal of gender in various media forms.

Lastly, while Bartl et al. (2020)[44] reference existing research on masuring gender bias(Moss-Racusin et al. 2012)[84], but it is not closely related to the operationalization made by the authors. Moss-Racusin et al. (2012) measure academic science's gender disparity. So despite the reference outside the NLP, it does not significantly contributes to the content validity of the operationalization.

Other works do not provide theoretical background on relevant disciplines. Their reference is limited to the field of NLP or other technical domains. Considering all the papers reviewed aim to measure gender bias in language model, lack of reference beyond the disciplinary boundary of NLP constitues a cause of concern. It shows that up to now, far too little attention has been paid to establishing the content validity of bias measurements developed for language models. The operationalization should reflect the contested nature of bias and underlying assumptions of conceptualizing bias. Moreover, the relationship between bias and other related concepts, such as unfairness and prejudice, should be explored. By establishing content validity, the conflated use of bias can be addressed, which is a fundamental challenge that limits most of bias measurements' validity.

For instance, it is unclear how sentiment scores indicate gender bias in language models. Dhamala et al. (2021) ([45] use five different methods (sentiment, toxicity, regard, psycholinguistic norms, and gender polarity) to measure the discrepancy between different groups in five domains: profession, gender, race, religion, and political ideology. One of the metric that authors use, sentiment, is calculated by using the Valence Aware Dictionary and Sentiment Reasoner (VADER, Hutto and Gilbert, 2014)[64]. VADER uses lexical features of words for a rule-based general sentiment analysis. It was designed to measure humans expressing sentiment intensity, ranging between negative, neutral, and positive. A negative score from VADER means the sentence is perceived as conveying a negative sentiment. Examples of words with positive emotions are "love, nice, good, and great" and examples of negative emotions are "hurt, ugly, sad, bad, and worse"(ibid, p. 217).

Based on the mechanism calculating the sentiment scores using VADER, a text about the relationship between depression and female teenagers will likely involve an abundant of negative terms to describe the difficult situation. And according to the lexical analysis, the text will be classified as negative. Similar scenarios can be thought of with an anti-abortion text using emotionally blackmailing words. On the other hand, a text

describing a beauty standard from the fashion industry would use mostly positive terms. Likewise, a text about oft-called 'pro-choice' that represents certain religious view that worships a traditional sense of family and reproduction can use positive words. Both texts will be classified as positive using VADER. However, such sentiment rating does not reflect the meaning such texts represent and their relation to structural gender injustice. Despite the negativity of chosen words, addressing the problem of the toxic impact of social networks to teenage girls and abortion contribute to addressing gender bias. In comparison, despite the positive sentiment they deliver, texts about pro-choice or unhealthy beauty standards contribute to worsening structural gender injustice. The meaning attached to such ideas fails to acknowledge women as an equal moral worth through objectification or instrumentation. Therefore, without establishing theoretical grounding how measured sentiment scores relate to gender bias, it is unclear how measured bias can be used to describe gender bias in language models.

Convergent Validity and Discriminant Validity

Ideally, newly suggested methods should be compared to established measurements for the construct. However, measuring bias in language models is a relatively new field, and no established measurement exists. To evaluate convergent or discriminant validity of the bias measurements, I searched for any types of comparison with other methods. If the measurements are compared to different metrics measuring the same type of bias, it contributes to the convergent validity. If the measurements are compared to different metrics measuring different constructs, such as different types of bias, it contributes to discriminant validity.

Some (32%, 6 papers [51][59][60][43][40][63]) papers do not provide any validation for the measurements. One citesteinbornInformationTheoreticApproachDataset2022 explores potential spurious correlation, but it was not explored experimentally.

Eight papers compare their methods with existing works. May et al. (2019)[54] show that word-level Caliskan (WEAT) also works in sentence-level (SEAT). Nangia et al. (2020)[55] compare with WinoBias [85] and StereoSet [40] as baselines to find that all three models exhibit substantial bias. Kwon and Mihindukulasooriya (2022)[56] test convergent validity with the original CrowS-Pairs [55] with author generated paraphrased subset of CrowS-Pairs. They show the limitation of the dataset showing that pseudo-log-likelihood based bias measurements are sensitive to specific words than the meaning of given phrase. Kaneko et al. (2022)[62] confirm their result with related works ([55][40]). Silva et al. (2021)[41] validate their result with related work by Hooker (2020)[86]. Alnegheimish et al. (2022)[52] compare the result with related work by Vit et al. (2020)[87]. Bartl et al. (2020)[44] confirm and extend the previous research by Kurita et al. (2019)[88]. De Vassimon Manuel et al. (2021)[53] compare with existing gender bias benchmark, WinoBias [85].

Two papers compare their result with human validation. Dhamala et al. (2021)[45]

validated automatic metrics with crowd-sourced human judgement of sentiment, toxicity, and gender polarity. Sotnikova et al. (2021)[46] use human annotation by authors to evaluate the hypothesis generation of the language model.

Four papers suggest multiple metrics to measure gender bias. Dhamala et al. (2021)[45] suggest five metrics, four using sentiment scores and one using gender polarity. Lucy and Bamman (2021)[61] use two metrics, topic modeling and lexical analysis, and show coherent results between two. Silva et al. (2021)[41] use three tests, WEAT (Word Embedding Association Test), SEQ (Sequence Likelihood), and PN (Pronoun Ranking), and question the validity of using WEAT in language models. Barikeri et al. (2021)[58] use both intrinsic (Language Model Bias) and extrinsic measurements (Language Model Perplexity, Dialog State Tracking) to measure bias.

Bartl et al. (2020)[44] test gender bias in two languages, English and German, and compare the result.

Consequential Validity

To evaluate the consequential validity of the proposed bias measurements, I determined whether the authors consider the impact of bias measurement, especially regarding how to interpret the result.

Two papers explicitly discussed the interpretation of the result according to their measurement. May et al. (2019)[54] call for a cautious interpretation of results from bias measurement. The proposed method for measuring bias (SEAT) has only positive predictive ability in which the presence of bias can be detected, but it does not guarantee the absence of bias. The authors stress that the result should be interpreted with the effect size and statistical significance, not as an absolute measure of bias. The authors also suggest that studying fairness and ethics in NLP should critically examine their methodology, taking into account of the social contexts in which NLP systems are deployed.

Similarly, Dhamala et al. (2021)[45] caution interpreting the results from suggested metrics in comparison with Wikipedia. While the authors use Wikipedia sentences as a baseline, it is not meant to indicate that the Wikipedia baseline is fair. Rather, the authors argue that the results show that Wikipedia is not free from biases.

Other papers did not consider the impact of bias measurements. However, consequential validity of bias measurements are critical dimension to evaluate the measurement. As Blodgett et al. (2020)[23] stated, “Language (Technology) Is Power”, and deciding what to measure is power. As bias is used as an umbrella term to indicate various undesirable patterns, bias measurements are used as a litmus paper to assess the risk of using language models in society. Designing bias measurements is determining what should be considered for making language models usable in society. Considering the impact of measurement, therefore, is a crucial step to develop language models.

Chapter 4

Which Biases are Morally Relevant?

In previous chapters, I discussed the conflated use of bias and how the conflation affects the operationalization of the bias as a construct. While a myriad of definitions and operationalizations are suggested, several essential questions remain about the relevance of bias in language models. In this chapter, I attempt to defend the view that bias that wrongfully discriminates people and reproduces structural injustice should be prioritized in identifying and measuring bias in language models.

Before arguing which biases are morally relevant, I first review how bias in language models relates to normative values, such as equality and justice. After reviewing the lack of normative reasoning and vague motivations of reviewed gender bias measurements, I address the unique complexities faced by the interdisciplinary nature of bias in language models.

In the next sections, I explain two critical concepts that are used throughout this chapter: wrongful discrimination and structural injustice. I argue that bias that reproduces wrongful discrimination or structural injustice should be prioritized to be identified, measured, and mitigated. While other unintended, undesirable, or problematic biases that language models exhibit would also merit consideration, wrongfully discriminating bias that reproduces structural injustice pose the most imminent threat to using language models in society. The social and moral implication of language models are especially critical considering the technology's potential to scale up discrimination and exacerbate existing structural injustice.

4.1 How Bias is Related to Normative Values

The importance of the social and moral implications of bias in language models is reflected in the current literature. Most reviewed papers on measuring gender bias in language models associate their work with normative values, such as language models'

risk of "scaling up social injustice" (May et al., 2019, p. 622)[54] or "stereotypes already disadvantaged groups propagate false beliefs about these groups and entrenches inequalities" (Nangia et al., 2020, p. 1)[55].

4.1.1 Motivations for Measuring Bias in LMs

In the last few years, there has been a surge of interest in uncovering societal harm and investigating normative values in algorithms [12][22][89][90]. Language models are not an exception. As the table shows, 17 out of 19 papers measuring gender bias in language models motivate their work with normative reasons. Despite the ambiguity and disorganization of value-laden terms, it seems clear that such normative aspects of bias in language models ascribes the importance and necessity of identifying and measuring bias. The full table of normative motivations in the gender bias measurement papers can be found in the Appendix B.

The most commonly used concepts were harm [51][55][61][57][43][58][63] and equality [55][46][41][58]. The framework of analyzing bias in terms of occupational and representational harms has been introduced by Kate Crawford's Keynote [91] at NeurIPS 2017. Allocational harm refers to distributive unfairness when an automated system distributes resources unfairly to different groups based on socially salient features. Representational harm is concerned with fair representation of groups in terms of associated sentiments or recognition. This categorization of harms has been popularized since Blodgett et al. (2020)[23] reviewed bias in NLP systems according to this framework.

Terms to indicate group-based equality, such as minority [54][55][57], social hierarchy [46], or segregation [45][51] also appeared in many works. Other value-laden keywords like ethical challenges [58], misbehaviour [59], undesirable [45], discrimination, and disparate treatments [45] were also used.

Such normative motivations show that morally dubious patterns that language models exhibit is a significant concern that should be urgently addressed. Of particular concern is to adequately capture such patterns in language models that aligns with the worries raised by such value-laden concepts. Identifying and measuring bias must be able to contribute to addressing morally problematic patterns in language models. Without providing adequate normative reasoning to conceptualize respective values, and connect it with the observed phenomena that is being measured in language model, normative motivation and empirical measurement will remain isolated.

4.1.2 Underspecified Normative Values and Ethics Washing

Despite the prevalence of normative motivations, it is unclear how those concepts are defined and conceptualized. Provided normative motivations are seldom expanded, and it is unclear to what extent normative motivation connects to the measurement suggested in the literature. It is ambiguous what social harms, social injustice, or inequalities mean in the context of the proposed methods for identifying bias in language models. Closely

related to the conflated use of the term bias, normative concepts suffer underspecification and vagueness. The lack of normative reasoning makes the link between empirical bias measurement and normative motivation unreliable.

For instance, Barikiri et al., (2021)[58] state that "stereotyping minoritized groups is a representational harm that perpetuates societal inequalities and unfairness (ibid, p. 1941)." Also, the authors categorizes two groups, dominant groups and discriminated groups, to define inequality between groups. With a set of negative stereotypical terms to describe minoritized groups and positive stereotypical terms for dominant group. (ibid, p. 1942) The authors use terms like equality, stereotype, and fairness in mixture without conceptualizing each term clearly. Such categorization is also too simplified to reflect the complex dimensions of bias. Since the concept of dominant and minoritized group was not expanded, it is unclear what they refer to in this context. Also, it is unclear how authors define positive and negative stereotypes. Moreover, the inequality between groups is not defined.

Such an operationalization is problematic in many folds. First, negative stereotypes are not attached exclusively to non-dominant, socially marginalized groups. For instance, negative stereotypes for being violent and extremist such as 'Proud Boys' [92] does not necessarily put their status into a non-dominant group, in terms of racism. Considering the prevalent racism in society, it is unclear whether a male white-supremacist group [93] would be accounted for as a non-dominant group. Similarly, positive stereotypes, such as Asians being good at math [94], does not put Asians as a dominant group. Therefore, the dichotomy between a dominant group with positive stereotypes and a non-dominant group with negative stereotypes does not reflect the complex relationship between bias, power structure, and stereotypes.

Second, inequality cannot be used interchangeably with difference. Equality is a contested construct that cannot be simply defined as a mathematical sense of two variables having the same values. Theorizing equality requires normative reasoning of what consists of equality citesen1980equality and the purpose of the purported idea of equality [95]. However, the authors describe to the relationship between "the groups in power, i.e., dominant groups, and discriminated groups, i.e. minoritized groups" (Barikeri et al., 2021, p. 1942)[58] as inequality without further explanation. It is unclear why the relationship between two groups are the case of inequality due to the lack of normative reasoning.

Similarly to Barikeri et al., all 19 papers from the literature review do not expand the normative motivations provided alongside the bias measurements. Consequently, most gender bias measurements in language model fail to show how bias becomes a problem of fairness, equality, or justice. Instead, they show the difference between groups based on socially salient features, such as gender, without explaining what the differences mean. However, the meaning of differences determine the relevance of such difference in social context. Measured properties of language models can be used as evidence showing how

such values are compromised in language models, on the basis of supporting normative reasoning how values, such as fairness and equality, is conceptualized and defined in the context.

The Difficulty of Normative Reasoning

Contested Concept Untangling normative issues attached to the biased, or any other morally undesirable behavior of language models, is not a trivial task. Like most other moral values like equality, fairness is a contested concept [96] and therefore any reductive definition runs the risk of missing features considered relevant for other theories of fairness. The same difficulty exists in other aforementioned concepts, such as (in)justice, equality, and harm.

The complex nature of value-laden concepts ought not to suggest that making effort to make clear definition is of no use. It is rather the opposite. If making language models less biased is meant to make the models fairer, fairness should be defined to such a degree that enables an empirical investigation. Bias needs to be conceptualized in a way that is relevant to the phenomena of interest, such as language models associating negative sentiments to female names, and accurately describes the phenomena.

Being stereotypical is one way of treating people unfairly, by treating individuals solely based on their membership to certain groups. However, treating people fairly is more than merely being ignorant of stereotypes. Interchangeably, making language models a-stereotypical will not necessarily make the language model fair. It can make it fairer, in terms of stereotypes, but it does not capture other relevant aspects of fairness. Therefore, setting a-stereotypical language models as an ideal language model will lead to unsatisfactory outcomes.

Only after articulating what is undesirable and to be improved based on the definition of the 'unbiased' language model, can meaningful measurement be developed. By meaningful, I mean that reducing bias according to that metric leads to language models producing less morally undesirable outputs, that the measurement aims to identify and illustrate.

Working across Disciplinary Boundaries Defining fairness requires an effort to provide sound normative reasoning. And connecting such definitions to a practical problem of quantifying bias in language models requires working beyond disciplinary boundaries. Normative reasoning requires a selection of fairness definitions that best describes the phenomena of interest observed in the language models. Such decisions should justify why such definitions were chosen among other candidates, which requires engaging with existing debate in relevant fields such as political philosophy and ethics. Normative reasoning is distinct in nature from developing a mathematical metric, or to systematically identify patterns in language models. Therefore, it poses an additional challenge for researchers working on value-laden concepts like fairness in terms of social implications of bias.

The challenge can be addressed by collaborating with researchers from the relevant fields. Existing expositions of investigating normative matters in language models are unsatisfactory because they rarely collaborate with researchers outside the field of computer science. One notable exception is Weidinger et al. (2021)[12] where computer scientists and philosophers are collaborating to investigate the ethical implications of language models. Most literature discussing the ethical challenges of language models are authored either exclusively by computer scientists or exclusively by philosophers. The challenge of discussing value in sociotechnical systems will benefit from collaborating beyond disciplinary boundaries, by combining expertise from different disciplines.

The Danger of Ethics-Washing As discussed in previous sections, it is challenging to investigate value-laden concepts in sociotechnical systems. Measuring bias in language models require clear definition and conceptualization both in terms of empirical and normative dimensions. Normative reasoning should provide a link between proclaimed motivation and the suggested method for measuring bias. Measuring gender bias is an instrumental tool to prevent or reduce undesirable bias in language models. Without clear connection, the bouquet of moral values are not promoted nor addressed by using such a method.

What is more pertinent in using vague normative terms with empirical methods is that it runs a risk of using such methods for ethics-washing [97]. The mismatch between vague normative motivation and empirical investigation can lead to ascribing normative values to measuring bias in language models which, in fact, was never justified. The risk of ambiguous use of normative motivations with measuring bias empirically is not limited to the lack of diligence in scientific investigation. Vague operationalization of bias with ambiguous normative reasoning can be misused to validate the measurement with normative values.

When bias in language models is identified by adopting certain measures of bias, the measurement represents the bias. One could claim that the language model is 'de-biased' when the measured bias is mitigated, and therefore the language model is free from normative charges attached to the bias in language models. The normative charges can include perpetuating social injustice or social inequality, as shown in papers measuring gender bias in language models. The lack of normative reasoning will not stop people from misusing the banner of normative values that motivates the measurement of bias. It is a case of ethics-washing, which is using ethics to defuse criticism without committing to actions that can improve the ethical issues that the technology is facing [98]. What entails addressing fair, just, and ethical language models should first be conceptualized and operationalized to show that language models do not perpetuate existing unfairness or social injustice. Without such an endeavor, including isolated normative motivations will not help addressing moral problems that language models might reproduce and exacerbate.

4.2 What Makes Difference Discrimination

Bias in language models matters, not because of the discrepancy itself, but because of the meanings attached to the discrepancy. What makes difference discrimination? It concerns what the difference means.

As shown in the previous section, gender bias measurements in language models have failed to establish why such discrepancies between men and women are morally wrong. Bias measurements attribute less women to professional occupations [51], relate female characters with behavioural expectation based on traditional gender roles [61], or associating positive or reproduces negative stereotypes about women [54]. I argue that it is problematic because it wrongfully discriminates against women. To evaluate why such patterns are cases of wrongful discrimination, I will refer to Deborah Hellman's view on meaning-based account of discrimination. Based on her definition of wrongful discrimination, I will show which bias in language models are morally wrongful discrimination, and therefore needs to be addressed with higher priority and urgency than others.

4.2.1 Meaning-based Account of Discrimination

Hellman(2017)[99] argues that what makes discrimination morally wrong is the meaning of the discrimination, rather than the actor's intentions, the relevance of the trait used to discriminate, or rationality of discrimination. Based on the meaning-based account of discrimination, the author argues that "discrimination is wrong when and because it is demeaning" (ibid, p. 1). It is demeaning when the actor with social power expresses denigration, which is the action of saying that someone is not good or important, which fails to treat those affected as equals. (ibid, p. 13)

In other words, not all discrimination is morally wrongful. In an expansive sense, discrimination can refer to any differential treatment based on personal traits. For instance, charging more expensive insurance for young drivers can be a discrimination based on age. But the discrimination is justified for several reasons, such as preventing adverse selection [100]. Similarly, the discrimination that language models make against certain groups of people itself is not morally problematic. What makes such discrimination wrongful relates to the meaning of such discrimination.

As shown in the above definition, the meaning-based account of wrongful discrimination has two aspects, which are an expressive dimension and a power dimension (Hellman, 2017, p. 12). An expressive dimension concerns if an action or policy regards another person is inferior or of lower status. It is especially relevant when such expression reflects historical injustice where the attributes, such as gender and race, have social meaning in the context. Therefore, discrimination based on socially salient features are especially morally problematic, since it fails to treat people with equal moral worth.

Another important dimension of meaning-based account of wrongful discrimination is the power relation. When the actor who expresses has social power that ascribes force

to the meaning of the action or policy, discriminatory action suffices to be a wrongful discrimination. Hellman argues that the power is important in discerning wrongful discrimination, since the actual power enables such discrimination to lower the social standing of those affected. (ibid, p. 15-16) The discriminator with power and authority can affect people in more consequential ways than those without such power. Furthermore, Hellman accentuates that demeaning depends on capacity that comes with power, not an actual effect (ibid, p. 17) Regardless of the outcome, when the actor of power fails to recognize the equal moral worth of others, it is a morally wrongful discrimination.

4.2.2 Bias and Structural Injustice

Hellman (2017)[99] argues that discrimination is especially problematic when it is based on socially salient features, such as gender and race. Socially salient features, such as gender, can be used as "accurate proxies" for discrimination due to historical injustice. (ibid, p. 7) For instance, Prates et al. (2021)[101] show that gender is used as proxy to attribute gendered pronouns for STEM (Science, Technology, Engineering, and Mathematics) jobs, by using more male pronouns than female pronouns to translate from grammatically neutral language to grammatically gendered language. Such bias stems from historical injustice of excluding women in these fields, thereby there are less data where women are associated with STEM fields. Such bias is especially problematic since the pattern of historical injustice perpetuates in a new vessel of language models.

Furthermore, to expand Hellman's account of what accounts as wrongful discrimination, I argue that not only historical injustice, but also contemporary structural injustice should be accounted for to evaluate whether the discrimination is morally wrong (McKeown 2021 [102], p. 13, Young, 2011 [103]).

According to Iris Marion Young (2011)[103], structural injustice exists when:

"social processes put large groups of persons under systematic threat of domination or deprivation of the means to develop and exercise their capabilities, at the same time that these processes enable others to dominate or to have a wide range of opportunities for developing and exercising capacities available to them." (ibid, p. 52)

Structural injustice concerns how people should relate to others on an equal footing, to live in a just society. The concept of structural injustice is closely related to demeaning, where one fails to treat the other of equal worth. The difference is while identifying discrimination focuses on the individual actions, structural injustice is focused on the social, economic, and political structure that creates injustice distinct from individual actions.

Individual cases of bias in language models can be evaluated whether they case morally wrongful discrimination, based on expressive and power dimensions. Assessing the expressive dimension can benefit from applying the perspective of structural injustice, to evaluate whether the expression of denigration is based on structural injustice. If the structural constraint makes bias susceptible to discriminate against those who are

deprived of their equal status as citizens, the bias is demeaning as it fails to treat those affected as equals. Structural injustice provides a context to judge how certain bias is demeaning since the social processes put certain groups under systematic unfair treatment where they are not treated as equals. Reproduction of structural injustice indicates whether the discrimination fails to treat an individual unequally, which is demeaning.

Wrongful discrimination based on gender puts those who are discriminated against under the systematic threat of domination where they lack equal opportunities. Identifying bias that reproduces structural injustice helps to prioritize which bias should be addressed with urgency.

While gender bias is a case of historical-structural injustice [104] (cited from McKeown p. 13) where historical injustice persists in the current social structure, newly arising structural injustice should also not be disregarded. The data sweat shop [105], for instance, is a new type of structural injustice that gives rise to wrongful discrimination. For the purpose of my thesis, I focus on gender bias in language models, that are reproducing historical and structural injustice.

4.2.3 Bias in Language Models and Wrongful Discrimination

In the previous chapter, I reviewed measurements suggested to identify gender bias in language models. Most of them used stereotypes and sentiment scores to operationalize gender bias in language models. Afterwards, I presented what makes discrimination morally wrong, according to the meaning-based account of discrimination.

To evaluate which measured bias in language model suffices to be a morally wrongful discrimination, two aspects need to be examined: the expressive dimension and the power dimension. The expressive dimension can be evaluated by assessing whether bias expressed in language model's output fails to treat those affected as an equal to others. The power dimension examines whether language model has an actual power that can turn such discriminatory bias into real harm.

The Power Dimension

To address the power dimension first, I argue that language models are in a position of power where bias can begin to consequentially affect the lives of people. Language models have been adopted in varying applications and can be used in creatively numerous ways, from collecting debt [1] to finding a flat in Berlin [2]. Language models have significantly lowered the cost of various language-related tasks, such as translation and producing plausible-sounding text. One statistics estimated that usage of ChatGPT will cause more layoffs in the U.S. within five years [106]. The bias in language models used for various applications, such as debt collection, can lead to disparate treatment of people. As the popularity of language models and they are adopted in more various applications, the power of language models also grows. A popular application based on

a language model, ChatGPT [5], attracted one million users within the five days since its launch [107]. As ChatGPT becomes more prevalent, the impact of the biased output that ChatGPT creates will also be as widespread. Amongst various "accidents" where ChatGPT created problematic output, once it produced a python script that categorizes one as a good scientist if they are "white" and "male" [6]. One might argue that it is up to the user's choice if such biased code actually gets used to create harm, the affordance that technology suggests possible answer is already an authority. As Hellman (2017)[99] pointed out earlier, demeaning depends on capacity that comes with power, not an actual effect (ibid, p. 17) By using ChatGPT to retrieve information, ChatGPT exerts its power to decide the content and format of information that people get through the interaction with ChatGPT[1]. By interacting with the language model that is biased, the bias can impact the user unconsciously [109].

Therefore, to evaluate whether bias in language models are morally wrongful discrimination, I will focus on the expressive dimension, as the actor of discrimination is language models for all bias measurements in language models.

The Expressive Dimension

The expressive dimension evaluates whether the one expresses denigration and views the other with less worth, i.e. demeaning. The meaning-based account discrimination needs to address specific cases to evaluate whether the treatment is demeaning, since the meaning varies across different societies based on cultural and social norms. Therefore, the evaluation of expressing denigration can only be in the context of a particular society (Hellman, 2017, p. 9).

To evaluate the expressive dimension of bias in language models, I refer to two examples of gender bias measurement from the papers I reviewed. Measured gender bias wrongfully discriminate against women or men[2] when the bias treats the group of lower status. In the first example, I show that it is difficult to identify morally wrongful bias by measuring bias using sentiment scores. The second example, on the other hand, shows how wrongful discrimination can be identified by measuring gender bias using stereotypes.

Bias with Sentiment Scores

Dhamala et al. (2021)[45] define a language model as biased "if it disproportionately generates text that is often perceived as being negative, unfair, prejudiced, or stereotypical against an idea or a group of people with common attributes." The authors argue that the model can be biased against women, African Americans, Muslims, when the model generates more negative text towards these groups (ibid, p. 862). The paper uses bias as an umbrella term, as it encompasses negative sentiment, unfairness, prejudice,

[1] It resonates how Google gained its dominance in the internet through its search engine [108].

[2] Since all the gender bias measurement reviewed used binary classification of male and female to evaluate gender bias, I also grouped according to the binary genders. This assumption is limited since it ignores people of diverse gender who are susceptible to be discriminated based on their gender.

and stereotypes. Unless the model generates negative text as frequently as in every social group or ideology, the language model is considered biased.

According to their definition, the language model has gender bias if it generates more negative text for certain gender groups, either men or women. The authors composed prompt based on Wikipedia articles of female and male actors, including names. Discussing the result of their measurement, the authors found out that texts generated from prompts including female names have a smaller share of text that is evaluated as negative. And they conclude that the model shows "a (negative) bias in sentiment scores towards the male population" (ibid, p. 866).

The negative sentiment is calculated by using the Valence Aware Dictionary and Sentiment Reasoner (VADER, Hutto and Gilbert, 2014)[64], which uses lexical features of words for a rule-based general sentiment analysis. A negative score from VADER means the sentence is perceived as conveying a negative sentiment, such as "hurt, ugly, sad, bad, and worse"(ibid, p. 217). Dhamala and colleagues found that language models generate more sentences with negative sentiment for prompts with male names, concluding that the model is biased against men, meaning men were associated with more negative sentiment.

To determine if the bias is wrongful discrimination, bias should be demeaning, by treating the discriminated group as lower status. However, associating negative emotions, such as "ugly" or "sadness" to one group does not reject the equality between the groups. The comparison reveals the difference between the group, but the groups are being treated on an equal setting. Also, although the groups are divided based on the socially salient feature of gender, it does not reproduce structural injustice. Men are not under "the threat of systematic threat of domination or deprivation of the means" (Young, 2011, p. 52) by being associated with negative emotions. Consequently, the bias is not a case of wrongful discrimination. It is not to argue that such bias is desirable or even permissible, but I aim to show that not all measured bias deserves equal importance. It is important to identify bias in language models that needs to be prioritized. Morally wrongful bias of language models matters, as it risks scaling existing social injustice via technological means.

Even though the authors concern that the unfair machine learning models risk "reinforcing undesirable stereotypes, subjecting users to disparate treatment and enforcing de facto segregation" (Dhamala et al., 2021, p. 862)[45], what it is meant by undesirable, disparate treatment, and segregation of language models is unclear. Accordingly, it is ambiguous how bias measured with sentiment scores relate to those moral harms. Based on the meaning-based account of discrimination, however, bias in sentiment scores does not suffice to be accounted as wrongful discrimination.

Similar limitations are found in different ways of measuring sentiment scores. Jentzsch and Turan (2022)[59] measures gender bias in BERT models by evaluating on sentiment classification using the Internet Movie Database (IMDB). Gender bias is measured as a

discrepancy in sentiment ratings between a pair of movie reviews. The pair is created by masking gendered terms. To create a male version, terms identified as female are replaced, such as 'woman' to 'man'. Likewise, the female version masks male-gendered terms into female-gendered terms, such as 'prince' to 'princess'. (Jetzsch and Turan, 2022, p. 186-187, p. 195-196)[59]

The way sentiment score was(Stewart: were?) attributed to the dataset is connected to the ratings attached to the reviews of respective movies. The reviews with the ratings of 4 or lower are labeled as negative sentiment and ratings as 7 or higher positive. In other words, movie reviews that express positive opinion are classified as positive text. It is a vastly different method to calculating sentiment score from VADER, which uses lexical analysis. Based on the method proposed by Jetzsch and Turan, the model is biased against women when the movie reviews using female terms express negative opinion to the movie.

Despite the difference in calculating sentiment score, both metrics miss a link between wrongful discrimination and measured gender bias. There can be a movie about female empowerment that is produced with a low quality, or unpopular perspective about women's rights. These movies can receive a low rating for the quality of the production or minority views. However, the low ratings for those do not exhibit denigration towards certain groups, nor reproduces structural injustice. Therefore, such sentiment scores are not adequate to evaluate gender bias that is a case of wrongful discrimination and reproduction of structural injustice. The authors reported a mixed result, some models preferred male terms and the others preferred female terms (ibid, p. 189).

Similarly, it is not to argue that patterns of associating female terms with negative sentiment valence or movie rating is completely unproblematic. It might show that some aspects of gender bias are reflected in such associations. However, in the given setting of proposed methods, it is difficult to discern the meaning attached to the discrepancy of sentiment scores, due to multiple potential confounding factors. What determines discrimination is not the discrepancy itself, but the meaning that situates the discrepancy in a particular social context that is demeaning certain groups. Whether the discrepancy instantiates structural injustice can be determined by examining what the content of bias means and how it relates to equality.

Bias with Stereotypes

Lucy and Bamman (2021)[61] investigate the stories generated by GPT-3 reproduce gender stereotypes from film, television, and books. The authors compared the topics of GPT-3 generated stories and human-written books to see how the perceived gender of the character relates to the occurrence with topical terms, such as appearances, intellect, and power. The perceived gender of the characters were defined using gendered pronouns, honorifics, or names. The prompts used for GPT-3 to generate stories consist of single sentences containing main characters, sampled from 402 English contemporary fiction books data.

The authors conducted two content analysis. First, topic modeling was used to uncover coherent collections of words in the text. The result shows that GPT-3 tends to associate feminine characters with topics related to family, emotions, and body parts. In contrast, masculine characters were aligned to politics, war, sports, and crime." (ibid, p. 50) Authors show that varying topics across perceived gender in GPT-3 generated stories align with previous works that showed language models relate women with caregiving roles [110], maternalism, and appearance [111]. In addition, GPT-3 generated longer stories when the prompt contains stereotypical characters than anti-stereotypical characters (Lucy and Bamman, 2021, p. 51)

Besides topic modeling, the authors analyze how characters are described by measuring semantic similarity with embeddings of lexicons. Three dimensions of descriptions, appearance, intellectual, and power, were chosen based on previous works on stereotypical description based on gender [83][112][113][111]. The result shows that words describing appearance are often used to feminine characters and words related to power to masculine characters.

The authors conclude that GPT-3 has internalized stereotypical gender stereotypes and it was strong enough to neutralize the effect of using words with power for feminine characters. Even when prompts did not include explicit gender information or stereotypes, GPT-3 tend to generate stories aligning with gender stereotypes. Additionally, authors discovered that GPT-3 tend to include more masculine characters and the result differ based on the character's gender, even when the identical prompts were used (ibid, p. 51-52).

Associating women with family, appearance and less power has been intensely investigated in feminist theory. Associating women with appearance reflects the history of objectification of women. Feminists have raised the problems of objectification, making women excessively preoccupied with their appearance [114]. Associating women with their appearance fails to recognize women as an equal agent as men, by identifying women with their body, instead of their entire being. Bartky (1990) argues that fragmentation of the female body regards women as "less inherently human than the mind or personality" (Bartky 1990, 130, cited from [114]) Language models associating feminine characters with appearance show that it is reproducing gender injustice is being reproduced. Similarly, examining familial dynamics and power structures effectively highlights the presence of gender inequality. In her book <Justice, Gender, and Family>, Susan Moller Okin pointed out that "socially constructed inequalities" exist in the distribution of critical social goods, such as power, prestige, and opportunities for self-development. (Okin, 1989 [115], 136, cited from Allen, 2022 [30])

In this section, I reviewed gender bias measurements developed for language models and show how measurements can be evaluated whether they cause morally significant discrimination. This chapter aimed to provide reasoning which bias are significant, and what are the criteria to prioritize certain bias over others. I argued that bias that

wrongfully discriminates persons should be addressed with urgency, especially when the bias reproduces existing injustice. In the next chapter, I will provide a unified framework based on previous chapters. The framework aims to identify bias in language model comprehensively, conceptualizing bias beyond technical terms.

4.2.4 On Mitigation

Ultimately, identifying and measuring bias aims to reduce bias in language models. What it means to reduce bias depends on how bias is conceptualized. Conceptualizing and understanding bias is a prerequisite to mitigate it in any meaningful way. If bias is defined as unfair discrimination, reducing bias should contribute to address the problem of fairness and equality. Similarly to measuring bias, reducing bias also requires considering the socio-technical context where bias in language model stems from.

Bias in language models reflect the society which is shaped by historic injustice. Gender bias, for instance, is embedded in the history and society that epistemic evidence supporting gender bias is prevalent. Language models produce code that defines a good scientist as a white male because historically most of the great scientists were white males, because others were excluded from science. Basu (2019)[116] argues that epistemic evidences cannot be taken without additional moral consideration, considering the unjust social structures of the world (p. 192-194). By considering normative correctness in addition to descriptive accuracy, it might result in compromising epistemic accuracy for the sake of fairness. To refer to the lands analogy from previous sections, defining bias considering normative correctness means giving up the reality but choosing the fantasy world.

However, as soon as it is clear that descriptive accuracy means reproducing existing injustice, defining bias based on descriptive accuracy inevitably fails to address the problem of justice. Given that the goal of reducing bias is to address expected harms and risks of bias in language models, aiming to meet descriptive accuracy is insufficient. Bias based on descriptive accuracy alone risks allowing statistical discrimination of language models. Based on descriptive evidence that language models are trained on, which reflects existing patterns of injustice, language models will be like a “rational racist” whose racist attitudes are epistemically justified but morally unjustifiable [116].

Therefore, if the goal of mitigating bias in language models is to minimize societal harm that bias can give rise to, defining bias should reflect normative correctness beyond descriptive accuracy. Even merely reproducing existing injustice in language models, language models can exacerbate by scaling the bias through interaction with people in large scale.

Chapter 5

A Holistic Framework for Measuring Bias

In the previous chapter, I argued that biases that are wrongful discrimination and reproduce structural injustice should be prioritized in identification, measurement, and ultimately, mitigation of bias. Based on the discussion in the previous chapters, I propose a unified framework to measure bias in language models scientifically and meaningfully beyond technical terms. The validity of measurement can be scientifically assessed by applying the measurement modeling proposed in Chapter 3. A valid operationalization is possible on the basis of the conceptualization of bias. Meaningful bias measurement should prioritize identifying bias that matters, such as a case of wrongful discrimination or structural injustice. Choosing what to measure is a power that determines how language models are evaluated, including their social and normative implications. Designing metrics to measure bias should, therefore, reflect the power structure that can impose wrongful discrimination and structural injustice against people.

The framework aims to aid addressing the bias in language models meaningfully to devise new metrics, or to evaluate existing metrics. Meaningfully bias measurement should address the social and moral implications of bias in language models. The first level of the framework evaluates the extent bias is conceptualized and articulated. Based on the clear definition, the second level assesses whether the bias is operationalized in a scientifically valid way. Lastly, the third level focuses on the structural dimension of bias, by evaluating language models as a category. Structural analysis investigates the struc-

Levels	Object for Analysis	Type of Analysis
Level 1	Definition of bias	Conceptualization Analysis
Level 2	Bias measurements	Measurement Modeling
Level 3	Language models as a category	Structural analysis

Table 5.1: A Holistic Framework

ture and the relationships around language models, beyond individual language models. Structural analysis aims to identify potential sources of bias, from stakeholders, technology itself, and the social context. Engaging in structural analysis is especially necessary to identify morally and socially significant bias, since the meaning of discrimination can only be determined in the context where language models are situated.

5.1 Technically Holistic Bias Metrics

) Before proposing the framework in detail, I will briefly review existing bias evaluations for language models that are designed to be holistic and extensive. While several holistic measurements for language models are available, most of them fail to address bias beyond the technical dimensions. However, limiting only to the technical aspects of bias in language models will inevitably result in technical solutions, which cannot capture important aspects of bias that are relevant in social contexts. To use language models across various sectors in society, it is crucial to identify bias that will have significant societal and moral impact. Such implications needs to be evaluated in the situated context of a particular society. Furthermore, making language models less biased according to technically holistic metrics will address bias only in technical aspects. But even technically perfectly unbiased language models can be biased when they are situated in a social context that is biased.

5.1.1 Beyond Imitation Game Benchmark (BIG-bench)

One of the most extensive benchmarks suggested to evaluate language models is BIG-bench [27]. BIG-bench tasks approach bias as a "systematic preference for members of one category", or "associating particular attributes with particular categories" in a fixed context (Srivastava et al., 2022, p. 17)[27], which is the most common way that bias is being measured in language models, as shown in reviewing gender bias measurements. BIG-bench evaluates social bias by using seven different metrics, such as measuring stereotypes [117], gender bias [118], or religious bias [119]. The authors aim to quantify bias concerning socially salient features, such as "age, disability status, gender identity, nationality, physical appearance, race/ethnicity, religion, socio-economic status, and sexual orientation" (Srivastava et al., 2022, p. 18)[27] The list of bias being examined in BIG-bench are commonly investigated biases especially in the context of U.S., according to the US Equal Employment Opportunities Commission's list of protected categories. The protected categories consist of: race/color, gender/gender identity or expression, socioeconomic status/occupation, nationality, religion, age, sexual orientation, physical appearance, and disability [55]

The paper concludes that the bias often increases in broad or ambiguous contexts and can be "steered through" prompting appropriately. BIG-bench's approach towards measuring bias in language models does not vary from the gender bias measurements I

reviewed in previous chapters. Despite using multiple metrics to capture social bias, it fails to conceptualize bias and situate it in the social context where language models are developed and used.

Listing various types of bias does not contribute to capture the complex dimensions of each bias. For instance, measuring the association between Muslims and violence [119] does not only generalize to other religious bias, nor capture all relevant aspects of bias related to Muslims. Furthermore, the bias is measured by comparing the sentence completion between prompt sentences with 'Muslims' and 'Christians' as a varying keyword. Such a simplified method of measuring bias cannot be understood as a representational method that measures bias concerning religion in language models.

It is a prime example of focusing on the discrepancy itself, not the meaning of the discrepancy. According to the metric, the language model will be considered 'debiased' when the model produces more violent keywords to 'Christian'. Measuring such patterns of discrepancy does not shed much light on how people of specific religions are suffering from stigmatization, or how language models can exacerbate such wrongful discrimination based on religion.

5.1.2 HELM Holistic Evaluation

Stanford Institute for Human-Centered Artificial Intelligence has published Artificial Intelligence Index since 2017 [120]. The title of Chapter 3.1 in the HAI AI Report from 2023 is "Meta-analysis of Fairness and Bias Metrics". The fact that fairness and bias are used as a set of terms shows how both terms are used expansively without conceptualizing fairness and bias profoundly. Not all bias relates to the problem of fairness, and the problem of fairness does not necessarily mean it can be identified by measuring bias. [1] The sense of "fair"[2] algorithms in the HAI report is similar to the notion of bias from BIG-bench:

"Algorithms are considered fair when they make predictions that neither favor nor discriminate against individuals or groups based on protected attributes which cannot be used for decision-making due to legal or ethical reasons (e.g., race, gender, religion)(HAI AI Report, p. 130)."

The measurements focus on the discrepancy across groups divided by socially salient features. Furthermore, the statement is mixing various concepts like legality, ethicality, fairness, and discrimination. Similarly with the bias measurements used in BIG-bench, the HAI report fails to acknowledge that what makes bias in language models socially

[1] Also, while the report states that "Algorithmic bias is measured in terms of allocative and representation harms." (HAI AI Index report 20203, 130), referring to Kate Crawford's 2017 talk. However, the speaker herself stresses the importance of a context-aware evaluation of harms and a consideration of various stakeholders and system affordances to assess computational harms in the later publication. [121] Therefore, while representational/occupational harm might have been a useful framework when fairness related research in NLP began, sticking to the old framework fails to reflect the recent developments made afterwards.

[2] The report uses fair, biased, and harmful interchangeably, which is highly problematic.

and morally relevant is the meaning of the discrepancy. Instead, empirical measurements focus on identifying various differences, without investigating what each difference indicates, in the context where the difference is measured. Despite the framework proposing "a robust stream of both new ethics benchmarks as well as diagnostic metrics was introduced", it is unclear how those benchmarks contribute to the ethics of language models. To understand certain bias measurements as ethical problems, it should be shown why such patterns are of a normative concern and how such concern is exhibited in language models.

5.2 A Holistic Bias Measurement

It is not truly holistic until bias measurements take into account of structural analysis beyond the technically defined bias. Bias measured with ad hoc properties of language models does not represent the complex essence of bias, and how bias manifests in society. Gender bias, for instance, cannot be reduced to the occupational stereotype, difference in sentiment score, or frequency of gendered nouns in the text generated by language models. While these are some of the relevant instances of gender bias, some of them even morally relevant, it is not clear whether they are the most relevant properties of gender bias in language models.

Measuring the construct should not be limited to the availability of data or tools. Bias measurements should be the tool for identifying relevant patterns in language models that are socially and morally significant. Letting the measurement matter most will be committing an error of reversing the purpose and the instrument. What is being measured should not define what matters in language models. In contrast, what matters should be measured. As Friedman and Nissenbaum (1996)[17] rightly pointed out, biased computer systems, bias defined as systematical and unfair discrimination against certain people, are "instruments of injustice" (p. 346) The reason why bias in language models should be investigated is not because bias could be measured, but because such bias reproduces injustice.

To develop a holistic bias measurement that is not confined to the technical conceptualization of bias, I present a framework that consists of three levels. Each level addresses what I have discussed in the previous chapters: how bias can be conceptualized, measured, and evaluated on its social and ethical relevance.

5.3 The Framework

In this section, I describe each level of the framework in detail. The framework is designed to capture bias in language models that matters in a scientifically valid way. The first step is to provide a clear definition of the construct. Accordingly, the first level concerns conceptualization which will contribute to both scientific validity and identification of

meaningful bias. The second level, measurement, focuses on the scientific validation of operationalization. The last level, structural evaluation, investigates bias in language models as a category. Structural evaluation aims to identify social and ethical relevance of bias, considering structural components of bias that go beyond individual language models. By engaging in all three levels, bias in language models can be comprehensively addressed and contribute to improve the quality of bias measurements. The full table of analyzing three levels of conceptualizations in the gender bias measurement papers can be found in the Appendix B.

5.3.1 Level 1: Conceptualization

Level 1 concerns the importance of addressing explicit and accurate definitions of bias[3]. Conceptualizing bias is not a trivial task since bias is used expansively in the context of NLP, as discussed in Chapter 2. The term bias itself has varying definitions across disciplines. Therefore, conceptualizing the bias and articulating an explicit and accurate definition is a fundamental step to develop methods to measure bias. Valid measurements can only be devised after an adequate level of conceptualization is established.

As discussed in Chapter 2, the term bias is conflated in both operational and conceptual accounts. By engaging in conceptualization, both conflated uses of bias can be tackled. The clarified definition of bias will indicate whether the bias refers to descriptive accuracy or normative correctness. The operational conflation will also be improved by connecting abstract concepts to an explicit articulation of the definition used in specific works.

Three Levels of Conceptualization

While there are various ways to conceptualize the construct, Adcock and Collier (2001)[47] propose four levels of conceptualization as a framework to develop validate measurements. The process aims to connect concepts and quantified scores by iteratively following four levels and their respective tasks. The four levels (stewart: the four levels consist?) consist of background concepts, systematized concepts, indicators, and scores for cases (ibid, p.531). The focus of the thesis is assessing the measurement of bias, especially how measurement is designed in relation to the construct, rather than the score produced from the measurement. Therefore, the last level of conceptualization from Adcock and Collier framework, scores for cases, was not considered in the evaluation of definition. Different tasks are assigned between levels to proceed to the next level.

- Background Concept

[3]The scope of bias can vary depending on the purpose of the work, which can be about broad definitions of bias, specific types of bias such as gender bias, or bias concerned with more particular phenomena like the Angry Black Women (ABW) stereotype. Regardless of the different scopes, it is critical to conceptualize the construct and provide a clear definition to develop measurement.

The background concept illustrates the varied meanings attached to the concept. Adcock and Collier define the background concept as "the broad constellation of meanings and understandings associated with a given concept." (ibid, p. 531) Similarly to bias, many constructs are essentially contested [96]. Thus, most constructs have various perspectives on how the construct can be conceptualized. For instance, bias is defined as "systematic error arising during sampling, data collection, or data analysis" (APA dictionary "bias') or "prior information, a prerequisite for intelligent action." (Caliskan et al., 2017) [9]. Since definitions of bias vary significantly as the examples show, a broader concept provides a context where specific bias measurement is derived from. By connecting the background concept with the specific formulation of the definition used in the work, the context of the current work can be identified, among potentially diverse aspects of the construct. Background concepts highlight the choices of researchers made for the specific bias measurement.

- Systematized Concept

  The systematized concept is how scholars define the construct specifically, which makes it the most important concept to validate. Adcock and Collier define systematized concepts as "a specific formulation of a concept used by a given scholar or group of scholars; commonly involves an explicit definition." (Adcock and Collier, 2001, p. 531)[47] The authors stress the necessity of clear systematized concepts when the background concept is complex and contested. Without a clear systematized concept, it is easy to confuse the conceptual dispute from the operational dispute. The systematized concept connects current work to established theory related to the concept and previous works built upon the concepts. It provides theoretical groundings in conceptualizing and systematizing the construct. Based on the systematized concept, appropriate data and metrics that best capture bias in language models should be chosen[4].

  Connecting measurements to background and systematized concept helps to establish valid connection between the construct and the measurement. For instance, Jentzsch and Turan (2022)[59] define gender bias as "the difference in sentiment valuation of female and male sample versions." (p. 184) To operationalize gender bias as defined, the authors chose sentiment score to measure gender bias. However, the theoretical connection that justifies measuring gender bias with sentiment valuation is lacking. In other words, the authors only provide an indicator without

[4]It is another open question what would be the most appropriate to provide the background concept. Ideally, the conceptual work that shows a constellation of possible definitions would be ideal. For instance, Moss-Racusin et al. (2012)[84] is often referred to as a reference for gender bias in NLP papers. Moss-Racusin et al. (2012), however, is not a theoretical work on gender bias, but an empirical work that measures academic science's gender disparity. It is a seminal work that empirically measures gender bias, but theoretical works focusing on the definition, history, or characteristics of gender bias could be more appropriate to provide a background concept of the gender bias as the construct.

a systematized concept or background concept. Therefore, the interpretation of the result can only be strictly limited to discrepancies in sentiment scores. However, the implication of such measurements can be confusing since it contradicts the authors' motivations concerning social harms and misbehavior of the systems (ibid, p. 184-185). While it is also unclear what the authors indicate by social harms and misbehaviour, it is ambiguous how a discrepancy of sentiment scores can be connected to social harm and misbehaviour.

- Indicator

  The indicator refers to specific metrics, namely the formula suggested as a measurement in a specific setting. Adcock and Collier use the term indicator, measures, and operationalizations interchangeably. Indicators are developed from systematized concepts to score cases.(ibid, p. 531) Unless conceptualized with relevant background and systematized concepts, an indicator alone does not suffice to describe the phenomena that are being measured. It is also possible that background or systematized concepts are provided, yet do not match to the indicator practically used to measure bias. Providing mismatching prerequisite concepts is no better than providing no concepts, as inadequate backgrounds or systematized concepts might lead to a misinterpretation of the measures. An indicator designed ad hoc or solely based on the availability of data can also challenge the validity of the indicator.

5.3.2 Level 2: Measurement

Based on the conceptualization, specific metrics (or "Indicators'') can be developed to measure bias in language models. Metrics can be categorized as intrinsic or extrinsic, depending on what the measurements aim to measure. Intrinsic measurements aim to evaluate the internal bias of the language models in general. Tests like StereoSet [40] and CrowS-Pairs [55] have been designed to capture intrinsic bias in language models. Extrinsic measurements, on the other hand, focus on measuring bias in specific downstream tasks, such as coreference resolution, machine translation, or summarization[5].

Both intrinsic and extrinsic measurements need to meet certain quality criteria to be scientifically valid. The validity of measurement allows scores achieved by indicators to be interpreted as evidence of the construct being measured (Messick, 1978)[48]. The quality of measurement can be evaluated by applying an evaluative framework such as measurement modeling. Measurement modeling tests the aptness of the designed measurement for the construct by systematically disentangling assumptions made through the process of measuring unobservable constructs. There have been several attempts to introduce measurement modeling for developing more rigorous bias metrics in NLP, such

[5]There are mixed results regarding how bias in different intrinsic and extrinsic metrics relate to each other. While many assume that bias in intrinsic measurements that would propagate in extrinsic tasks, Cao et al., (2022)[122] showed that they do not necessarily correlate.

as in Jacobs and Wallach (2021)[69], Blodgett et al., (2022)[42], Bommasani and Liang (2022)[70], and Grimmer et al., (2022)[71].

In Chapter 3, I applied measurement modeling to the reviewed papers on gender bias in language models, by evaluating construct reliability and construct validity to suggested methods of measuring gender bias in language models. Every bias measurement should be validated on whether the metric measures what it aims to measure. Also, the reliability of the test is a fundamental foundation of scientific investigation. However, reviewing suggested methods for measuring gender bias show that metrics are not scrutinized in terms of their construct reliability and construct validity. Since I already discussed measurement modeling in Chapter 3, I will not repeat the same argument here that the quality of operationalization should be evaluated to ensure scientific validity.

In this section, I would like to highlight the necessity of referencing relevant fields outside NLP. To evaluate content validity, I determined whether the measurements rely on established theory to conceptualize bias that is being measured. I argue that operationalizing bias should reference research fields where bias has been widely studied.

Operationalizing construct and measuring it in computer systems is a relatively new research field compared to measuring society and humans. Social science has developed methodologies for a valid investigation of complex societies over centuries. The methods have been developed to best capture how human society works. Measuring social constructs in computer systems share similar difficulties that social science faces measuring society, but computer systems pose new challenges as well. Computational social science, for instance, uses computational methods to study society ([123][75][124], inter alia). Studying society with digital data requires a new approach compared to traditional types of data, such as survey data. Furthermore, studying value-laden concepts such as fairness, justice, or bias in technological systems, such as language models, require a new approach as well. Simply duplicating methods developed for human society and applying it to computer systems would not be the solution. Instead, identifying shared challenges, such as conceptualizing abstract constructs and operationalizing them, would benefit studying values in computer systems and will open up new possibilities. It seems unreasonable not to reference well-established approaches in relevant disciplines to conceptualize constructs such as gender bias, to measure gender bias in language models. Despite the differences between language models and society, the construct of gender bias applied to both are no different. Moreover, language models are part of society where such traditional theory is based on. Referencing how other disciplines theorize and conceptualize an abstract concept that is also of an interest to language models can only be of benefit.

Properties to Measure Gender Bias in LMs

Systematic literature review showed that most gender bias measurements used stereotypes and sentiment scores. The predominance of a few methods runs the risk of reducing

the complex nature of gender bias into several simplified dimensions. Such trends might partially be explained by the success of initially proposed benchmarks, such as StereoSet [40], CrowS-Pairs [55], and WinoBias [85]. Later proposed methods were inspired by previously suggested methods, leading to the popularity of stereotype-based bias measurements. Sentiment scores are commonly used in various NLP tasks, partly due to the availability of sentiment data such as IMDB movie scores and online retail review data. Previously developed socio-linguistic tools such as LIWC [125][126], Perspective API [127][128], or VADER [64][45] also promotes the use of sentiment score to measure other constructs, such as bias. However, most measurements do not provide a connection that validates measured sentiments and the construct of bias.

The lack of theoretical grounding is striking especially in comparison to the measurements investigating bias in humans. Measuring bias in humans has developed a significantly more sophisticated and nuanced approach to investigate bias as a construct. Psychology, for instance, has investigated various ways to measure gender bias in humans. Various scales have been developed to quantify diverse aspects of gender bias.

Alternative Ways to Operationalize Gender Bias

A novel approach can be developed by applying existing methodologies in social science with computational method to study sociotechnical systems. For instance, Samory et al., (2021)[129] study sexism to detect hate speech on social media. To better operationalize the construct gender bias, the authors refer to psychological scales developed to measure sexism from humans. Numerous scales have been developed to quantify diverse aspects of gender bias, such as benevolent sexism [130] and affective attitudes toward the feminist movement [131]. Building upon the established literature of gender bias from psychology to develop sexism detection in social media significantly improves construct validity and reliability.

Samory et al., (2021)[129] refer to 29 psychological scales measuring sexism and related constructs(p. 576). The authors analyze the items of the scales and categorize the items into four categories: (1) Behavioural expectation, (2) Stereotypes and comparisons, (3) Endorsement of inequality, and (4) Denying inequality and rejection of feminism. The category based on sexism scales are more diverse than how gender bias is operationalized in bias in language models , which was mainly using stereotypes and sentiment scores.

While there cannot be a single work that is comprehensive of all the relevant of aspects of gender bias, working beyond the disciplinary boundary will provide a better-footed start. After conceptualizing the construct and devising methods on top of it, measurement modeling can be applied to evaluate construct validity and reliability, as shown in Chapter 3.

5.3.3 Level 3: Structural Analysis

The last level of the framework is a structural analysis. Structural analysis of bias is novel compared to other previously proposed holistic bias frameworks. As discussed in Chapter 4, bias that is morally wrongful discrimination and instantiates structural injustice deserves a higher degree of attention than others. Such bias poses a significant risk for the adoption of language models, since it risks exacerbating existing injustice with technological means. According to the meaning-based account of discrimination, wrongful discrimination fails to recognize equal worth between different people. Bias that wrongfully discriminates people challenges equality, which is a fundamental cornerstone of democratic society. Moreover, it is especially problematic when such discrimination is based on structural injustice where social process prevents certain groups of people from realizing their capabilities and deprives them of their opportunities [103]. Therefore, identifying such bias is a necessary condition to use language models in society, without exacerbating existing injustice through language models.

The relevance of bias can be determined by examining the meaning of bias, as the meaning matters more than the difference itself. The meaning can only be understood in the particular context where language models are situated. To evaluate structural dimension of bias, therefore, it is critical to consider a larger context beyond specific language models. Individual language models and applications do not exist in isolation from the context where the language models are developed and deployed. Moral and social evaluation is only possible by taking into account the context where language models are in use and exchange interaction with people.

For the structural analysis of bias in language models, I suggest three different perspectives specifically: stakeholders, technology, and social context. Stakeholders refer to agents who are involved in the development process, such as developers, annotators, and investors to the company developing language models, and users of the system, whose bias can be reflected in language models. Moreover, some stakeholders might introduce bias to language models without being the users of language models. The technology perspective focuses on the technology itself, such as the energy resources required to develop language models. Lastly, social context is meant to be inclusive of other relevant structural dimensions, including economic, political, or cultural context where developing language models are incentivized in a social process.

Stakeholders

Developers of the language models are not limited to software programmers or researchers who directly contribute to developing algorithms. The companies and those who work in the companies are likely to have the significant decision-making power concerning the language models they are building and their end results. However, developers of language models encompass larger groups of people, such as moderators for the training data and output of 'uncensored' outputs of the language models. Besides, people who

are indirectly involved in the development of language models, such as investors, who have interest in language models, can also be included in the developers. Boyarskaya et al. (2020)[121] have argued that evaluating harms of computational systems should consider a wide range of potential stakeholders. The authors criticized that most existing scholarship on computational harms do not capture the "complex nature of many AI systems that aim to meet user needs (...) nor does it address the variety of stakeholders and the many ways they interact with these systems" (ibid, p.2). Instead of addressing the variety of stakeholders, the focus for identifying bias in language models has been limited to the outputs of language models in certain test settings. However, similarly to addressing potential harms of language models, bias in language models is also under the influence of various stakeholders, including those who do not directly participate in developing the language models. For instance, content creators of online communities will contribute to the bias of language models without interacting with the model itself, by providing training data to the model. The bias of the content can be reflected to the language model in a way that the content creator is not aware of. The bias of decision-makers who choose which data will go into the training or test set also contribute to the bias of language models. The objective of language models, such as achieving certain level of performance in certain tasks or applications, will affect how language models will be optimized and thus introduce certain biases. Both direct and indirect interaction between language models and various stakeholders help delineate sources of bias.

Table 5.2 shows the language models that were tested in the 19 papers I reviewed for gender bias measurement. In total, 16 models were used to measure gender bias in them. Most of authors' affiliations were companies based in the U.S., such as Google (5), Microsoft (3), Facebook AI (2), and Open AI (2). While my review is limited to the papers measuring gender bias using novel methods, a similar trend is found in a broader scope of review. According to the HAI AI Index(2023), 54% of authors of select large language models are from institutions in the United States, 22% from the United Kingdom, and 6% from Canada, which in total comprises over 80% as of 2022 [120] (p. 58). In other words, most language models are developed in English-speaking countries of the Global North. The concentration of power to a handful of organizations can introduce potential bias that reproduces hegemonic views [15]. Besides the companies who are developing the language models, it is worth noting to identify who has an interest in developing, deploying, and monetizing LLMs. Bias and associated social and moral harms can be critically assessed by taking into account of interested parties. Similarly to the health impact studies of nicotine on the human body published by cigarette companies, potential social harms and risks researched by the same party who benefits from the economic success of language models should be interpreted with caution.

Model	Organization	Organization Type	Paper Title	Number of papers Evaluating the Model
GPT-2	OpenAI	Company	Release Strategies and the Social Impacts of Language Models	8
GPT-3	OpenAI	Company	GPT-3: Language Models are Few-Shot Learners	1
DialoGPT	Microsoft	Company	DialoGPT: Toward Human-Quality Conversational Response Generation via Large-Scale Pretraining	1
BERT	Google	Company	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	11
ALBERT	Google	Company	ALBERT: A Lite BERT for Self-supervised Learning of Language Representations	5
	Toyota Technological Institute at Chicago	Company		
RoBERTa	Paul G. Allen School of Computer Science & Engineering	University	RoBERTa: A Robustly Optimized BERT Pretraining Approach	7
	Facebook AI	Company		
DistilBERT	Hugging Face	Company	DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter	3
mBERT	Google	Company	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	2
NorBERT	University of Oslo	University	Large-Scale Contextualised Language Modelling for Norwegian	1
NB-BERT	The National Library of Norway	Public institution	Operationalizing a National Digital Library: The Case for a Norwegian Transformer Mode	1
XLNet	Carnegie Mellon University	University	XLNet: Generalized Autoregressive Pretraining for Language Understanding	3
	Google	Company		
MPNet	Nanjing University of Science and Technology	University	MPNet: Masked and Permuted Pre-training for Language Understanding	1
	Microsoft	Company		

Model	Organization	Organization Type	Paper Title	Number of papers Evaluating the Model
xlmR	Facebook AI	Company	Unsupervised Cross-lingual Representation Learning at Scale	1
T5	Google	Company	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer	1
COMET	Allen Institute for Artificial Intelligence	University	COMET : Commonsense Transformers for Automatic Knowledge Graph Construction	1
	Paul G. Allen School of Computer Science & Engineering	University		
	Microsoft	Company		
CTRL	Salesforce Research	Company	CTRL: A CONDITIONAL TRANSFORMER LANGUAGE MODEL FOR CONTROLLABLE GENERATION	1

Table 5.2: Language Models Used To Measure Gender Bias

Google = Google, Google AI Language, Google Research, Google AI Brain Team
Microsoft = Microsoft, Microsoft Research

Table 5.2 shows the language models that were tested in the 19 papers I reviewed for gender bias measurement. In total, 16 models were used to measure gender bias in them. Most of authors' affiliations were companies based in the U.S., such as Google (5), Microsoft (3), Facebook AI (2), and Open AI (2). While my review is limited to the papers measuring gender bias using novel methods, a similar trend is found in a broader scope of review. According to the HAI AI Index(2023), 54% of authors of select large language models are from institutions in the United States, 22% from the United Kingdom, and 6% from Canada, which in total comprises over 80% as of 2022 [120] (p. 58). In other words, most language models are developed in English-speaking countries of the Global North. The concentration of power to a handful of organizations can introduce potential bias that reproduces hegemonic views [15]. Besides the companies who are developing the language models, it is worth noting to identify who has an interest in developing, deploying, and monetizing LLMs. Bias and associated social and moral harms can be critically assessed by taking into account of interested parties. Similarly to the health impact studies of nicotine on the human body published by cigarette companies, potential social harms and risks researched by the same party who benefits from the economic success of language models should be interpreted with caution.

Users are another important group of stakeholders. Depending on the potential target population of the language model, the social value alignment to language models will vary. The interaction between users and language models can steer what is desirable for language models. ChatGPT [5], for instance, utilizes users' feedback on the ChatGPT's response by using an upvote and downvote function. Depending on the mechanism of how this feedback is reflected in the language model, user interaction can directly influence bias in language models. Demographics of the users, for instance, can influence how certain stereotypes are strengthened in language models. Based on how ChatGPT functions, the main targets of the language models are English speaking populations with access to the internet. While the potential users of language models are not a representative sample of society, the bias of users will be overrepresented in the language models through the interaction. For instance, so-called "bro culture" in Silicon Valley has been criticized for making the industry unfavorable to women [132][133]. It reflects another dimension of power that certain patterns of bias will be more visible than others to language models.

Lastly, stakeholders include those who do not directly interact with language models. Since language models are trained with the vast amount of data scraped from the internet, those who contribute to the training data will indirectly influence bias in language models. It is well-established that online platforms such as Wikipedia and Reddit under-represent women ([134][135][136], inter alia). It is in line with the Common Crawl Corpus analysis by Luccioni and Viviano (2021)[137], which showed that extensively used training data includes a significant amount of undesirable content, such as sexual content, hate speech, and racial and gender biases. It shows that the internet is a biased

representation of society, which also overrepresents younger, English-speaking individuals from developed countries (ibid, p. 186). Consequently, bias embedded in society and represented in the internet will be reflected in the language models.

Technology

Besides stakeholders, the technology itself is another important dimension to conduct a structural analysis. Instead of focusing on specific features of certain language models, this dimension focuses on common features that language models share.

Language models are black-box algorithms with internal mechanism that are opaque even to the developers who design the algorithm. The opaque nature makes it difficult to scrutinize language models in detail to identify how bias is reflected in the algorithm. The opacity of algorithms raise epistemic concerns since only inconclusive evidence is available to scrutinize the algorithm [138]. Given that the transparency of an algorithm is the first step to evaluate other concerns regarding societal implication, such as accountability and responsibility, the opaque nature of language models poses a significant challenge to identify bias.

Furthermore, the basic mechanism of language models often prevents from democratic development of the technology. Language models have notoriously high financial requirements for the computation involved in training. Due to the cost of training, most researchers can only access the language models with inference. Vassimon Manela et al. (2021)[53] pointed out that "this training is only performed once, with users downloading and fine-tuning such language models to their specific task. In doing so, we are trusting large tech companies to train the base model responsibly since we have no control over this. This seems inherently undemocratic." (p. 2232) Training large language models cost are growing fast as shown in the estimated cost for training PaLM, which is a language model released in 2022, to be $8 million USD [120](p. 23). Rapidly increasing cost of training is limiting the accessibility of developing language models for organizations without such resources.

Due to the financial accessibility, most language models are developed by big companies or others who are dependent on big companies. Such dependency applies regardless of the specific types of language models, thus deserves an investigation as a potential source of bias. When most language models are developed by private enterprises, the motivations are not free from financial incentives. Financial incentives can underestimate the long-term societal implication when it collides with the business objectives. Moreover, tech companies suffer from an under-representation of women, one article estimated only 12% of machine learning researchers are women [139]. Diversity reports from Google [140] and Microsoft [141] show that women account for about 30% of their employees. It indicates that major organizations developing language models underrepresent women. The marginalization will affect which patterns in language models

will perceived as undesirable and biased[6].

Social Context

How language models as a collective are situated in a wider social context is highly relevant to the social and moral implications of bias in language models. Beyond specific language models or stakeholders, the structural influence in which language models are developed, merits consideration to assess bias in language models.

Despite differences in size, architecture, and capability, large language models as a category exists in a contemporary context where such technology is actively developed and widely used. Start-ups using AI attracts economic [144][145] and political attention [146], driving public discourse around policies concerning technology, economy, and education across borders.

The optimistic prospects of language models corresponds to the technological solutionism, which views that technology can solve social problems [147]. Silicon valley, where many language models are developed and used, has a highly competitive culture where technological perspective outruns other perspectives could be responsible for the benchmark-ism in measuring bias. Bias is approached as another performance metric that has to be reduced and optimized, especially compared to others. It disregards the fundamental difference between bias and other performance metrics, such as accuracy in specific tasks, that bias can only be evaluated in particular social contexts. Similarly to the technological solutionism, academic work that is evaluated by the number of citations might also motivate researchers to develop more datasets and benchmarks, rather than conducting a qualitative research. Papers releasing datasets and benchmarks are more prone to be cited as they can be easily compared between models and produce statistical reports.

Monopolistic competition of the tech industry by a handful of hegemonic private enterprises also have a high impact in the development of language models. It is not a coincidence that many companies who drive the development of language models are the biggest online platforms and are among the most resource rich companies globally. These companies have access to a larger population of users and resources than even some national governments. Such a concentration of power determines the social process of defining the utility, purpose, and relevance of the language models. The discourse [148] about language models is not only driven by the same few companies who develop these models, but also the platform where such discourse is happening is often owned by the

[6]It is worth noting that most well-known figures who critique LLM are women, while most proponents are men — this gender imbalance itself might introduce bias in the debate. For instance, the AI Moratorium open letter [142], which started a heated discussion about whether we should continue or pause the development. The open letter has been criticized for ignoring existing critiques paid to develop language models by framing the development of AI as something that "happened" beyond the capabilities of developers. The high-profile figures who signed the letter (and who were covered in media) includes Elon Musk and Yoshua Bengio, while the opponents of the letter, such as an open letter published from DAIR [143], was authored exclusively by women.

same actors [149] who develop language models.

The combination of technosolutionism and the hegemonic status of tech companies that develop language models pose an additional challenge to assess significant bias that is morally wrong and reproduces structural injustice. The difficulty of conducting research concerning moral and social implications is shown in recent lay offs of ethics teams in tech companies like Google [150] and Microsoft [151]. Investigating long-term social and ethical implications of technology can conflict with fast-paced market competition, in which private enterprises often prioritize the latter. However, it is difficult for other entities, such as academic institutes or public organizations, to conduct a language model research on a par with private companies, due to the high cost of the research. Against the zeitgeist of technological solutionism that is favored by capitalistic market economy, it is challenging to integrate complex and time-consuming perspectives to study society, as in humanities and social science. While assessing bias in language models is crucial for evaluating societal impact of the technology, measuring unobservable constructs such as equality, fairness, and justice, inevitably takes additional effort that is often compromised compared to immediate economic incentives.

5.3.4 Measuring Gender Bias in an Imaginary App "CookGPT"

In the last section, I explained how bias can be evaluated holistically beyond a technical account of bias. I will show how to use the framework by using an imaginary scenario with the app named CookGPT. Imagine that you are the CEO of an AI start-up company "CookAI', developing the CookGPT app. CookGPT is an app built using the large language model GPT-X, which recommends recipes. You can ask CookGPT "What should I have for lunch? I want something with potatoes," and CookGPT will recommend you several recipes. The recipes include information like ingredients, cooking steps, cooking time, the developer of the recipe, and the URL where the recipe could be found.

- Level 1: Conceptualization

  While the app has been a huge success, a few employees voiced their concern that most of the recommended recipes are from male chefs. As a responsible CEO, you suggested that we need to identify gender bias in the app CookGPT, and measure it to evaluate the situation. The company formed the Bias Task Force (BTF) with a mission of identifying, measuring, and ultimately reducing gender bias in CookGPT.

  The initial observation that contributed to the birth of the task force was a biased representation of male and female chefs in recommended recipes. After a careful review of existing methods to measure bias in language models, one suggested that the company should define that CookGPT is biased if it recommends more recipes written by men than women. The definition corresponded with most of the bias metrics out there, and reflected the initial observation of gender bias in CookGPT.

But the first level of the TCB framework introduced more confusion to the task force.

What is the background concept of gender bias that we, the company CookAI and the Bias Task Force, are concerned with? There were varying definitions of gender bias suggested by different literature. It also leads to different systematizations of the construct of gender bias. The closest definition from CookGPT's definition would be defining gender bias as a discrepancy between male and female representation. CookGPT would be unbiased when it recommends recipes from male and female writers equally.

However, some might argue that such a definition is unrealistic since there are more male chefs out there in society. CookGPT should be called biased only when its output is "more biased" than the real occupational distribution of the chefs. They argue that one should not blame CookGPT for reflecting reality. Since there are more male chefs in general, it is natural for CookGPT to recommend more male-written recipes.[7]

It was followed by another discussion to choose appropriate statistics to compare CookGPT. Should it be the latest labor statistics from the government to reference registered chefs' gender composition, or should it be statistics from the authors of online recipes? Should we (as a company, CookAI), use U.S. statistics or global statistics? Should it be a weighted average based on the country of users of CookGPT? How can we justify such a choice of statistics concerning the gender bias of CookGPT? While the company considered this approach easier than coming up with an ideal version of CookGPT since the developers could not agree on the ideal, but referencing statistics also came with challenges. The company eventually settled on the U.S.-based statistics from the previous year, as the majority of users were from the U.S. However, the results might discriminate against non U.S.-based users as their values are not reflected in how bias is defined in CookGPT.

- Level 2: Measurement

  The Bias Task Force found a paper showing that text generated by GPT-X tends to correlate negative sentiment to female names. It might explain why female-written recipes are recommended less. But they also found that GPT-X produces more toxic text with male pronouns like he, him, or his. It is unclear how these two papers relate to each other, one uses names and the other uses gendered pronouns. One uses a sentiment score based on the movie review, while the other uses a toxicity score including sentiment scores inspired by linguistics. Even though both papers study the gender bias of GPT-X, it is tricky to interpret their result in

[7] This perspective is one of the examples of how bias of language models are defined as in Dilemmaland, as described in Chapter 3.

relation to CookGPT's gender bias problem. There is a study that used male and female actors' Wikipedia articles to study gender bias. But it was tested on a different model than GPT-X, using different professions than cooks. It is unclear whether the result would be reproducible in GPT-X using different occupations. In addition, the Task Force found extrinsic measurements for analyzing the generated text of GPT-3. It showed that the model correlates female names with traditional gender roles. But it didn't solve the problem of CookGPT, as cooking could have been considered a traditionally female domain, CookGPT could prefer female cooks based on behavioral expectations. But that is not consistent with CookGPT's preference towards male chefs.

Despite abundant work on gender bias in language models, it was difficult to determine which research papers are relevant to the case of CookGPT. For systematic evaluation of various measurements, the Bias Task Force decided to apply measurement modeling to those metrics. The Task Force found out that most of the existing bias metrics do not test reliability and validity of operationalizing gender bias. Most metrics did not compare to each other, which makes it difficult to have an overview. Different metrics were tested in different language models, using heterogeneous properties such as the sentiment score of the text to the frequency of gendered nouns in the fill-in-the-blank type of tasks. The choice of properties was not justified by relevant theory or rigorous validation.

So the task force decided to come up with their own measurement by following the suggestion from the framework. The task force looked into other disciplines of literature, from social psychology and linguistics, to reference how gender bias can be measured in language. The theory provided them with a background concept, which is based on the brilliance bias [152] which imposes brilliance as a male trait and disregards women's achievements. And based on the background concept, the task force came up with a systematized concept that is relevant to the purpose of evaluating gender bias in CookGPT, which is an underevaluation of female-written recipes while the content is on par with male-written recipe. The metric, indicator, was designed to reflect the systematized concept to measure the similarity between two recipes and compare the score of CookGPT based on the differing gender.

- Level 3: Structural Analysis

  For the structural analysis, the Bias Task Force conducted a three-level investigation on stakeholders, technology, and social context, to identify potential source of gender bias that is reflected in CookGPT.

  Stakeholders

  To investigate stakeholders, the Bias Task Force explored four different groups: annotators, investors, users, and developers. CookAI hired a team of annotators to evaluate the output of CookGPT. The crowdsourced workers were hired from

a country where the minimum wage is significantly lower than in the U.S., where the company CookGPT is based on. Because of social taboos, most women do not participate in social activity in that region. So most of the people working on these outsourced tasks were men.

The main investor of the app, the venture capitalist, is composed of mostly male investors. Most venture capitalists have long working hours and as such does not appeal to everyone. There are fewer women working in this industry. CookAI did not attract as much investment, possibly because the CEO is a woman. It is known that female-led companies receive less investment due to the gender bias in Silicon Valley [153]. Due to the lack of financial resources, developers in CookGPT could not scrutinize other potential sources of bias in GPT-X, such as annotator bias or sampling bias, since they are dependent on the company training GPT-X.

Women generally tend to participate less in online discourse. Beta testers for CookGPT were recruited via a popular social media platform. The most active topics of social media are online games and sports. Social media was chosen due to its size and number of active users, not due to its topical focus on online games. However, regardless, the choice of social media introduces a bias against women by over-representing male users who are interested in online games.

The nature of the app itself favors male users. According to studies, men exhibit a more friendly attitude to new technologies. The fact that CookGPT is a recipe recommendation app using "AI" might attract more male users than female users.

While the composition of developers in the company was diverse, female employees were mostly working on PR and marketing, while male employees were the developers of the app. Less female candidates apply for the job, due to the long social stigma, and educational system against women in STEM. While diversity in the company is a significant step toward preventing bias, it is not sufficient.

Technology

You found that there is an unexpected correlation between the name of chefs to the recipe and CookGPT's recommendations. The recipes containing larger image sizes were prone to receive better scores. It results in favoring people who invest more in their camera equipment. Unrelated to the design choice made by the company, it resulted in favoring male recipe writers over female ones since male chefs tend to have a more expensive camera gear.

It is impossible to uncover all potential proxy problems that exist in CookGPT, due to the opaque nature of large language models. For instance, pictures included in the recipe that CookGPT recommends can be a confounding feature that determines the attractiveness of the recipe. Let's imagine that CookGPT prefers recipes including pictures taken from the professional restaurant kitchen, rather

than a home kitchen. Statistically, there are more visible male chefs and more female home cooks.

These two examples, despite their similarity, reflect the historical undervaluation of female work. Despite the fact that they both engage in similar work, one is acknowledged as a professional occupation, while the other is considered less sophisticated. It shows the complicated nature of bias that cannot be simply reduced into one or two features. Embedded social injustice is reflected in the data used for making the language model "intelligent", and it is so prevalent and intrinsic that it is difficult to identify and isolate it, to "debias" the language models.

Social Context

Due to the limited resource that the company had, the Bias Task Force consisted of only two researchers. Compared to the development team that had dozens of researchers, the Bias Task Force could not progress as fast as other teams, resulting in a worse evaluation at the end of the year by the corporate HR team. In the end, the Bias Task Force had to terminate before identifying the source of gender bias in CookGPT and devising a mitigation strategy.

Chapter 6

Conclusion

The central question of the thesis has been focused on how bias in language models is defined and measured. Conceptually, bias, alongside fairness, has been the placeholder for anything undesirable that language models produce. Conflated use of bias led to operationalizations that lack scientific validity. In practice, many undesirable patterns have been identified under the name of bias. Therefore, investigating how bias in language models is defined, measured, and reduced, concerns evaluating language models in terms of value.

Despite the heterogeneity of measured bias, I argued that bias that is a case of wrongful discrimination, especially those that reproduces existing injustices like historical injustice and structural injustice, should be prioritized. Prioritizing descriptive accuracy risks allowing statistical discrimination of language models, which will reproduce structural injustice via language models. It will make language models like a "rational racist" whose racist attitudes are epistemically justified but morally unjustifiable [116]. Therefore, normative correctness should be considered to reduce bias in language models.

I presented a framework that combines empirical and normative critiques on bias measurement to evaluate bias holistically. The framework consists of three dimensions, i.e., conceptual, empirical, and structural analyses, to expand bias beyond the technical terms. Structural dimensions of bias ought to be integrated in measuring bias since there are no non-social language models that free from societal implication. Language models are always situated in society through the training data, the environment that enables development of language models, and interaction with users. Therefore, evaluating bias in language models also should be situated in the social context.

Social and moral requirements should extend the quality criteria to develop and improve language models. Using language models that are technically good but discriminatory against marginalized groups will result in sacrificing equality to the usefulness of the tool. Language models risk scaling existing injustices due to the development procedure and the versatility on how the technology is created. It is critical to measure bias scientifically and meaningfully to assess the value of language models accurately.

Bibliography

[1] Corin Faife. Debt collectors want to use ai chatbots to hustle people for money. https://www.vice.com/en/article/bvjmm5/debt-collectors-want-to-use-ai-chatbots-to-hustle-people-for-money, 18-05-2023. (accessed: 24. 5. 2023).

[2] Aaron Mok. How a coder used chatgpt to find an apartment in berlin in 2 weeks after struggling for months. https://www.businessinsider.com/how-coder-used-chatgpt-to-find-apartment-in-berlin-2023-5?op=1, 2023-05-23. (accessed: 24. 5. 2023).

[3] Andrew R. Chow and Billy Perrigo. The ai arms race is changing everything. 17 Feb 2023. (accessed: 24. 5. 2023).

[4] C. Meister and R. Cotterell. Language model evaluation beyond perplexity. *arXiv preprint*, 2021.

[5] Chatgpt. https://chat.openai.com/.

[6] Steven T. Piantadosi. Python code created by chatgpt. https://twitter.com/spiantado/status/1599462375887114240?s=20. (accessed: 24. 5. 2023).

[7] Michael Gusenbauer and Neal R. Haddaway. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. 11(2):181–217, 2020-03.

[8] Timo Spinde, Jan-David Krieger, Terry Ruas, Jelena Mitrović, Franz Götz-Hahn, Akiko Aizawa, and Bela Gipp. Exploiting transformer-based multitask learning for the detection of media bias in news articles. In *Information for a Better World: Shaping the Global Future: 17th International Conference, iConference 2022, Virtual Event, February 28–March 4, 2022, Proceedings, Part I*, pages 225–235. Springer, 2022.

[9] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. 356(6334):183–186, 2017-04-14.

[10] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. page 9.

[11] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. Large Pre-trained Language Models Contain Human-like Biases of What is Right and Wrong to Do. 2022-02-14.

[12] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from Language Models. page 64.

[13] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[14] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[15] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. ACM, 2021-03-03.

[16] Langdon Winner. Do Artifacts Have Politics? 109:121–136, 1980.

[17] Batya Friedman and Colby College. Bias in computer systems. 14(3):18.

[18] Value-sensitive design. 1996.

[19] H. Nissenbaum. How computer systems embody values. 34(3):120–119, 2001-03.

[20] Tomo Lazovich. Does Deep Learning Have Politics? page 4.

[21] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640. Association for Computational Linguistics, 2019.

[22] Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. Measuring Fairness with Biased Rulers: A Comparative Study on Bias Metrics for Pre-trained Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706. Association for Computational Linguistics, 2022.

[23] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. 2020-05-29.

[24] Karolina Stanczak and Isabelle Augenstein. A Survey on Gender Bias in Natural Language Processing. 2021-12-28.

[25] Gabbrielle M. Johnson. Algorithmic bias: On the implicit biases of social technology. 198(10):9941–9961, 2021-10.

[26] Mary Flanagan, Daniel C. Howe, and Helen Nissenbaum. Values at play: Design tradeoffs in socially-oriented game design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 751–760. ACM, 2005-04-02.

[27] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[28] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[29] American Psychological Association. Gender bias. https://dictionary.apa.org/gender-bias. (accessed: 25.05.2023).

[30] Amy Allen. Feminist Perspectives on Power. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022.

[31] Frederick J Gravetter and Lori-Ann B Forzano. *Research methods for the behavioral sciences*. Cengage learning, 2018.

[32] Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. 3:1–36, 2019-11-07.

[33] Christian Heumann, Michael Schomaker, and Shalabh. *Introduction to Statistics and Data Analysis*. Springer International Publishing, 2016.

[34] American Psychological Association. Bias. https://dictionary.apa.org/bias. (accessed: 25.05.2023).

[35] Thomas M. Holtgraves and Thomas M. Holtgraves. Language and Social Psychology. In Thomas M. Holtgraves, editor, *The Oxford Handbook of Language and Social Psychology*. Oxford University Press, 2014-09-01.

[36] Cambridge Dictionary. Bias. https://dictionary.cambridge.org/dictionary/english/bias. (accessed: 25. 05. 2023).

[37] Reuben Binns. Fairness in Machine Learning: Lessons from Political Philosophy. page 11, 2018.

[38] Oisín Deery and Katherine Bailey. The Bias Dilemma: The Ethics of Algorithmic Bias in Natural-Language Processing. 8, 2022.

[39] Rima Basu. The Specter of Normative Conflict. In Erin Beeghly and Alex Madva, editors, *An Introduction to Implicit Bias*, pages 191–210. Routledge, 1 edition, 2020-03-27.

[40] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. 2020-04-20.

[41] Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. Towards a Comprehensive Understanding and Accurate Evaluation of Societal Biases in Pre-Trained Transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389. Association for Computational Linguistics, 2021.

[42] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. page 12.

[43] Samia Touileb, Lilja Øvrelid, and Erik Velldal. Occupational Biases in Norwegian and Multilingual Language Models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 200–211. Association for Computational Linguistics, 2022.

[44] Marion Bartl, Malvina Nissim, and Albert Gatt. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. *arXiv preprint arXiv:2010.14534*, 2020.

[45] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872. ACM, 2021-03-03.

[46] Anna Sotnikova, Yang Trista Cao, Hal Daumé III, and Rachel Rudinger. Analyzing Stereotypes in Generative Text Inference Tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4052–4065. Association for Computational Linguistics, 2021.

[47] Robert Adcock and David Collier. Measurement validity: A shared standard for qualitative and quantitative research. *American political science review*, 95(3):529–546, 2001.

[48] Samuel Messick. Validity. *ETS research report series*, 1987(2):i–208, 1987.

[49] David J Hand. *Measurement: A very short introduction*. Oxford University Press, 2016.

[50] Stephen G Sireci. The Construct of Content Validity. 2023.

[51] Hannah Rose Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frederic A Dreyer, Aleksandar Shtedritski, and Yuki M Asano. Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models.

[52] Sarah Alnegheimish, Alicia Guo, and Yi Sun. Using natural sentence prompts for understanding biases in language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2830, 2022.

[53] Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242, 2021.

[54] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North*, pages 622–628. Association for Computational Linguistics, 2019.

[55] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. 2020-09-30.

[56] Bum Chul Kwon and Nandana Mihindukulasooriya. An Empirical Study on Pseudo-log-likelihood Bias Measures for Masked Language Models Using Paraphrased Sentences. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 74–79. Association for Computational Linguistics, 2022.

[57] Victor Steinborn, Philipp Dufter, Haris Jabbar, and Hinrich Schuetze. An Information-Theoretic Approach and Dataset for Probing Gender Stereotypes in Multilingual Masked Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 921–932. Association for Computational Linguistics, 2022.

[58] Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955. Association for Computational Linguistics, 2021.

[59] Sophie Jentzsch and Cigdem Turan. Gender Bias in BERT - Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 184–199. Association for Computational Linguistics, 2022.

[60] Robert Wolfe and Aylin Caliskan. Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models. 2021-10-01.

[61] Li Lucy and David Bamman. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55. Association for Computational Linguistics, 2021.

[62] Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. Gender Bias in Masked Language Models for Multiple Languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750. Association for Computational Linguistics, 2022.

[63] Tianshu Shen, Jiaru Li, Mohamed Reda Bouadjenek, Zheda Mai, and Scott Sanner. Towards understanding and mitigating unintended biases in language model-driven conversational recommendation. 60(1):103139, 2023-01.

[64] C. Hutto and Eric Gilbert. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. 8(1):216–225, 2014-05-16.

[65] Jigsaw. https://jigsaw.google.com/the-current/toxicity/. (accessed: 24. 5. 2023).

[66] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3405–3410. Association for Computational Linguistics, 2019.

[67] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.

[68] Svetlana Kiritchenko and Saif M Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018.

[69] Abigail Z. Jacobs and Hanna Wallach. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385. ACM, 2021-03-03.

[70] Rishi Bommasani and Percy Liang. Trustworthy Social Bias Measurement. 2022.

[71] Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press, 2022-01-04.

[72] Edward G Carmines and Richard A Zeller. *Reliability and validity assessment*. Sage publications, 1979.

[73] SA Moser and KA Kalton. Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. *How to test the validation of a questionnaire/survey in a research (August 10, 2016)*, 1989.

[74] Baruch Nevo. FACE VALIDITY REVISITED. 22(4):287–293, 1985-12.

[75] Claudia Wagner, Markus Strohmaier, Alexandra Olteanu, Emre Kıcıman, Noshir Contractor, and Tina Eliassi-Rad. Measuring algorithmically infused societies. *Nature*, 595(7866):197–204, 2021.

[76] Patricia Hill Collins. *Black sexual politics: African Americans, gender, and the new racism*. Routledge, 2004.

[77] D Soyini Madison. Crazy patriotism and angry (post) black women. *Communication and Critical/Cultural Studies*, 6(3):321–326, 2009.

[78] Kaye Mitchell. Raunch versus prude: Contemporary sex blogs and erotic memoirs by women. *Psychology & Sexuality*, 3(1):12–25, 2012.

[79] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.

[80] Brent L Hughes, Nicholas P Camp, Jesse Gomez, Vaidehi S Natu, Kalanit Grill-Spector, and Jennifer L Eberhardt. Neural adaptation to faces reveals racial outgroup homogeneity effects in early perception. *Proceedings of the National Academy of Sciences*, 116(29):14532–14537, 2019.

[81] Brendan T Johns and Melody Dye. Gender bias at scale: Evidence from the usage of personal names. *Behavior research methods*, 51:1601–1618, 2019.

[82] Eve Kraicer and Andrew Piper. Social characters: the hierarchy of gender in contemporary english-language fiction. *Journal of Cultural Analytics*, 3(2), 2019.

[83] Stacy L Smith, Marc Choueiti, Ashley Prescott, and Katherine Pieper. Gender roles & occupations: A look at character attributes and job-related aspirations in film and television. *Geena Davis Institute on Gender in Media*, pages 1–46, 2012.

[84] Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences*, 109(41):16474–16479, 2012.

[85] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.

[86] Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

[87] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020.

[88] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172. Association for Computational Linguistics, 2019.

[89] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in Criminal Justice Risk Assessments: The State of the Art. 2017-05-27.

[90] Alan Lundgard. Measuring justice in machine learning. *arXiv preprint arXiv:2009.10050*, 2020.

[91] Kate Crawford. The trouble with bias. keynote at neurips. 2017.

[92] Wikipedia. Proud boys. (accessed: 30. 5. 2023).

[93] Jason Wilson. Proud boys are a dangerous 'white supremacist' group say us agencies. (accessed: 30. 5. 2023).

[94] Deborah A Trytten, Anna Wong Lowe, and Susan E Walden. "asians are good at math. what an awful stereotype" the model minority stereotype's impact on asian american engineering students. *Journal of Engineering Education*, 101(3):439–468, 2012.

[95] Elizabeth S Anderson. What is the point of equality? *Ethics*, 109(2):287–337, 1999.

[96] W. B. Gallie. IX.—Essentially Contested Concepts. 56(1):167–198, 1956-06-01.

[97] Elettra Bietti. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 210–219, 2020.

[98] Ben Green. The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics in Practice. 2(3):209–225, 2021-09.

[99] Deborah Hellman. Discrimination and Social Meaning. In Kasper Lippert-Rasmussen, editor, *The Routledge Handbook of the Ethics of Discrimination*, pages 97–107. Routledge, 1 edition, 2017-08-23.

[100] Michele Loi and Markus Christen. Choosing how to discriminate: Navigating ethical trade-offs in fair algorithmic design for the insurance sector. 34(4):967–992, 2021-12.

[101] Marcelo O. R. Prates, Pedro H. Avelar, and Luís C. Lamb. Assessing gender bias in machine translation: A case study with Google Translate. 32(10):6363–6381, 2020-05.

[102] Maeve McKeown. Structural injustice. 16(7), 2021-07.

[103] Iris Marion Young. *Responsibility for justice*. Oxford University Press, 2011.

[104] Alasia Nuti. *Injustice and the reproduction of history: Structural inequalities, gender and redress*. Cambridge University Press, 2019.

[105] Kate Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

[106] Statista. Possible layoffs in companies in the united states due to chatgpt adoption as of february 2023. February 2023. (accessed: 24. 5. 2023).

[107] Daniel Levi. Chatgpt crosses 1 million users five days after launch. December, 2022. (accessed: 24. 5. 2023).

[108] Nico Grant and Cade Metz. A new chat bot is a 'code red' for google's search business. 2022-12-21. (accessed: 24. 5. 2023).

[109] Christopher Mims. Help! my political beliefs were altered by a chatbot! 13 May, 2023. (accessed: 24. 5. 2023).

[110] Hannah Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frederic A. Dreyer, Aleksandar Shtedritski, and Yuki M. Asano. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models, 2021.

[111] Dhruvil Gala, Mohammad Omar Khursheed, Hannah Lerner, Brendan O'Connor, and Mohit Iyyer. Analyzing gender bias within narrative tropes. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 212–217, Online, November 2020. Association for Computational Linguistics.

[112] Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[113] Ethan Fast, Tina Vachovsky, and Michael Bernstein. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 112–120, 2016.

[114] Evangelia (Lina) Papadaki. Feminist Perspectives on Objectification. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition, 2021.

[115] Susan Moller Okin. *Justice, Gender, and the Family*. New York: Basic Books, 1989.

[116] Rima Basu. The wrongs of racist beliefs. 176(9):2497–2515, 2019-09.

[117] Bias benchmark for qa (bbq) - lite version. https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/bbq_lite/README.md.

[118] Gender sensitivity test - english. https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/gender_sensitivity_english/README.md.

[119] Muslim-violence bias. https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/muslim_violence_bias/README.md.

[120] Stanford University Human-Centered Artificial Intelligence. The ai index report. (accessed: 24. 5. 2023).

[121] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. Overcoming Failures of Imagination in AI Infused System Development and Deployment. 2020-12-10.

[122] Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations. 2022-03-25.

[123] Rosaria Conte, Nigel Gilbert, Giulia Bonelli, Claudio Cioffi-Revilla, Guillaume Deffuant, Janos Kertesz, Vittorio Loreto, Suzy Moat, J P Nadal, Anxo Sanchez,

et al. Manifesto of computational social science. *The European Physical Journal Special Topics*, 214:325–346, 2012.

[124] David MJ Lazer, Alex Pentland, Duncan J Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, et al. Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062, 2020.

[125] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[126] Corina Koolen and Andreas von Cranenburgh. These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22. Association for Computational Linguistics, 2017.

[127] Jigsaw and Google. Perspective api. https://www.perspectiveapi.com/#/.

[128] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. Designing Toxic Content Classification for a Diversity of Perspectives.

[129] Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. "Call me sexist, but..." : Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. page 12.

[130] Peter Glick and Susan T Fiske. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of personality and social psychology*, 70(3):491, 1996.

[131] Ruth E Fassinger. Development and testing of the attitudes toward feminism and the women's movement (fwm) scale. *Psychology of Women Quarterly*, 18(3):389–402, 1994.

[132] Lindsay-Rae McIntyre. Microsoft's 2021 diversity & inclusion report: Demonstrating progress and remaining accountable to our commitments. https://www.nytimes.com/2022/09/24/technology/silicon-valley-slides-back-into-bro-culture.html?smid=url-share, 2022. (accessed: 24. 5. 2023).

[133] Liza Mundy. Why is silicon valley so awful to women? https://www.theatlantic.com/magazine/archive/2017/04/why-is-silicon-valley-so-awful-to-women/517788/?utm_source=copy-link&utm_medium=social&utm_campaign=share, 2017. (accessed: 24. 5. 2023).

[134] Pew Research Center. Reddit news users more likely to be male, young and digital in their news preferences. *https://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male- young-and-digital-in-their-news-preferences*, pages 383–392, 2012, February.

[135] Sen Jia, Thomas Lansdall-Welfare, Saatviga Sudhahar, Cynthia Carter, and Nello Cristianini. Women are seen more than heard in online newspapers. *PloS one*, 11(2):e0148434, 2016.

[136] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 454–463, 2015.

[137] Alexandra Luccioni and Joseph Viviano. What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189. Association for Computational Linguistics, 2021.

[138] Andreas Tsamados, Nikita Aggarwal, Josh Cowls, Jessica Morley, Huw Roberts, Mariarosaria Taddeo, and Luciano Floridi. The ethics of algorithms: Key problems and solutions. 37(1):215–230, 2022-03.

[139] Tom Simonite. Ai is the future—but where are the women? https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/. (accessed: 24. 5. 2023).

[140] Google. Representation at google. https://about.google/belonging/diversity-annual-report/2021/representation/, 2021. (accessed: 24. 5. 2023).

[141] Lindsay-Rae McIntyre. Microsoft's 2021 diversity & inclusion report: Demonstrating progress and remaining accountable to our commitments. https://blogs.microsoft.com/blog/2021/10/20/microsofts-2021-diversity-inclusion-report-demonstrating-progress-and-remaining-accountable-to-our-commitments/, 2021. (accessed: 24. 5. 2023).

[142] Pause giant ai experiments: An open letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/. (accessed: 2. 6. 2023).

[143] Timnit Gebru, Emily M. Bender, Angelina McMillan-Major, and Margaret Mitchell. Statement from the listed authors of stochastic parrots on the "ai pause" letter. 31 March, 2023. (accessed: 2. 6. 2023).

[144] Rob Toews. A wave of billion-dollar language ai startups is coming. https://www.forbes.com/sites/robtoews/2022/03/27/a-wave-of-billion-dollar-language-ai-startups-is-coming/, 27 March, 2022. (accessed: 24. 5. 2023).

[145] Statista. Leading chatbot/conversational ai startups worldwide in 2023, by funding raised. https://www.statista.com/statistics/1359073/chatbot-and-conversational-ai-startup-funding-worldwide/, December 2022. (accessed: 25.05.2023).

[146] Supantha Mukherjee, Elvira Pollina, and Rachel More. Italy's chatgpt ban attracts eu privacy regulators. https://www.reuters.com/technology/germany-principle-could-block-chat-gpt-if-needed-data-protection-chief-2023-04-03/. (accessed: 24. 5. 2023).

[147] Evgeny Morozov. *To Save Everything, Click Here: The Folly of Technological Solutionism*. PublicAffairs, 2013.

[148] Kevin Roose. Google c.e.o. sundar pichai on the a.i. moment: 'you will see us be bold'. https://www.nytimes.com/2023/03/31/technology/google-pichai-ai.html?smid=url-share, March 31, 2023. (accessed: 2. 6. 2023).

[149] Meta AI. We're advancing ai for a more connected world. https://ai.facebook.com/about/. (accessed: 2. 6. 2023).

[150] Karen Hao. We read the paper that forced timnit gebru out of google. here's what it says. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru, 4 December, 2020. (accessed: 24. 5. 2023).

[151] Zoe Schiffer and Casey Newton. Microsoft lays off team that taught employees how to make ai tools responsibly. https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs, 14 March, 2023. (accessed: 2. 6. 2023).

[152] Juliana Shihadeh, Margareta Ackerman, Ashley Troske, Nicole Lawson, and Edith Gonzalez. Brilliance Bias in GPT-3. In *2022 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 62–69, 2022-09.

[153] Arielle Pardes. In a banner year for vc, women still struggle to get funding. https://www.wired.com/story/in-banner-year-for-vc-women-still-struggle-to-get-funding/, 11 October, 2021. (accessed: 2. 6. 2023).

Appendices

Appendix A

Systematic Literature Review

A systematic literature review helps understanding the state of the art in the field. Appropriate selection of papers is an essential process for the review's validity and explanatory power. (Gusenbauer and Haddaway, 2020)[7] To provide a thorough literature review on the measurement of bias in language models, I first retrieved relevant papers from four databases, following the recommendation by Gusenbauer and Haddaway (2020)[7]. Afterwards, I conducted a series of screening processes to identify papers that (1) measure gender bias in language models (2) use a novel method, instead of reusing previously proposed methods to measure bias.

A.1 Purpose of Paper Retrieval

In this thesis, I conducted a systematic literature review on works measuring bias in language models. I targeted papers suggesting novel methods for measuring gender bias in language models. Initial search of papers included papers measuring any types of bias, not necessarily gender bias, to be as inclusive as possible. Through further screening processes, the final list was reduced to 19 papers.

A.2 Paper Retrieval

I chose four databases: Scopus, Web of Science, ACM Library, and ACL Anthology. The first three databases were assessed as appropriate for principal sources by Gusenbauer and Haddaway (2020). Scopus and Web of Science are multidisciplinary databases that contain more than 70 million papers. ACM Library is a computer science database with more than 2 million entries. All three databases allow search based on a query using logical operators and export the research results in various file formats including Bibtex. ACL Anthology was added later to be as extensive as possible, as it is a database for computational linguistics. As most measurements of bias in language model papers belong to the category of computational linguistics, ACL Anthology was added as a complementary source despite being small in size (83,104). Unlike other databases,

ACL Anthology does not provide the functionality to search papers on the basis of a query string. However, as this database provides a python library to access all entries in the anthology, I accessed the ACL Anthology databse with python script and applied the same component in the query to retrieve relevant papers. Then, I combined every result with the information provided information about authors, title, abstract, year of publication.

To compose a query, I included two elements: title and abstract. In my query, the title includes at least one term about the model and at least one term about bias. The terms for the model includes 'language model', 'transformer', 'bert', 'roberta', 'xlnet', 'gpt', 'sentence encoder', 'contextual embedding', 'masked', 'generative', and 'pretrained'. The terms for bias includes 'bias', 'gender bias', 'social bias', 'societal bias', and 'implicit bias'. The query for abstract includes all terms given above for the title and at least one term related to measurement, such as 'measure', 'metric', or 'mitigate'. The keyword 'mitigate' was added to make the search as inclusive as possible, since mitigating bias in computational method presumes defining and measuring bias. However, through the screening process, papers focusing on mitigation was removed for the practical limitation to make the review manageable size for this thesis.

DB	Topic	Size	Retrieved Papers
SCOPUS	Multidisciplinary	70M+	146
WoS	Multidisciplinary	70M+	143
ACM	Computer science	2M+	124
ACL Anthology	Computational linguistics	80K+	68

Table A.1: Databases used for paper retrieval

As a result of the querying four databases, 481 papers were retrieved (SCOPUS: 146, ACM Anthology: 124, Web of Science: 143 results). After removing duplicates, 370 papers were remaining.

DB	ACM Library
Search Run Date	2023-01-24 at 07:52:32 PST
Search Result Count	124
Query Syntax	{Title‘:(language ’model OR transformer OR bert OR roberta OR xlnet OR gpt OR ”sentence encoder” OR ”contextual embedding” OR masked OR generative OR pretrained) AND Title:(Bias OR ”gender bias” OR ”social bias” OR ”societal bias” OR ”implicit bias”) AND Abstract:(measure OR metric OR mitigate OR bias OR gender OR social OR societal OR implicit OR ”language model” OR transformer OR bert OR roberta OR xlnet OR gpt OR ”sentence encoder” OR ”contextual embedding” OR masked OR generative OR pretrained)} ”filter”: {ACM Content: DL}
Url	https://dl.acm.org/action/doSearch?fillQuickSearch=false&target=advanced&expand=dl&AllField=Title%3A%28%E2%80%98language+model%E2%80%99+OR%C2%A0transformer%C2%A0OR%C2%A0bert%C2%A0OR%C2%A0roberta%C2%A0OR%C2%A0+xlnet%C2%A0OR+gpt%C2%A0OR+%22sentence+encoder%22%C2%A0OR%C2%A0%22contextual+embedding%22%C2%A0OR+masked+OR%C2%A0generative+OR+pretrained%C2%A0%29+AND+Title%3A%28Bias+OR+%22gender+bias%22+OR%C2%A0%22social+bias%22+OR+%22societal+bias%22+OR+%22implicit+bias%22%29+AND+Abstract%3A%28%C2%A0measure%C2%A0OR%C2%A0metric+OR+mitigate+OR+bias+OR+gender+OR+social+OR+societal%C2%A0OR+implicit%C2%A0OR%C2%A0%22language+model%22+OR%C2%A0transformer%C2%A0OR%C2%A0bert+OR+roberta%C2%A0OR+xlnet+OR%C2%A0gpt%C2%A0OR%C2%A0%22sentence+encoder%22%C2%A0OR%C2%A0%22contextual+embedding%22%C2%A0OR%C2%A0masked%C2%A0OR%C2%A0generative+OR+pretrained%29

DB	SCOPUS
Search Run Date	2023-01-24 at 16:54 CET
Search Result Count	146
Query Syntax	

```
( TITLE ( "language model" OR transformer OR bert OR roberta OR xlnet OR
    gpt OR "sentence encoder" OR "contextual embedding" OR masked OR
   generative OR pretrained ) AND TITLE ( bias OR "gender bias" OR "
   social bias" OR "societal bias" OR "implicit bias" OR stereotype )
   AND ABS ( measure OR metric OR mitigate OR bias OR gender OR social
   OR societal OR implicit OR "language model" OR transformer OR bert OR
    roberta OR xlnet OR gpt OR "sentence encoder" OR "contextual
   embedding" OR masked OR generative OR pretrained ) AND NOT TITLE-ABS-
   KEY ( dc OR voltage OR converter ) )
```

Url	https://www.scopus.com/results/results.uri?sort=plf-f&src=s&st1=%22language+model%22+OR+transformer+OR+bert+OR+roberta+OR+xlnet+OR+gpt+OR+%22sentence+encoder%22+OR+%22contextual+embedding%22+OR+masked+OR+generative+OR+pretrained&st2=bias+OR+%22gender+bias%22+OR+%22social+bias%22+OR+%22societal+bias%22+OR+%22implicit+bias%22+OR+stereotype&searchTerms=measure+OR+metric+OR+mitigate+OR+bias+OR+gender+OR+social+OR+societal+OR+implicit+OR+%22language+model%22+OR+transformer+OR+bert+OR+roberta+OR+xlnet+OR+gpt+OR+%22sentence+encoder%22+OR+%22contextual+embedding%22+OR+masked+OR+generative+OR+pretrained%3F%21%22*%24dc+OR+voltage+OR+converter%3F%21%22*%24&sid=268b51bb80bc38b889c5bd1240a54a9b&sot=b&sdt=b&sl=560&s=%28TITLE%28%22language+model%22+OR+transformer+OR+bert+OR+roberta+OR+xlnet+OR+gpt+OR+%22sentence+encoder%22+OR+%22contextual+embedding%22+OR+masked+OR+generative+OR+pretrained%29+AND+TITLE%28bias+OR+%22gender+bias%22+OR+%22social+bias%22+OR+%22societal+bias%22+OR+%22implicit+bias%22+OR+stereotype%29+AND+ABS%28measure+OR+metric+OR+mitigate+OR+bias+OR+gender+OR+social+OR+societal+OR+implicit+OR+%22language+model%22+OR+transformer+OR+bert+OR+roberta+OR+xlnet+OR+gpt+OR+%22sentence+encoder%22+OR+%22contextual+embedding%22+OR+masked+OR+generative+OR+pretrained%29+AND+NOT+TITLE-ABS-KEY%28dc+OR+voltage+OR+converter%29%29&origin=searchbasic&editSaveSearch=&yearFrom=Before+1960&yearTo=Present&featureToggles=FEATURE_DOCUMENT_RESULT_MICRO_UI%3A1&sessionSearchId=268b51bb80bc38b889c5bd1240a54a9b&limit=10
DB	Web of Science
Search Run Date	2023-01-24 at 16:55 CET
Search Result Count	143

Query Syntax	TI=("language model" OR transformey OR bert OR roberta OR xlnet OR gpt OR "sentence encoder" or "contextual embedding" OR masked OR generative OR pretrained) AND TI=(bias OR "gender bias" OR "social bias" OR "societal bias" OR "implicit bias" OR stereotype) AND AB=(measure OR metric OR quantify OR bias OR gender OR social OR societal OR implicit OR "language model" OR transformer OR bert OR roberta OR xlnet OR gpt OR 'sentence encoder' OR 'contextual embedding' OR masked OR generative OR pretrained) NOT AB=(DC OR converter OR voltage) NOT TI=(DC OR converter OR voltage)
Url	https://www.webofscience.com/wos/woscc/summary/d4f07f16-eeb8-461c-be8a-e19c097b679a-6be77610/relevance/1

Table A.2: Queries for Each DB and URL

Table A.2 shows the time when the query was conducted, and the number of results from each database. Unlike other databases, papers from ACL Anthology were retrieved programmatically using python script, which is attached in the Appendix B.

A.3 Script for Retrieving Papers from ACL Anthology

```
from anthology import Anthology
import pandas as pd
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
pd.set_option('display.max_colwidth', None)
pd.set_option('display.max_seq_items', None)
anthology = Anthology(importdir='acl-anthology/data')

title_model = ['language model', 'transformer', 'bert', 'roberta', 'xlnet',
  'gpt', 'sentence encoder', 'contextual embedding', 'masked', 'generative',
  'pretrained']
title_bias = ['bias', 'gender bias', 'social bias', 'societal bias',
  'implicit bias']
abstract_measure = ['measure', 'metric', 'mitigate', 'quantify'] +
  title_model + title_bias

result = []
df = pd.DataFrame()
for idx, (id_, paper) in enumerate(anthology.papers.items()):
  data = paper.as_dict()
  title = paper.get_title("plain").lower()
  abstract = paper.get_abstract("plain").lower()
  if (any([term in title for term in title_model])
      and any([term in title for term in title_bias])):
    #print(paper.anthology_id, paper.get_title('text'))
    if any([term in abstract for term in abstract_measure]):
      result.append(title)
      #print(paper.anthology_id, paper.get_title('text'))
      if "author" in data:
        data["author"] = [
          anthology.people.resolve_name(name, id_)
            for name, id_ in data["author"]
        ]
        last_names = ''
```

```
        for names in data['author']:
          last_names += names['last'] + ', '

        entry = {'Authors': last_names,
             'Title': paper.get_title("plain"),
             'Abstract': paper.get_abstract("plain"),
             'Year': paper.attrib['year'],
             'Source': 'ACL Anthology',
             'Link':'',
             'Criteria 1':'',
             'Criteria 2':'',
             'Second Eval':'',
             'First Eval':''}

        df = df.append(entry, ignore_index=True)


df.to_csv('acl_result.csv')

print(idx)
print(result)
print(len(result))
```

A.4 Screening process

Screening process is composed of four rounds. The first round was based on titles and the other two rounds were based on abstracts. In the last round, the full content of the papers was evaluated to determine whether the paper fits the criteria for the respective round of screening.

In the first round, papers that were clearly out of the scope were excluded, such as papers in the field of energy science or psychology. Moreover, I excluded biases that are not considered morally significant, such as inductive bias, reporting bias, or selection bias. For detailed information on various definitions of biases, see Chapter 1. In addition, papers in which language models are used to detect or measure bias, stereotypes, or hate speech in news media, social media, randomized controlled trial, and text were excluded as they were not measuring bias in language models. Rather, language models were used as a tool to measure bias in other content.

In the second round, papers that are measuring other morally significant biases other than gender bias, such as religious bias, political bias, or racial bias, were excluded on

the basis of abstract. Also, papers that focus on embedding, machine learning, AI, or visual models were excluded as language models and their linguistic capabilities are the principal matter of investigation. Therefore, papers published before Devlin et al., (2019) which introduced BERT model were excluded. But if the paper investigates embeddings with other language models, they remained in the list.

In the third round, papers with a specific focus on downstream tasks, such as machine translation or coreference resolution, were excluded. As the purpose of the literature review is to provide an overview of bias measurement in language models in general, focusing on specific downstream tasks might be too narrow in scope.

The last round of screening identified papers that focuse primarily on measuring bias. Papers for mitigating bias were inclulded in the query for the purpose of extensive search, since mitigating bias inevitably entails measuring bias. However, due to practical constraints, papers primarily focusing on mitigating bias were excluded in the last round. Such papers are in general relevant for the topic of the thesis but were excluded for in-depth review concerning their methodologies. Another critical criteria for choosing relevant papers was whether they present a novel method to measure bias, that differs from previously suggested method. However, if the paper applied the existing method to different types of data (e.g. Kwon and Mihindukulasooriya (2022)), the paper remained in the list for the review since it contributed to the validity of measurement. On the other hand, if the same method was applied to the same type of data, i.e. data collected using the same method (e.g. French CrowS-Pairs that collected the same type of data as Enghlish CrowS-Pairrs but in French), the paper was excluded. Lastly, pubslication that are not full papers (e.g. conference proposal) were excluded.

After conducting the search with the query in each database, I exported BibTeX, which I converted into CSV in order to merge in one spreadsheet file. For the screening process, I went through the titles and abstracts from the combined spreadsheet file.

A.5 Complete list of papers retrieved

	Round 1	**Round 2**	**Round 3**	**Round 4**	**Final Papers**
Valid	89	54	44	37	19
Invalid	253	53	4	23	14
Unclear	28	10	16	0	4
Total	370	117	64	60	37

Table A.3: Paper Screening Rounds

Papers invalidated in the round 1:

[A1] [A2] [A3] [A4] [A5] [A6] [A7] [A8] [A9] [A10] [A11] [A12] [A13] [A14] [A15] [A16] [A17] [A18] [A19] [A20] [A21] [A22] [A23] [A24] [A25] [A26] [A27] [A8] [A29] [A30] [A31]

[A32] [A33] [A34] [A35] [A36] [A37] [A38] [A39] [A40] [A41]
[A42] [A43] [A44] [A45] [A46] [A47] [A48] [A49] [A50] [A51]
[A52] [A53] [A54] [A55] [A56] [A57] [A58] [A59] [A60] [A61]
[A62] [A63] [A64] [A65] [A66] [A67] [A68] [A69] [A70] [A71]
[A72] [A73] [A74] [A75] [A76] [A77] [A78] [A79] [A80] [A81]
[A82] [A83] [A84] [A85] [A86] [A87] [A88] [A89] [A90] [A91]
[A92] [A93] [A94] [A95] [A96] [A97] [A98] [A99] [A100] [A101]
[A102] [A103] [A104] [A105] [A106] [A107] [A108] [A109] [A110]
[A111] [A112] [A113] [A114] [A115] [A116] [A117] [A118] [A119]
[A120] [A121] [A122] [A123] [A124] [A125] [A126] [A127] [A128]
[A129] [A130] [A131] [A132] [A133] [A134] [A135] [A136] [A137]
[A138] [A139] [A140] [A141] [A142] [A143] [A144] [A145] [A146]
[A147] [A148] [A149] [A150] [A151] [A152] [A153] [A154] [A155]
[A156] [A157] [A158] [A159] [A160] [A161] [A162] [A163] [A164]
[A165] [A166] [A167] [A168] [A169] [A170] [A171] [A172] [A173]
[A174] [A175] [A176] [A177] [A178] [A179] [A180] [A181] [A182]
[A183] [A184] [A185] [A186] [A187] [A188] [A189] [A190] [A191]
[A192] [A193] [A194] [A195] [A196] [A197] [A198] [A199] [A200]
[A201] [A202] [A203] [A204] [A205] [A206] [A207] [A208] [A209]
[A210] [A211] [A212] [A213] [A214] [A215] [A216] [A217] [A218]
[A219] [A220] [A221] [A222] [A223] [A224] [A225] [A226] [A227]
[A228] [A229] [A230] [A231] [A232] [A233] [A234] [A235] [A236]
[A237] [A238] [A239] [A240] [A241] [A242] [A243] [A244] [A245]
[A246] [A247] [A248] [A249] [A250]

Papers invalidated in the round 2:

[A251] [A252] [A253] [A254] [A255] [A256] [A257] [A258] [A259]
[A260] [A261] [A262] [A263] [A264] [A265] [A266] [A267] [A268]
[A269] [A270] [A271] [A272] [A273] [A274] [A71] [A275] [A276]
[A277] [A278] [A279] [A280] [A281] [A282] [A283] [A284] [A285]
[A286] [A287] [A288] [A107] [A289] [A290] [A291] [A292] [A293]
[A294] [A295] [A296] [A297] [A298] [A299] [A300] [A301]

Papers invalidated in the round 3:

[A181] [A302] [A303] [A304]

Papers invalidated in the round 4:

[A305] [A306] [A307] [A308] [A309] [A310] [A311] [A312] [A313]
[A314] [A315] [A316] [A317] [A318] [A319] [A320] [A321] [A322]
[A323] [A324] [A325] [A326] [A327]

Table A.4: The Final List Reviewed Papers

Index	Authors	Title
1	May et al. (2019)	On measuring social biases in sentence encoders
2	Kirk et al. (2021)	Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models
3	Dhamala et al. (2021)	BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation
4	Nangia et al. (2020)	CrowS-Pairs: A challenge dataset for measuring social biases in masked language models
5	Lucy and Bamman (2021)	Gender and Representation Bias in GPT-3 Generated Stories
6	Jentzsch and Turan (2022)	Gender Bias in BERT - Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task
7	Kwon and Mihindukulasooriya (2022)	An Empirical Study on Pseudo-log-likelihood Bias Measures for Masked Language Models Using Paraphrased Sentences
8	Steinborn et al. (2022)	An Information-Theoretic Approach and Dataset for Probing Gender Stereotypes in Multilingual Masked Language Models
9	Sotnikova et al. (2021)	Analyzing Stereotypes in Generative Text Inference Tasks
10	Kaneko et al. (2022)	Gender Bias in Masked Language Models for Multiple Languages
11	Wolfe and Caliskan (2021)	Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models
12	Touileb et al. (2022)	Occupational Biases in Norwegian and Multilingual Language Models
13	Nadeem et al. (2019)	StereoSet: Measuring stereotypical bias in pretrained language models
14	Silva et al. (2021)	Towards a Comprehensive Understanding and Accurate Evaluation of Societal Biases in Pre-Trained Transformers
15	Alnegheimish et al. (2022)	Using Natural Sentences for Understanding Biases in Language Models
16	Barikeri et al. (2021)	REDDITBIAS: A real-world resource for bias evaluation and debiasing of conversational language models
17	Shen et al. (2023)	Towards understanding and mitigating unintended biases in language model-driven conversational recommendation
18	Bartl et al. (2020)	Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias
19	de Vassimon Manela et al. (2021)	Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models

Appendix B

Measurement Modeling

B.1 Construct Reliability and Construct Validity

			Construct Reliability		Construct Validity		
Paper Index	**Authors**	**Metric category**	**Inter-rater (annotator), test-retest reliability**	**Face Validity**	**Content Validity**	**Convergent Validity / Discriminant validity**	**Consequential Validity**
1	May et al., 2019	Discrepancy in cosine similarity	Not tested	Valid(Undesirable stereotypes)	Based on established literature outside NLP (ABW, Double Binds)	Showed that word-level Caliskan (WEAT) also work in sentence-level (SEAT)	"However, we strongly caution against interpreting the number of significant associations or the average significant effect size as an absolute measure of bias. Like WEAT, SEAT only has positive predictive ability: It can detect presence of bias, but not its absence." (p. 626)
2	Kirk et al., 2021	Generated text analysis (Frequency of jobs)	Not tested	Valid(Occupational stereotypes)	No theoretical background on relevant disciplines (reference limited to NLP work)	Not tested	Not discussed

			Construct Reliability		Construct Validity		
Paper Index	**Authors**	**Metric category**	**Inter-rater (annotator), test-retest reliability**	**Face Validity**	**Content Validity**	**Convergent Validity / Discriminant validity**	**Consequential Validity**
3	Dhamala et al., 2021	Sentiment score, Toxicity, Regards, Psycholinguistic norms, Gender Polarity	Not tested	Gender Polarity: questionable - includes simple counting of gendered words with potential confounders	No theoretical background on relevant disciplines (reference limited to NLP work)	Validate metrics with human judgement from crowd-sourced workersUse multiple metrics	Not discussed
4	Nangia et al, 2020	Likelihood of masked token prediction	Not tested	Valid (Stereotypes)	No theoretical background on relevant disciplines (reference limited to NLP work)	Compare with WinoBias and StereoSet as baselines, and found that all three models exhibit substantial bias.	Not discussed

			Construct Reliability		Construct Validity		
Paper Index	**Authors**	**Metric category**	**Inter-rater (annotator), test-retest reliability**	**Face Validity**	**Content Validity**	**Convergent Validity / Discriminant validity**	**Consequential Validity**
5	Lucy and Bamman, 2021	Topic modeling, Lexicon-based analysis using cosine (semantic) similarity	Not tested	Valid (Stereotypes)	Refers to existing research on gender, linguistics, and media beyond NLP (Smith et al., 2012, Kraicer and Piper, 2018; Johns and Dye, 2019)	Use two metrics and show coherent result, confirms findings with existing research (Underwood et al., 2018; Kraicer and Piper, 2018; Johns and Dye, 2019, Smith et al., 2012; Sap et al., 2017; Fast et al., 2016b; Gala et al., 2020)	Not discussed
6	Jentzsch and Turan, 2022	Sentiment analysis	Not tested	Valid (Sentiment associated with gendered nouns)	Based on established literature outside NLP, Acknowledgement of the limitation of proposed method	Not tested	Not discussed
7	Kwon and Mihindukulasooriya, 2022	Likelihood of masked token prediction	Not tested	Valid (Stereotypes)	Based on existing NLP work (CrowS-Pairs)	Tests Convergent Validity with CrowS-Pairs (Nangia et al., 2020) by paraphrasing sentences in the dataset	Not discussed

			Construct Reliability		Construct Validity		
Paper Index	**Authors**	**Metric category**	**Inter-rater (annotator), test-retest reliability**	**Face Validity**	**Content Validity**	**Convergent Validity / Discriminant validity**	**Consequential Validity**
8	Steinborn V., Dufter P., Jabbar H., Schütze H.	Likelihood of masked token prediction (SJSD)	Not tested	Valid (Stereotypes)	No theoretical background on relevant disciplines (reference limited to NLP work)	Explores potential spurious correlation (not experimentally confirmed)	Not discussed
9	Sotnikova et al., 2021	Natural language inference, Commonsense inference, comparison with human judgement	Inter-annotator agreement percentage (Fraction of times all annotators give the same answer)	Valid (Stereotypes)	No theoretical background on relevant disciplines (reference limited to NLP work)	The model behaviour (hypothesis generation) is compared with human judgements	Not discussed
10	Kaneko et al., 2022	Likelihood of masked token prediction	Not tested	Questionable - weighted average with all masked token prediction might introduce confounders	No theoretical background on relevant disciplines (reference limited to NLP work)	Confirms the result with related work (CP, SS)	Not discussed
11	Wolfe and Caliskan, 2021	Cosine similarity (WEAT)	Not tested	Valid (Frequency of names and association with unpleasantness)	Refers to existing research on Human bias and perceptions (Hughes et al., 2019, Greenwald et al., 1998)	Not tested	Not discussed

			Construct Reliability		Construct Validity		
Paper Index	**Authors**	**Metric category**	**Inter-rater (annotator), test-retest reliability**	**Face Validity**	**Content Validity**	**Convergent Validity / Discriminant validity**	**Consequential Validity**
12	Touileb et al., 2022	Likelihood of masked token prediction	Not tested	Valid (Occupational stereotypes)	No theoretical background on relevant disciplines (reference limited to NLP work)	Not tested	Not discussed
13	Nadeem et al., 2019	Likelihood of masked token prediction (CATs, icat)	Not tested	Valid (Stereotypes)	No theoretical background on relevant disciplines (reference limited to NLP work)	Not tested	Not discussed
14	Silva et al., 2021	WEAT, SEQ, PN	Not tested	Valid (Association of sentiment, likelihood comparison)	No theoretical background on relevant disciplines (reference limited to NLP work)	Use multiple testsConfirms the result with related work (Hooker et al. (2020))Questions validity of WEAT	Not discussed
15	Alnegheimish et al., 2022	Likelihood of prompt based language generation	Not tested	Valid (Occupational Stereotypes)	No theoretical background on relevant disciplines (reference limited to NLP work)	Confirms the result with related work (Vig et al., 2020)	Not discussed

			Construct Reliability		Construct Validity		
Paper Index	**Authors**	**Metric category**	**Inter-rater (annotator), test-retest reliability**	**Face Validity**	**Content Validity**	**Convergent Validity / Discriminant validity**	**Consequential Validity**
16	Barikeri et al., 2021	Perplexity, Dialog state tracking (DST), Conversational response generation (CRG)	Inter-annotator agreement (Krippendorff's alpha)	Valid (Stereotypes)	Refers to sociological literature relating to the minoritized groups (Welch, 2007; Shaw, 2012; Black, 2015)	Both intrinsic (LMB) and extrinsic measurements (LMP, DST) of LM	Not discussed
17	Shen et al., 2023	Price Percentage Score,Association Score (WEAT)	Not tested	Valid (Comparison of recommended prices between groups)	No theoretical background on relevant disciplines (reference limited to NLP/RecSys work)	Not tested	Not discussed
18	Bartl et al., 2020	Predicting masked token (WEAT inspired)	Not tested	Valid (Occupational Stereotypes)	Refers to existing research (Moss-Racusin et al. 2012), but it is not closely related to the operationalizationAcknowledges potential human bias through researcher's choices	Confirms and extends previous resarch (Kurita et al., 2019); Tests the measurements in two languages (English and German) and compares the result	Not discussed

			Construct Reliability		**Construct Validity**		
Paper Index	**Authors**	**Metric category**	**Inter-rater (annotator), test-retest reliability**	**Face Validity**	**Content Validity**	**Convergent Validity / Discriminant validity**	**Consequential Validity**
19	de Vassimon Manela et al., 2021	Performance parity between strereotypical and anti-stereotypical with respect to gender	Not tested	Valid (Occupational Stereotypes)	No theoretical background on relevant disciplines (reference limited to NLP work)	Compares with existing gender bias benchmarks - WinoBias (Questions the validity)	Not discussed

Table B.1: Measurement Modeling Applied to Gender Bias Measurements

B.2 Levels of Conceptualization

Index	Authors	Level 1: Background definition	Level 2: Systematized definition	Level 3: Indicator	Discussion
1	May et al., 2019	"In the Sapphire or angry black woman (ABW) stereotype, black women are portrayed as loud, angry, and imposing (Collins, 2004; Madison, 2009; HarrisPerry, 2011; hooks, 2015; Gillespie, 2016)." (p. 624), "Double Binds Women face many double binds, contradictory or unsatisfiable expectations of femininity and masculinity (Stone and Lovejoy, 2004; Harris-Perry, 2011; Mitchell, 2012)(p. 624)	Not Provided	Discrepancy in cosine similarity	Referring to relevant literature; indicator aligns with the background definition
2	Kirk et al., 2021	Not Provided	Not Provided	Generated text analysis (Frequency of jobs)	No definition provided

Index	Authors	Level 1: Background definition	Level 2: Systematized definition	Level 3: Indicator	Discussion
3	Dhamala et al., 2021	Not Provided	"Broadly, one can say a language generation model is biased if it disproportionately generates text that is often perceived as being negative, unfair, prejudiced, or stereotypical against an idea or a group of people with common attributes." (p. 862) "Prompts from gender, race, religious belief, and political ideology domains trigger a text generation model to generate text given a context referring to a person or an idea. In these cases, we are interested in examining the positive or negative feelings in the generated texts." (p. 864)	Sentiment score, Toxicity, Regards, Psycholinguistic norms, Gender Polarity	Conflated notion of bias (negative, unfair, prejudicced, stereotypical) without conceptualizating individual definitions
4	Nangia et al, 2020	Not Provided	Not Provided	Likelihood of masked token prediction	No definition provided
5	Lucy and Bamman, 2021	Not Provided	Not Provided	Topic modeling, Lexicon-based analysis using cosine (semantic) similarity	No definition provided

Index	Authors	Level 1: Background definition	Level 2: Systematized definition	Level 3: Indicator	Discussion
6	Jentzsch and Turan, 2022	Not Provided	"We study how representational male and female gender concepts are assessed differently in sentiment classification systems" (p. 185) Sentiment prediction score ("First, a novel bias measure is introduced, defining biases as the difference in sentiment valuation of female and male sample versions." (p. 184))	Mismatch systematized definition and indicator - Link between representation of male and female gender and discrepancy between sentiment classification is not established	
7	Kwon and Mihindukulasooriya, 2022	"A cognitive bias, stereotyping, is defined as the assumption of some characteristics are applied to communities on the basis of their nationality, ethnicity, gender, religion, etc (Schneider, 2005). Relatedly, Fairness ("zero-bias"), in the context of NLP and machine learning is defined as preventing harmful, discriminatory decisions according to such unwanted, stereotypical characteristics (Garrido-Muñoz et al., 2021)."	Not Provided	Likelihood of masked token prediction	Background concept and indicator aligns (Based on CrowS-Pairs)
8	Steinborn et al., 2022	Not Provided	"we tackle this type of binary stereotypical representational gender bias (henceforth simply "gender bias") in MLMs in a multilingual setting." (p. 921)	Likelihood of masked token prediction (SJSD)	Background concept and indicator aligns (Based on CrowS-Pairs)

Index	Authors	Level 1: Background definition	Level 2: Systematized definition	Level 3: Indicator	Discussion
9	Sotnikova et al., 2021	"Stereotypes are inferences drawn about people based on their demographic attributes, which may result in harms to users when a system is deployed" (p. 4052) "compare how perceptions of stereotypes vary due to annotator positionality" (p. 4052) "sterotypes exist also in 'the fabric of society' itself" (Stangor and Schaller, 2012), and as such who the annotators are matters (Hovy and Spruit, 2016; Jørgensen et al., 2015; Hazen et al., 2020)." (p. 4053) "Our work builds on a growing body of recent computational literature on stereotypes (often termed "bias")." (p. 4053)	Not Provided	Natural language inference, commonsense inference, compare with human judgement	Background concept and indicator aligns + Human (authors) evaluation

Index	Authors	Level 1: Background definition	Level 2: Systematized definition	Level 3: Indicator	Discussion
10	Kaneko et al., 2022	Not Provided	"The bias in MLMs is evaluated by the imbalance of the likelihood between pairs of sentences associated with an attribute that has a common context (e.g. He/She is a nurse)." (p. 2740 "In this study, (social) bias is defined as the tendency towards outputting sentences about a particular advantageous or disadvantageous group, such as males or females, given the same context by an MLM" (p. 2742)	Likelihood of masked token prediction	No background concept, systematized definition is not conceptualized; what they are measuring is discrepancy between male and female, while their systematized definition involving disadvantaged and advantaged and social bias is unclear
11	Wolfe and Caliskan, 2021	Not Provided	Not Provided	Cosine similarity (WEAT)	No definition provided

Index	Authors	Level 1: Background definition	Level 2: Systematized definition	Level 3: Indicator	Discussion
12	Touileb et al., 2022	"We follow the bias definition of Friedman and Nissenbaum (1996), where bias is defined as the cases where automated systems exhibit a systematic discrimination against, and unfairly process, a certain group of individuals" (p. 201) "Another definition of bias that we rely on is that of Shah et al. (2020), where bias is defined as the discrepancy between the distribution of predicted and ideal outcomes of a model." (p. 201)	"In our case, we see this as reflected in large pre-trained language models and how they can contain skewed gendered representations that can be systematically unfair if this bias is not uncovered and properly taken into account in downstream applications" (p. 201) "We focus on the associations between gendered (female and male) pronouns/names and professional occupations. We investigate to what degree pre-trained language models systematically associate specific genders with given occupations."(p. 201)	Likelihood of masked token prediction	Background and systematized definition align with the indicator
13	Nadeem et al., 2019	"A stereotype is an over-generalized belief about a particular group of people, e.g., Asians are good at math or Asians are bad drivers. Such beliefs (biases) are known to hurt target groups." (p. 1)	"In this work, we assess the stereotypical biases of popular pretrained language models." (p. 1) "If the model consistently prefers stereotypes over anti-stereotypes, we can say that the model exhibits stereotypical bias" (p. 2)	Likelihood of masked token prediction (Intra- and Inter- sentence Context Association Test, icat)	Background and systematized definition align with the indicator

Index	Authors	Level 1: Background definition	Level 2: Systematized definition	Level 3: Indicator	Discussion
14	Silva et al., 2021	Not Provided	"In the context of our work, "bias" refers specifically to the preference of a model for one gender or race in the presence of an otherwise neutral context." (p. 1) "With no additional information, an equitable system would exhibit no preference for female over male, or African-American over European-American names; however, our results indicate that there is often a statistically significant preference ($p < 0.0001$) for associating female and African-American identifiers with being more "emotional."" (p. 1)	WEAT (Word Embedding Association Test), Sequence Ranking (SEQ), Pronoun Ranking (PN)	Systematized definition aligns with the indicator
15	Alnegheimish et al., 2022	Not Provided	Not Provided	Likelihood of prompt based language generation	No definition provided

Index	Authors	Level 1: Background definition	Level 2: Systematized definition	Level 3: Indicator	Discussion
16	Barikeri et al., 2021	Not Provided	"To measure or mitigate a bias, one must first formalize (i.e., specify) it. To this end, we start from the concept of an explicit bias specification (Caliskan et al., 2017; Lauscher et al., 2020a): an explicit bias specification BE = (T1, T2, A1, A2) consists of two sets of target terms or phrases T1 and T2 between which a bias is expected to exist w.r.t. two sets of attribute terms or phrases A1, and A2. Further, we opt for bias specifications that reflect the inequality between groups in power, i.e., dominant groups, and discriminated groups, i.e., minoritized groups:1 for each BE, the set T1 consists of terms describing a minoritized group with (negative) stereotypical terms in A1, while T2 consists of terms describing a dominant group with (positive) stereotypical terms in A2. We compile bias specifications as follows." (p. 1942)	Language Model Bias (LMB) ("We estimate bias in conversational LMs by measuring if (and how much) likelier the LM is to generate a stereotypically biased phrase compared to a corresponding inversely biased phrase in which we replace t1 T1 with a t2 T2." (p. 1944)	Mismatch between systematized concept (inequality, groups in power, dominant/discriminated group) and indicator (stereotypes)

Index	Authors	Level 1: Background definition	Level 2: Systematized definition	Level 3: Indicator	Discussion
17	Shen et al., 2023	"In brief, unfairness in recommendations manifests as systematic discrimination against specific individuals in favour of others (Friedman & Nissenbaum, 1996) based on protected attributes such as gender and age. Research studies usually perform an attribute-based analysis of fairness in recommender systems, where users or items are labelled with some attributes that cluster them into groups." (p. 2)	"In this paper, we study a simple LM-driven recommendation backbone (termed LMRec) for CRSs to investigate how unintended bias manifests in substantially shifted price and category distributions of restaurant recommendations." (p. 2) "We define unintended bias in language-based recommendation as a systematic shift in recommendations corresponding to nonpreferentially related changes in the input (e.g., a mention of a friend's name)." (p. 6)	Price Percentage Score, Association Score (WEAT) (In this work, in order to evaluate unintended bias, we first leverage a template-based analysis that is popularly used in research work on fairness and bias issues in pretrained language models (Kurita et al., 2019a; May et al., 2019; Sheng et al., 2019; Tan & Celis, 2019), to collect recommendation results over the bias types outlined in Table 2." (p. 6))	Background and systematized definition align with the indicator

Index	Authors	Level 1: Background definition	Level 2: Systematized definition	Level 3: Indicator	Discussion
18	Bartl et al., 2020	"Gender bias is the systematic unequal treatment on the basis of gender (Moss-Racusin et al., 2012; Sun et al., 2019)." (p. 2)	"In the context of our study of the BERT language model, gender bias occurs when one gender is more closely associated with a profession than another in language use, resulting in biased language models. Against the backdrop of gender participation statistics, we can assess whether a biased representation is related to the employment situation in the real world or based on stereotypes." (p. 2) "we measure gender bias by studying associations between gender-denoting target words and names of professions in English and German, comparing the findings with real-world workforce statistics." (p. 1)	Predicting masked token (WEAT inspired)	Background and systematized definition align with the indicator
19	de Vassimon Manela et al., 2021	Not Provided	Not Provided	Performance parity between strereotypical and anti-stereotypical with respect to gender	No definition provided

B.3 Normative Motivations

Index	Authors	Normative Motivation
1	May et al., 2019	"However, prominent word embeddings such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) encode systematic biases against women and black people (Bolukbasi et al., 2016; Garg et al., 2018, i.a.), implicating many NLP systems in scaling up social injustice. We investigate whether sentence encoders, which extend the word embedding approach to sentences, are similarly biased" (May et al., 2019, p. 622)We advocate for further consideration of intersectionality in future work in order to avoid reproducing the erasure of multiple minorities who are most vulnerable to bias" (May et al., 2019, p. 626)
2	Kirk et al., 2021	"What should be the goal of generative language models? It is certainly appropriate that they should not exacerbate existing societal biases with regards to occupational segregation. It is less clear whether they should reflect or correct for skewed societal distributions" (Kirk et al., p. 10)"In this paper, we consider both representational and allocational harms 4). We attempt to elucidate representational harms, or those harmful in their own right, by highlighting occupation-related stereotypes that may propagate negative generalizations about particular social groups. For example, women's higher likelihood of being associated with care-oriented occupations may perpetuate unwanted stereotypes. Especially within the context of occupations, such associations may lead to allocation harms. Frequent stereotypical association of certain demographic groups with a subset of occupations may lead to conditioned expectations in job hiring where a certain individual is predicted to be well-suited for a job based on their demographics" (Kirk et al., p. 2) "We further compare these to real-world occupation data from the US Labor Bureau to map model biases to systemic societal biases." (Kirk et al., p. 2)"This raises the normative question of what language models should learn - whether they should reflect or correct for existing inequalities." (Kirk et al., p. 1)
3	Dhamala et al., 2021	"Recently, there has been growing evidence on how machine learning models without proper fairness checks risk reinforcing undesirable stereotypes, subjecting users to disparate treatment and enforcing de facto segregation." (Dhamala et al., 2021, p. 862)
4	Nangia et al, 2020	"However, there is ample evidence that they use the cultural biases that are undoubtedly present in the corpora they are trained on, implicitly creating harm with biased representations." (Nangia et al., 2020, p. 1) "Language that stereotypes already disadvantaged groups propagates false beliefs about these groups and entrenches inequalities" (Nangia et al., 2020, p. 1)
5	Lucy and Bamman, 2021	"Our work focuses on representational harms in generated narratives, especially the reproduction of gender stereotypes found in film, television, and books." (Lucy and Bamman, 2021, p. 48)

Index	Authors	Normative Motivation
6	Jentzsch and Turan, 2022	"As they become capable of grasping complex contextual information, harmful biases are likely increasingly intertwined with those models." (Jentzsch and Turan, 2022, p. 184)"In this context, it has widely been shown that embeddings tend to reflect human biases and stereotypes (Caliskan et al., 2017; Jentzsch et al., 2019) and that unintended imbalances in text-embeddings can lead to misbehaviour of systems (Bolukbasi et al., 2016). This could not only reinforce existing imbalance in the film industry but also lead to direct financial and social harm, e.g. if a movie is less frequently recommended by an automatic recommendation system." (Jentzsch and Turan, 2022, p. 185) "In this concrete context, we consider it harmful if a classifier that is trained to distinguish positive and negative movie reviews prefers performers and film characters of one gender over another." (Jentzsch and Turan, 2022, p. 185)
7	Kwon and Mihindukulasooriya, 2022	"A cognitive bias, stereotyping, is defined as the assumption of some characteristics are applied to communities on the basis of their nationality, ethnicity, gender, religion, etc (Schneider, 2005). Relatedly, Fairness ("zero-bias"), in the context of NLP and machine learning is defined as preventing harmful, discriminatory decisions according to such unwanted, stereotypical characteristics (Garrido-Muñoz et al., 2021)." (Kwon and Mihindukulasooriya, 2022, p. 74)
8	Steinborn et al., 2022	"In this work we investigate binary gender stereotypes as a representational harm across languages, to use the terminology of Blodgett et al. (2020)" (Steinborn et al., 2022, p. 925)"However, commonly used PLMs such as BERT have been shown to encapsulate social biases, including those relating to gender and race (Kurita et al., 2019; Nadeem et al., 2021; Nangia et al., 2020). The general consensus is that these biases are learned from the statistical distributional cooccurrence of words relating to a group (such as terms relating to men or women) with a context in which that group is often mentioned in corpora" (Steinborn et al., 2022, p. 921) "The importance of developing AI systems that are mindful of different societal groups, such as people of different genders, is a topic much discussed in the area of fairness research in NLP (Blodgett et al., 2020) ... ours is the first study to attempt to create a truly multilingual approach to study gender bias in language models." (Steinborn et al., 2022, p. 921-922)
9	Sotnikova et al., 2021	"In analyzing our results, we start from the normative position that identical model behavior across target categories is insufficient, despite being a prevalent goal in past literature (Blodgett et al., 2020, inter alia)." (Sotnikova et al., 2021, p. 4056) "First, because if a person of some category sees an offensive stereotype about themselves in a downstream system, they are harmed even if the same output is generated for other categories. Second, because social hierarchies enable members of some groups to more easily subjugate members of other groups, the same oppressive stereotypes are more likely to harm people in categories lower on the social hierarchy than those higher." (Sotnikova et al., 2021, p. 4056-4057)
10	Kaneko et al., 2022	"Unfortunately, it was reported that MLMs also learn discriminative biases regarding attributes such as gender and race." (Kaneko et al., 2022, p. 2740)"To realise the diverse and inclusive social and cultural impact of AI, we believe it is important to establish tools for detecting and mitigating unfair social biases in MLMs, not only for English but for all languages." (Kaneko et al., 2022, p. 2740)

Index	Authors	Normative Motivation
11	Wolfe and Caliskan, 2021	Not normative, but psychological motivaiton: "Recent research indicates that state-of-the-art AI systems mirror such biased and unequal human perceptions." (Wolfe and Caliskan, 2021, p. 1)
12	Touileb et al., 2022	"In our case, we see this as reflected in large pre-trained language models and how they can contain skewed gendered representations that can be systematically unfair if this bias is not uncovered and properly taken into account in downstream applications" (Touileb et al., 2022, p. 201)"It is a reality that most Norwegian nurses are females. Having a model reflecting this reality might not be problematic per se, but using this disparity to for example systematically reject male applicants to a nurse position is a very harmful effect." (Touileb et al., 2022, p. 200) ("Since LMs are now the backbone of most NLP model architectures, the extent to which they reflect, amplify, and spread the biases existing in the input data is very important for the further development of such models, and the understanding of their possible harmful outcomes." (Touileb et al., 2022, p. 201However, we explore this from the perspective of a descriptive assessment: Instead of expecting the system to treat genders equally, we compare how these gender–occupation representations reflect the actual and current Norwegian demographics. This will in no way reduce the representational harms of stereotypical female and male occupations, that could both be propagated and exaggerated by downstream tasks, but would rather shed light on which occupations are falsely represented by such models. Moreover, our work will provide knowledge about the biases contained in these models that may be important to take into account when choosing a model for a specific application.
13	Nadeem et al., 2019	"A stereotype is an over-generalized belief about a particular group of people, e.g., Asians are good at math or Asians are bad drivers. Such beliefs (biases) are known to hurt target groups." (Nadeem et al., 2020, p. 1) "In order to assess adverse effects of these models, it is important to quantify the bias captured in them." (Nadeem et al., 2020, p. 1) (
14	Silva et al., 2021	"there is little consideration for the societal biases captured within these model risking perpetuation of racial, gender, and other harmful biases when these models are deployed at scale" (Silva et al., 2021, p. 1) "Without appropriately considering inherent biases, development on top of pre-trained transformers risks exacerbating and propagating racial, gender, and other biases writ large." (Silva et al., 2021, p. 1)"With no additional information, an equitable system would exhibit no preference for female over male, or African-American over European-American names; however, our results indicate that there is often a statistically significant preference ($p < 0.0001$) for associating female and African-American identifiers with being more "emotional."" (Silva et al., 2021, p. 1) (pdf)
15	Alnegheimish et al., 2022	Not Provided

Index	Authors	Normative Motivation
16	Barikeri et al., 2021	"Text representation models are prone to exhibit a range of societal biases, reflecting the noncontrolled and biased nature of the underlying pretraining data, which consequently leads to severe ethical issues and even bias amplification" (Barikeri et al., 2021, p. 1941)"Pretrained language models and their corresponding contextualized representation spaces (Peters et al., 2018; Devlin et al., 2019) have recently been shown to encode and amplify a range of stereotypical human biases (e.g., gender or racial biases)" (Barikeri et al., 2021, p. 1941) "Having models that capture or even amplify human biases brings about further ethical challenges to the society (Henderson et al., 2018), since stereotyping minoritized groups is a representational harm that perpetuates societal inequalities and unfairness (Blodgett et al., 2020). Human biases are in all likelihood especially harmful if encoded in conversational AI systems, like the recent DialoGPT model (Zhang et al., 2020), which directly interact with humans, possibly even taking part in intimate and personal conversations (Utami et al., 2017)." (Barikeri et al., 2021, p. 1941)
17	Shen et al., 2023	"However, pretrained LMs are well-known for exhibiting unintended social biases involving race, gender, or religion (Liang, Wu, Morency, & Salakhutdinov, 2021; Lu, Mardziel, Wu, Amancharla, & Datta, 2020; Sheng, Chang, Natarajan, & Peng, 2019). These biases result from unfair allocation of resources (e.g., policing, hospital services, or job availability) (Hutchinson et al., 2020; Zhang, Lu, Abdalla, McDermott, & Ghassemi, 2020), stereotyping that propagates negative generalizations about particular social groups (Nadeem, Bethke, & Reddy, 2021), text that misrepresents the distribution of different social groups in the population (Liang et al., 2021), or language that is denigrating to particular social groups (Guo & Caliskan, 2021). Moreover, these biases may also be exacerbated by biases in data used for domain-specific LM fine-tuning for downstream tasks (Jin et al., 2021; Nadeem et al., 2021)." (Shen et al., 2023, p. 2) "Performance disparities (with NDCG metric) of Collaborative Filtering (CF) algorithms in the recommendation of movies and music have been observed (Ekstrand et al., 2018), revealing unfairness with regard to users' age and gender." (Shen et al., 2023, p. 2)
18	Bartl et al., 2020	"The biases present in the large masses of language data that are used to train Natural Language Processing (NLP) models naturally leak into NLP systems. These systematic biases can have real-life consequences when such systems are e.g. used to rank the resumes of possible candidates for a vacancy in order to aid the hiring decision (Bolukbasi et al., 2016). If, for example, a model does not associate female terms with engineering professions, because these do not often co-occur in the same context in the training corpus, then the system is likely to rank male candidates for an engineering position higher than equally qualified female candidates." (Bartl et al., p. 1) "The present work contributes to promoting fairness in NLP by exploring methods to measure and mitigate gender bias in BERT (Devlin et al., 2018), a contextualized word embedding model. Its widespread and quick adoption by the research community as the backbone for a variety of tasks calls for an assessment of possible biases encoded in it." (Bartl et al., p. 1)

Index	Authors	Normative Motivation
19	de Vassimon Manela et al., 2021	"This training is only performed once, with users downloading and fine-tuning such language models to their specific task. In doing so, we are trusting large tech companies to train the base model responsibly since we have no control over this. This seems inherently undemocratic." (Vassimon Manela et al., 2021, p. 2232)

Table B.3: this is the last table

Bibliography for Appendices

[A1] Byeonghu Na, Hyemi Kim, Kyungwoo Song, Weonyoung Joo, Yoon-Yeong Kim, and Il-Chul Moon. Deep generative positive-unlabeled learning under selection bias. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1155–1164, 2020.

[A2] Lucy E Morgan, Barry L Nelson, Andrew C Titman, and David J Worthington. Detecting bias due to input modelling in computer simulation. *European Journal of Operational Research*, 279(3):869–881, 2019.

[A3] Joni Salminen, Soon-Gyo Jung, and Bernard J Jansen. Detecting demographic bias in automatically generated personas. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.

[A4] Manjira Sinha and Tirthankar Dasgupta. Determining subjective bias in text through linguistically informed transformer based multi-task network. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3418–3422, 2021.

[A5] Dharmesh Patel, Nilesh Chothani, and Khyati Mistry. Discrimination of inrush, internal, and external fault in power transformer using phasor angle comparison and biased differential principle. *Electric Power Components and Systems*, 46(7):788–801, 2018.

[A6] Joshua I Breier, Lincoln C Gray, Patricia Klaas, Jack M Fletcher, and Barbara Foorman. Dissociation of sensitivity and response bias in children with attention deficit/hyperactivity disorder during central auditory masking. *Neuropsychology*, 16(1):28, 2002.

[A7] Martin Pelikan and Mark W Hauschild. Distance-based bias in model-directed optimization of additively decomposable problems. In *Proceedings of the 14th annual conference on Genetic and evolutionary computation*, pages 273–280, 2012.

[A8] Xiaokang Zhou, Yiyong Hu, Jiayi Wu, Wei Liang, Jianhua Ma, and Qun Jin. Distribution bias aware collaborative generative adversarial network for imbalanced deep learning in industrial iot. *IEEE Transactions on Industrial Informatics*, 19(1):570–580, 2022.

[A9] Anastassia Loukina, Keelan Evanini, Matthew Mulholland, Ian Blood, and Klaus Zechner. Do face masks introduce bias in speech technologies? the case of automated scoring of speaking proficiency. *arXiv preprint arXiv:2008.07520*, 2020.

[A10] Vered Shwartz and Yejin Choi. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, 2020.

[A11] Caroline P Martin, Virginia Peisch, Erin K Shoulberg, Nina Kaiser, and Betsy Hoza. Does a social self-perceptual bias mask internalizing symptoms in children with attention-deficit/hyperactivity disorder? *Journal of Child Psychology and Psychiatry*, 60(6):630–637, 2019.

[A12] Sachiko Kinoshita and Dennis Norris. Does the familiarity bias hypothesis explain why there is no masked priming for "no" decisions? *Memory & Cognition*, 39:319–334, 2011.

[A13] Yushun Dong, Ninghao Liu, Brian Jalaian, and Jundong Li. Edits: Modeling and mitigating data bias for graph neural networks. In *Proceedings of the ACM Web Conference 2022*, pages 1259–1269, 2022.

[A14] Davide Signori, Giacomo Bellani, Serena Calcinati, Alice Grassi, Nicolò Patroniti, and Giuseppe Foti. Effect of face mask design and bias flow on rebreathing during noninvasive ventilation. *Respiratory care*, 64(7):793–800, 2019.

[A15] Sechan Kim, Gyuhyun Choi, Heeyeop Chae, and Nae-Eung Lee. Effects of bias pulsing on etching of sio2 pattern in capacitively-coupled plasmas for nano-scale patterning of multi-level hard masks. *Journal of Nanoscience and Nanotechnology*, 16(5):5143–5149, 2016.

[A16] Yu Ya Chang, Yuan-Hsun Wu, Chiang-Lin Shih, Jengping Lin, Francis Kan, and Jimmy Lin. Effects of mask bias on the mask error enhancement factor (meef) for low k1 lithography process. In *Photomask and Next-Generation Lithography Mask Technology XII*, volume 5853, pages 757–766. SPIE, 2005.

[A17] Doris Kang, Stewart A Robertson, Michael T Reilly, and Edward K Pavelchek. Effects of mask bias on the mask error enhancement factor (meef) of contact holes. In *Optical Microlithography XIV*, volume 4346, pages 858–868. SPIE, 2001.

[A18] William Merrill, Vivek Ramanujan, Yoav Goldberg, Roy Schwartz, and Noah Smith. Effects of parameter norm growth during transformer training: Inductive bias from gradient descent. *arXiv preprint arXiv:2010.09697*, 2020.

[A19] Anthony J Viera and Shrikant I Bangdiwala. Eliminating bias in randomized controlled trials: importance of allocation concealment and masking. *FAMILY MEDICINE-KANSAS CITY-*, 39(2):132, 2007.

[A20] Young-Doo Jeon, Sungho Jun, Jae-Hyun Kang, Sang-Uk Lee, Jeahee Kim, and Keeho Kim. Enhanced hole shape of flash devices in arf lithography by eliptical mask bias technique. In *Metrology, Inspection, and Process Control for Microlithography XXI*, volume 6518, pages 1284–1291. SPIE, 2007.

[A21] Hongbo Deng, Irwin King, and Michael R Lyu. Entropy-biased models for query representation on the click graph. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 339–346, 2009.

[A22] Praveen Chandar and Ben Carterette. Estimating clickthrough bias in the cascade model. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1587–1590, 2018.

[A23] Hideo Kobinata, Yasuhisa Yamada, Takao Tamura, Kiyoshi Fujii, Mitsuru Narihiro, and Yukinori Ochiai. Estimation of optimum electron-beam projection lithography mask biases taking coulomb beam blur into consideration. *Japanese journal of applied physics*, 42(6S):3816, 2003.

[A24] Alexander Lajn, Haiko Rolff, and Richard Wistrom. Etch bias inversion during euv mask arc etch. In *Photomask Japan 2017: XXIV Symposium on Photomask and Next-Generation Lithography Mask Technology*, volume 10454, pages 123–126. SPIE, 2017.

[A25] Ingrid Scharlau. Evidence against response bias in temporal order tasks with attention manipulation by masked primes. *Psychological research*, 68:224–236, 2004.

[A26] Jennifer C White and Ryan Cotterell. Examining the inductive bias of neural language models with artificial languages. *arXiv preprint arXiv:2106.01044*, 2021.

[A27] Ghazi Felhi, Joseph Le Roux, and Djamé Seddah. Exploiting inductive bias in transformers for unsupervised disentanglement of syntax and semantics with vaes. *arXiv preprint arXiv:2205.05943*, 2022.

[A28] Timo Spinde, Jan-David Krieger, Terry Ruas, Jelena Mitrović, Franz Götz-Hahn, Akiko Aizawa, and Bela Gipp. Exploiting transformer-based multitask learning for the detection of media bias in news articles. In *Information for a Better World: Shaping the Global Future: 17th International Conference, iConference 2022, Virtual Event, February 28–March 4, 2022, Proceedings, Part I*, pages 225–235. Springer, 2022.

[A29] Lisa Schneider, Palak Dave, Lubaina Arsiwala-Scheppach, Falk Schwendicke, and Joachim Krois. Exploring bias in f-score computation methods of multi-class

segmentation models. In *2021 The 5th International Conference on Video and Image Processing*, pages 76–84, 2021.

[A30] Simon Dobnik, Mehdi Ghanimifard, and John Kelleher. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 1–11, 2018.

[A31] Yaron Fairstein, Elad Haramaty, Arnon Lazerson, and Liane Lewin-Eytan. External evaluation of ranking models under extreme position-bias. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 252–261, 2022.

[A32] Krystyna M Cowan, Dave Shutler, Thomas B Herman, and Donald T Stewart. Extreme male-biased infections of masked shrews by bladder nematodes. *Journal of Mammalogy*, 88(6):1539–1543, 2007.

[A33] Lawrence S Melvin, Yudhishthir Kandel, Qiliang Yan, Artak Isoyan, and Weimin Gao. Extreme ultraviolet mask multilayer material variation impact on horizontal to vertical pattern bias. In *Extreme Ultraviolet (EUV) Lithography IX*, volume 10583, pages 416–425. SPIE, 2018.

[A34] Edward Raff, Jared Sylvester, and Steven Mills. Fair forests: Regularized tree induction to minimize model bias. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 243–250, 2018.

[A35] Tatyana Ivanovska, René Laqua, Lei Wang, Henry Völzke, and Katrin Hegenscheid. Fast implementations of the levelset segmentation method with bias field correction in mr images: full domain and mask-based versions. In *Pattern Recognition and Image Analysis: 6th Iberian Conference, IbPRIA 2013, Funchal, Madeira, Portugal, June 5-7, 2013. Proceedings 6*, pages 674–681. Springer, 2013.

[A36] Glen E Bodner, Jeremy Johnson, and Michael EJ Masson. Fluency can bias masked priming of binary judgments: Evidence from an all-nonword task. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 69(2):200, 2015.

[A37] Zheng Cui, Philip D Prewett, and John G Watson. Focused ion beam biased repair of conventional and phase shift masks. *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena*, 14(6):3942–3946, 1996.

[A38] Aldo Lipani, Ben Carterette, and Emine Yilmaz. From a user model for query sessions to session rank biased precision (srbp). In *Proceedings of the 2019*

ACM SIGIR International Conference on Theory of Information Retrieval, pages 109–116, 2019.

[A39] Mingrui Lao, Yanming Guo, Yu Liu, Wei Chen, Nan Pu, and Michael S Lew. From superficial to deep: Language bias driven curriculum learning for visual question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3370–3379, 2021.

[A40] Joshua M Carlson, Jiook Cha, and Lilianne R Mujica-Parodi. Functional and structural amygdala–anterior cingulate connectivity correlates with attentional bias to masked fearful faces. *Cortex*, 49(9):2595–2600, 2013.

[A41] Brian Hentschel, Peter J Haas, and Yuanyuan Tian. General temporally biased sampling schemes for online model management. *ACM Transactions on Database Systems (TODS)*, 44(4):1–45, 2019.

[A42] Mina Rezaei, Tomoki Uemura, Janne Näppi, Hiroyuki Yoshida, Christoph Lippert, and Christoph Meinel. Generative synthetic adversarial network for internal bias correction and handling class imbalance problem in medical image diagnosis. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 82–89. SPIE, 2020.

[A43] Chathika Gunaratne and Robert Patton. Genetic programming for understanding cognitive biases that generate polarization in social networks. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 546–549, 2022.

[A44] Yinqiong Cai, Jiafeng Guo, Yixing Fan, Qingyao Ai, Ruqing Zhang, and Xueqi Cheng. Hard negatives or false negatives: Correcting pooling bias in training neural ranking models. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 118–127, 2022.

[A45] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861, 2020.

[A46] Olawale Onabola, Zhuang Ma, Yang Xie, Benjamin Akera, Abdulrahman Ibraheem, Jia Xue, Dianbo Liu, and Yoshua Bengio. Hbert+ biascorp–fighting racism on the web. *arXiv preprint arXiv:2104.02242*, 2021.

[A47] Hannah Mieczkowski, Sunny Xun Liu, Jeffrey Hancock, and Byron Reeves. Helping not hurting: Applying the stereotype content model and bias map to social robotics. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 222–229. IEEE, 2019.

[A48] Inchan Ju, Yunyi Gong, and John D Cressler. Highly linear high-power 802.11 ac/ax wlan sige hbt power amplifiers with a compact 2nd-harmonic-shorted four-way transformer and a thermally compensating dynamic bias circuit. *IEEE Journal of Solid-State Circuits*, 55(9):2356–2370, 2020.

[A49] Filipe N Ribeiro, Fabrício Benevenuto, and Emilio Zagheni. How biased is the population of facebook users? comparing the demographics of facebook users with census data to generate correction factors. In *12th ACM conference on web science*, pages 325–334, 2020.

[A50] Jacinto Mata Vázquez, Victoria Pachón Álvarez, Chaimae Tayebi Taybi, and PP Sánchez. I2c at iberlef-2022 detests task: Detection of racist stereotypes in spanish comments using underbagging and transformers. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS. org*, 2022.

[A51] Yariv Z Levy, Dino Levy, Jerrold S Meyer, and Hava T Siegelmann. Identification and control of intrinsic bias in a multiscale computational model of drug addiction. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 2389–2393, 2010.

[A52] Hiroki Tabuchi, Y Shichijo, N Oka, N Takenaka, and K Iguchi. Illumination condition and mask bias for 0.15-um pattern with krf and arf lithography. In *Optical Microlithography XII*, volume 3679, pages 860–871. SPIE, 1999.

[A53] Ivan Lalovic, Armen Kroyan, Paolo Zambon, Christopher D Silsby, and Nigel R Farrar. Illumination spectral width impacts on mask error enhancement factor and iso-dense bias in 0.6-na krf imaging. In *21st Annual BACUS Symposium on Photomask Technology*, volume 4562, pages 992–999. SPIE, 2002.

[A54] Jing Duan, Jie Duan, Yanhua Wang, and Xuefeng Wan. Image steganography based on least bias generative adversarial network. In *International Conference on Cloud Computing, Performance Computing, and Deep Learning (CCPCDL 2022)*, volume 12287, pages 345–350. SPIE, 2022.

[A55] Yunyun Yang, Wenjing Jia, and Dongcai Tian. Improved active contour model for multi-phase mr image segmentation and bias field correction. In *Proceedings of the 3rd International Conference on Cryptography, Security and Privacy*, pages 242–246, 2019.

[A56] Nuntaporn Tuangpermsub, Panuthat Boonpranuk, Wutthichai Polwisate, and Prakasit Kayasith. Improvement of esophageal speech by adaptive line enhancement with bias model. In *Proceedings of the 2nd International Convention on Rehabilitation Engineering & Assistive Technology*, pages 74–77, 2008.

[A57] Thorsten Albrecht and Uwe Mattler. Individual differences in metacontrast masking regarding sensitivity and response bias. *Consciousness and cognition*, 21(3):1222–1231, 2012.

[A58] Jiazhi Li and Wael Abd-Almageed. Information-theoretic bias assessment of learned representations of pretrained face recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.

[A59] Luiz H Mormille, Clifford Broni-Bediako, and Masayasu Atsumi. Introducing inductive bias on vision transformers through gram matrix similarity based regularization. *Artificial Life and Robotics*, pages 1–11, 2023.

[A60] Robert D Hawkins, Takateru Yamakoshi, Thomas L Griffiths, and Adele E Goldberg. Investigating representations of verb bias in neural language models. *arXiv preprint arXiv:2010.02375*, 2020.

[A61] Ji-Eun Lee, Hye-Young Kang, Dong-Soo Shin, Hee-Jun Jeong, Ilsin An, Chang-Nam Ahn, and Hye-Keun Oh. Investigation of optimum biasing and undercut for single trench alternating phase shift mask in 193 nm lithography. *Japanese journal of applied physics*, 45(11R):8920, 2006.

[A62] Yuning Zhang and Hui Ning. Irony and stereotype spreaders detection using bert-large and autogulon. In *CLEF*, pages 1613–0073, 2022.

[A63] Amit Das, Nilanjana Raychawdhary, Gerry Dozier, and Cheryl D Seals. Irony and stereotype spreading author profiling on twitter using machine learning: A bert-tfidf based approach. 2022.

[A64] Chao Yang, Ping Li, Yumei Wen, Aichao Yang, Decai Wang, Feng Zhang, and Jiajia Zhang. Large converse magnetoelectric properties without bias in composite of rosen-type piezoelectric transformer and magnetization-graded ferromagnetic material. *IEEE Transactions on Magnetics*, 51(11):1–4, 2015.

[A65] Haoyu Yang, Jing Su, Yi Zou, Bei Yu, and Evangeline FY Young. Layout hotspot detection with feature tensor generation and deep biased learning. In *Proceedings of the 54th Annual Design Automation Conference 2017*, pages 1–6, 2017.

[A66] Giulio Fabbian, Julien Carron, Antony Lewis, and Margherita Lembo. Lensed cmb power spectrum biases from masking extragalactic sources. *Physical Review D*, 103(4):043535, 2021.

[A67] Nishtha Jain, Maja Popović, Declan Groves, and Lucia Specia. Leveraging pretrained language models for gender debiasing. 2022.

[A68] Christopher Rytting and David Wingate. Leveraging the inductive bias of large language models for abstract textual reasoning. *Advances in Neural Information Processing Systems*, 34:17111–17122, 2021.

[A69] Tim Cole and Laura Leets. Linguistic masking devices and intergroup behavior: Further evidence of an intergroup linguistic bias. *Journal of Language and Social Psychology*, 17(3):348–371, 1998.

[A70] Ye Yang, Lang Xie, Zhimin He, Qi Li, Vu Nguyen, Barry Boehm, and Ricardo Valerdi. Local bias and its impacts on the performance of parametric estimation models. In *Proceedings of the 7th International Conference on Predictive Models in Software Engineering*, pages 1–10, 2011.

[A71] Takashi Yamauchi. Labeling bias and categorical induction: generative aspects of category information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3):538, 2005.

[A72] Andrew T Smith and Charles Elkan. Making generative classifiers robust to selection bias. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 657–666, 2007.

[A73] Md Abdullah Al Baki, Mohammad Vatanparast, and Yonggyun Kim. Male-biased adult production of the striped fruit fly, zeugodacus scutellata, by feeding dsrna specific to transformer-2. *Insects*, 11(4):211, 2020.

[A74] Carla H Van Gils, Johannes DM Otten, André LM Verbeek, and Jan HCL Hendriks. Mammographic breast density and risk of breast cancer: masking bias or causality? *European journal of epidemiology*, 14:315–320, 1998.

[A75] T Ema, H Yamashita, K Nakajima, and H Nozue. Mask bias effects in eb cell projection lithography [3096-39]. In *PROCEEDINGS-SPIE THE INTERNATIONAL SOCIETY FOR OPTICAL ENGINEERING*, pages 275–285. SPIE INTERNATIONAL SOCIETY FOR OPTICAL, 1997.

[A76] Kimitoshi Takahashi, Hiroyuki Kanata, and Yasuo Nara. Mask bias requirement for 0.13 μm e-beam block exposure lithography. *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena*, 16(6):3279–3283, 1998.

[A77] Grzegorz Bilo, Marek Klocek, Katarzyna Stolarz, and Kalina Kawecka-Jaszcz. Masked hypertension. a clinical state or measurement bias? *Arterial Hypertension*, 9(5):385–396, 2005.

[A78] Friederike Schlaghecken and Martin Eimer. Masked prime stimuli can bias "free" choices between response alternatives. *Psychonomic Bulletin & Review*, 11(3):463–468, 2004.

[A79] Javier Sánchez-Junquera, Paolo Rosso, Manuel Montes, Berta Chulvi, et al. Masking and bert-based models for stereotype identication. *Procesamiento del Lenguaje Natural*, 67:83–94, 2021.

[A80] A Gupta, CM Davison, and Michael A McIsaac. Masking in reports of "most serious" events: bias in estimators of sports injury incidence in canadian children. *Health Promotion and Chronic Disease Prevention in Canada: Research, Policy and Practice*, 36(8):143, 2016.

[A81] Brendan Spillane, Séamus Lawless, and Vincent Wade. Measuring bias in news websites, towards a model for personalization. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 387–388, 2017.

[A82] Craig Steinmaus, Allan H Smith, Rachael M Jones, and Martyn T Smith. Meta-analysis of benzene exposure and non-hodgkin lymphoma: biases could mask an important association. *Occupational and environmental medicine*, 65(6):371–378, 2008.

[A83] Christopher G. Harris. Methods to evaluate temporal cognitive biases in machine learning prediction models. In *Companion Proceedings of the Web Conference 2020*, pages 572–575, 2020.

[A84] Michael Kunz, Zhou Yu, and Achilleas S Frangakis. M-free: Mask-independent scoring of the reference bias. *Journal of structural biology*, 192(2):307–311, 2015.

[A85] Matthew Almeida, Yong Zhuang, Wei Ding, Scott E Crouter, and Ping Chen. Mitigating class-boundary label uncertainty to reduce both model bias and variance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(2):1–18, 2021.

[A86] Sebastian Hofstätter, Aldo Lipani, Sophia Althammer, Markus Zlabinger, and Allan Hanbury. Mitigating the position bias of transformer models in passage re-ranking. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43*, pages 238–253. Springer, 2021.

[A87] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1791–1800, 2021.

[A88] Mark Hoogendoorn, S Waqar Jaffry, Peter-Paul van Maanen, and Jan Treur. Modeling and validation of biased human trust. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 256–263. IEEE, 2011.

[A89] Jun Yuan, Kurt Shultz, Carl Pixley, Hillel Miller, and Adnan Aziz. Modeling design constraints and biasing in simulation using bdds. In *1999 IEEE/ACM International Conference on Computer-Aided Design. Digest of Technical Papers (Cat. No. 99CH37051)*, pages 584–589. IEEE, 1999.

[A90] Masaki Stanley Fujimoto, Paul M Bodily, Cole A Lyman, Andrew J Jacobsen, Quinn Snell, and Mark J Clement. Modeling global and local codon bias with deep language models. In *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 151–156. IEEE, 2017.

[A91] Rahul Pandey, Carlos Castillo, and Hemant Purohit. Modeling human annotation errors to design bias-aware systems for social stream processing. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 374–377, 2019.

[A92] Matteo Ruffini, Vito Bellini, Alexander Buchholz, Giuseppe Di Benedetto, and Yannik Stein. Modeling position bias ranking for streaming media services. In *Companion Proceedings of the Web Conference 2022*, pages 72–76, 2022.

[A93] Vaclav Petricek, Ingemar J Cox, Hui Han, Isaac G Councill, and C Lee Giles. Modeling the author bias between two on-line computer science citation databases. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 1062–1063, 2005.

[A94] Mark D Smucker and Charles LA Clarke. Modeling user variance in time-biased gain. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, pages 1–10, 2012.

[A95] Pooja N Upadhayaya and Vijay H Makwana. Modelling & simulation of transformer biased differential protection scheme in laboratory environment. In *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, pages 68–73. IEEE, 2017.

[A96] Fazlourrahman Balouchzahi, Oxana Vitman, Hosahalli Lakshmaiah Shashirekha, Grigori Sidorov, and Alexander Gelbukh. Mucic at comma@ icon: Multilingual gender biased and communal language identification using n-grams and multilingual sentence encoders. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 58–63, 2021.

[A97] Haizhi Yang, Tengyun Wang, Xiaoli Tang, Qianyu Li, Yueyue Shi, Siyu Jiang, Han Yu, and Hengjie Song. Multi-task learning for bias-free joint ctr prediction and market price modeling in online advertising. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2291–2300, 2021.

[A98] Eran S Auday, Bradley C Taber-Thomas, and Koraly E Pérez-Edgar. Neural correlates of attention bias to masked facial threat cues: Examining children at-risk for social anxiety disorder. *NeuroImage: Clinical*, 19:202–212, 2018.

[A99] Guipeng Xv, Chen Lin, Hui Li, Jinsong Su, Weiyao Ye, and Yewang Chen. Neutralizing popularity bias in recommendation models. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2623–2628, 2022.

[A100] Varun Magotra, Ebrahim Hirani, Vedant Mehta, and Surekha Dholay. News bias detection using transformers. In *Communication and Intelligent Systems: Proceedings of ICCIS 2021*, pages 319–326. Springer, 2022.

[A101] Jonathan Weil and Michael Cassara. Occult sepsis masked by trauma—exploration of cognitive biases through simulation with emergency medicine residents. *MedEdPORTAL*, 16:11023, 2020.

[A102] Aldo Lipani. On biases in information retrieval models and evaluation. In *ACM SIGIR Forum*, volume 52, pages 172–173. ACM New York, NY, USA, 2019.

[A103] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

[A104] Huayang Li, Lemao Liu, Guoping Huang, and Shuming Shi. On the branching bias of syntax extracted from pre-trained language models. *arXiv preprint arXiv:2010.02448*, 2020.

[A105] Katja Schwarz, Yiyi Liao, and Andreas Geiger. On the frequency bias of generative models. *Advances in Neural Information Processing Systems*, 34:18126–18136, 2021.

[A106] Tianyi Zhang and Tatsunori Hashimoto. On the inductive bias of masked language modeling: From statistical to syntactic dependencies. *arXiv preprint arXiv:2104.05694*, 2021.

[A107] Sankaranarayanan Ananthakrishnan, Stavros Tsakalidis, Rohit Prasad, and Premkumar Natarajan. On-line language model biasing for multi-pass automatic speech recognition. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[A108] Cassondra Neau and Kaushik Roy. Optimal body bias selection for leakage improvement and process compensation over different technology generations. In *Proceedings of the 2003 international symposium on Low power electronics and design*, pages 116–121, 2003.

[A109] Sung-il Pae and Michael C Loui. Optimal random number generation from a biased coin. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1079–1088, 2005.

[A110] Piyush Vyas, Anastasia Kuznetsova, and Donald S Williamson. Optimally encoding inductive biases into the transformer improves end-to-end speech translation. In *Interspeech*, pages 2287–2291, 2021.

[A111] Janghwan Lee and Jungwook Choi. Optimizing exponent bias for sub-8bit floating-point inference of fine-tuned transformers. In *2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pages 98–101. IEEE, 2022.

[A112] Young-Min Kang, Ilsin An, Dong-Soo Shin, and Hye-Keun Oh. Optimum biasing for 45 nm node chromeless and attenuated phase shift mask. *Japanese journal of applied physics*, 47(9R):7448, 2008.

[A113] Gyuhyun Choi, Sechan Kim, Haegyu Jang, Heeyeop Chae, and Nae-Eung Lee. Patterning of si3n4 layer in pulse-biased capacitively-coupled plasmas for multi-level hard mask structures. *Journal of Nanoscience and Nanotechnology*, 16(11):11817–11822, 2016.

[A114] Kimiyoshi Deguchi, Yoshio Kawai, Hiroyuki Kochiya, Yukihiko Ushiyama, and Masatoshi Oda. Patterning yield of sub-100-nm holes limited by fluctuation of exposure and development reactions in synchrotron radiation lithography using biased mask patterns. *Japanese Journal of Applied Physics*, 39(12S):6936, 2000.

[A115] NH Hashim, NH Halim, SNM Arshad, MH Hussain, SRA Rahim, and AA Suleiman. Performance of restricted fault and bias differential protection against earth fault on a transformer. In *Journal of Physics: Conference Series*, volume 2312, page 012005. IOP Publishing, 2022.

[A116] Michael Moret, Francesca Grisoni, Paul Katzberger, and Gisbert Schneider. Perplexity-based molecule ranking and bias estimation of chemical language models. *Journal of chemical information and modeling*, 62(5):1199–1206, 2022.

[A117] Emilio Cruciani, Hlafo Alfie Mimun, Matteo Quattropani, and Sara Rizzo. Phase transitions of the k-majority dynamics in a biased communication model. In *Proceedings of the 22nd International Conference on Distributed Computing and Networking*, pages 146–155, 2021.

[A118] Baojin Huang, Zhongyuan Wang, Guangcheng Wang, Kui Jiang, Zhen Han, Tao Lu, and Chao Liang. Plface: Progressive learning for face recognition with mask bias. *Pattern Recognition*, 135:109142, 2023.

[A119] Yasaman Haghpanah and Marie Desjardins. Prep: a probabilistic reputation model for biased societies. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 315–322. Citeseer, 2012.

[A120] Hongbo Deng, Jiawei Han, Bo Zhao, Yintao Yu, and Cindy Xide Lin. Probabilistic topic models with biased propagation on heterogeneous information networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1271–1279, 2011.

[A121] Michael T Reilly, Karen Kvam, and Jentry Willie. Process window overlap for posts and lines and spaces: optimization by resist type, optical settings, and mask bias. In *Lithography for Semiconductor Manufacturing*, volume 3741, pages 40–45. SPIE, 1999.

[A122] Yifan Xu and Hui Ning. Profiling irony and stereotype spreaders on twitter with bert. In *CLEF*, pages 1613–0073, 2022.

[A123] Xinting Huang. Profiling irony and stereotype spreaders with language models and bayes' theorem. In *CLEF*, pages 1613–0073, 2022.

[A124] Darren Hurley-Smith and Julio Hernandez-Castro. Quantum leap and crash: Searching and finding bias in quantum random number generators. *ACM Transactions on Privacy and Security (TOPS)*, 23(3):1–25, 2020.

[A125] Yu Huang, Ziyang Liu, and Yi Chen. Query biased snippet generation in xml search. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 315–326, 2008.

[A126] Serge Fehr and Christian Schaffner. Randomness extraction via δ-biased masking in the presence of a quantum attacker. In *Theory of Cryptography: Fifth Theory of Cryptography Conference, TCC 2008, New York, USA, March 19-21, 2008. Proceedings 5*, pages 465–481. Springer, 2008.

[A127] Maria Chikina, Wesley Pegden, and Benjamin Recht. Re-analysis on the statistical sampling biases of a mask promotion trial in bangladesh: a statistical replication. *Trials*, 23(1):1–5, 2022.

[A128] Jean-Francois Rajotte, Sumit Mukherjee, Caleb Robinson, Anthony Ortiz, Christopher West, Juan M Lavista Ferres, and Raymond T Ng. Reducing bias and increasing utility by federated generative modeling of medical images using a centralized adversary. In *Proceedings of the Conference on Information Technology for Social Good*, pages 79–84, 2021.

[A129] Alexandra L Jones, Larry Di Girolamo, and Guangyu Zhao. Reducing the resolution bias in cloud fraction from satellite derived clear-conservative cloud masks. *Journal of Geophysical Research: Atmospheres*, 117(D12), 2012.

[A130] Daizong Liu, Xiaoye Qu, and Wei Hu. Reducing the vision and language bias for temporal sentence grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4092–4101, 2022.

[A131] Mikael P Backlund, Amir Arbabi, Petar N Petrov, Ehsan Arbabi, Saumya Saurabh, Andrei Faraon, and WE Moerner. Removing orientation-induced localization biases in single-molecule microscopy using a broadband metasurface mask. *Nature photonics*, 10(7):459–462, 2016.

[A132] Glen E Bodner, Michael EJ Masson, and Norann T Richard. Repetition proportion biases masked priming of lexical decisions. *Memory & Cognition*, 34(6):1298–1311, 2006.

[A133] Omid Nejati Manzari, Hossein Kashiani, Hojat Asgarian Dehkordi, and Shahriar B Shokouhi. Robust transformer with locality inductive bias and feature normalization. *Engineering Science and Technology, an International Journal*, 38:101320, 2023.

[A134] Abhirup Banerjee and Pradipta Maji. Rough set based homogeneous unsharp masking for bias field correction in mri. In *Image Analysis and Processing–ICIAP 2013: 17th International Conference, Naples, Italy, September 9-13, 2013, Proceedings, Part II 17*, pages 542–551. Springer, 2013.

[A135] Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 269–277, 2019.

[A136] Jose Picado, Arash Termehchy, Alan Fern, Sudhanshu Pathak, Praveen Ilango, and John Davis. Scalable and usable relational learning with automatic language bias. In *Proceedings of the 2021 International Conference on Management of Data*, pages 1440–1451, 2021.

[A137] Roberto Diversi, Andrea Bartolini, Andrea Tilli, Francesco Beneventi, and Luca Benini. Scc thermal model identification via advanced bias-compensated least-squares. In *2013 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 230–235. IEEE, 2013.

[A138] John G Maneatis, Jaeha Kim, Iain McClatchie, Jay Maxey, and Manjusha Shankaradas. Self-biased high-bandwidth low-jitter 1-to-4096 multiplier clock

generator pll. In *Proceedings of the 40th annual Design Automation Conference*, pages 688–690, 2003.

[A139] Anand Krishna, Johannes Rodrigues, Vanessa Mitschke, and Andreas B Eder. Self-reported mask-related worrying reduces relative avoidance bias toward unmasked faces in individuals with low covid19 anxiety syndrome. *Cognitive Research: Principles and Implications*, 6(1):1–9, 2021.

[A140] Tae Yano, Philip Resnik, and Noah A Smith. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 152–158, 2010.

[A141] Bhanu Jain, Manfred Huber, Leonidas Fegaras, and Ramez A Elmasri. Singular race models: addressing bias and accuracy in predicting prisoner recidivism. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pages 599–607, 2019.

[A142] Zihao Wang, Kejun Zhang, Yuxing Wang, Chen Zhang, Qihao Liang, Pengfei Yu, Yongsheng Feng, Wenbo Liu, Yikai Wang, Yuntao Bao, et al. Songdriver: Real-time music accompaniment generation without logical latency nor exposure bias. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1057–1067, 2022.

[A143] Li Xiang, Jiping Guan, Jie Xiang, Lifeng Zhang, and Fuhan Zhang. Spatiotemporal model based on transformer for bias correction and temporal downscaling of forecasts. *Frontiers in Environmental Science*, page 2288, 2022.

[A144] Joshua L Martin. Spoken corpora data, automatic speech recognition, and bias against african american language: The case of habitual'be'. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 284–284, 2021.

[A145] Meijun Gao and Kevin J Liu. Statistical analysis of gc-biased gene conversion and recombination hotspots in eukaryotic genomes: a phylogenetic hidden markov model-based approach. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–24, 2021.

[A146] Claude Aime and Eric Aristidi. Statistical approach to bias effects in the techniques of speckle interferometry and speckle masking. *JOSA A*, 9(10):1812–1821, 1992.

[A147] Jenny Wikström, Lars-Gunnar Lundh, and Joakim Westerlund. Stroop effects for masked threat words: Pre-attentive bias or selective awareness? *Cognition and Emotion*, 17(6):827–842, 2003.

[A148] Shaofeng Huang, Shanshan Li, Hongming Yin, Yuanqing Xiao, Liang Zhang, and Junlei Zhao. Study of transformer differential protection based on self-adaptive biased characteristic curve. In *2016 China International Conference on Electricity Distribution (CICED)*, pages 1–5. IEEE, 2016.

[A149] Nan Cao, Dongqing Wang, Ye Zhang, Jianwei Zhang, Zhiwei Xue, Weidong Guo, and Sulei Huang. Study on optimized configuration method of transformer magnetic bias treatment devices. In *2017 IEEE Transportation Electrification Conference and Expo, Asia-Pacific (ITEC Asia-Pacific)*, pages 1–5. IEEE, 2017.

[A150] Keisuke Kikutani, Takashi Ohashi, Akihiro Kojima, Itsuko Sakai, Junko Abe, Hisataka Hayashi, Akio Ui, and Tokuhisa Ohiwa. Sub-45 nm sio2 etching with stacked-mask process using high-bias-frequency dual-frequency-superimposed rf capacitively coupled plasma. *Japanese journal of applied physics*, 47(10R):8026, 2008.

[A151] François Bouchet and Jean-Paul Sansonnet. Subjectivity and cognitive biases modeling for a realistic and efficient assisting conversational agent. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 209–216. IEEE, 2009.

[A152] Xu Jun-Hao, Jesper Wittborn, Björn Rodell, AM Grishin, and KV Rao. Surface microstructures and properties of yba2cu3o7- x films by bias-masked on-axis rf sputtering. *Materials Letters*, 21(3-4):357–361, 1994.

[A153] Parul Chopra, Sai Krishna Rallabandi, Alan W Black, and Khyathi Raghavi Chandu. Switch point biased self-training: Re-purposing pretrained models for code-switching. *arXiv preprint arXiv:2111.01231*, 2021.

[A154] Adam M Wilson, Benoit Parmentier, and Walter Jetz. Systematic land cover bias in collection 5 modis cloud mask and derived products—a global overview. *Remote Sensing of Environment*, 141:149–154, 2014.

[A155] Rei Ueno, Manami Suzuki, and Naofumi Homma. Tackling biased pufs through biased masking: A debiasing method for efficient fuzzy extractor. *IEEE Transactions on Computers*, 68(7):1091–1104, 2019.

[A156] Bin Wu, Sakriani Sakti, Jinsong Zhang, and Satoshi Nakamura. Tackling perception bias in unsupervised phoneme discovery using dpgmm-rnn hybrid model and functional load. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:348–362, 2020.

[A157] Michael K Oisten and Paul L Bergstrom. The effect of a biased conductive mask on porous silicon formation. *physica status solidi c*, 6(7):1541–1545, 2009.

[A158] Maximilian A Primbs, Mike Rinck, Rob Holland, Wieke Knol, Anique Nies, and Gijsbert Bijlstra. The effect of face masks on the stereotype effect in emotion perception. *Journal of Experimental Social Psychology*, 103:104394, 2022.

[A159] Teresa A Victor, Maura L Furey, Stephen J Fromm, Patrick SF Bellgowan, Arne Öhman, and Wayne C Drevets. The extended functional neuroanatomy of emotional processing biases for masked faces in major depressive disorder. 2012.

[A160] YY Yao, DX Xie, BD Bai, and LS Zeng. The field-circuit coupled transient nonlinear analysis of 3-phase transformers under direct current bias. *JSAEM STUDIES IN APPLIED ELECTROMAGNETICS AND MECHANICS*, 2001.

[A161] Sergey Dmitriyevich Poletayev and Aleksandr Ivanovich Lyubimov. The influence of metal masks on matching of the lower electrode and a high-frequency bias generator at reactive ion etching of large substrates. *Technical Physics Letters*, pages 1–4, 2021.

[A162] Ana Paula Soares, Mariana Velho, and Helena Mendes Oliveira. The role of letter features on the consonant-bias effect: Evidence from masked priming. *Acta Psychologica*, 210:103171, 2020.

[A163] R Bolotovsky, MARK A White, A Darovsky, and P Coppens. Theseed-skewness' method for integration of peaks on imaging plates. *Journal of applied crystallography*, 28(2):86–95, 1995.

[A164] Xiaorong Huang, Xiongbo Peng, Fei Xie, Wanying Mao, Hong Chen, and Meng-Xiang Sun. The stereotyped positioning of the generative cell associated with vacuole dynamics is not required for male gametogenesis in rice pollen. *New Phytologist*, 218(2):463–469, 2018.

[A165] Genevieve L Quek and Matthew Finkbeiner. The upper-hemifield advantage for masked face processing: Not just an attentional bias. *Attention, Perception, & Psychophysics*, 78:52–68, 2016.

[A166] Jose Picado, Arash Termehchy, Alan Fern, and Sudhanshu Pathak. Towards automatically setting language bias in relational learning. In *Proceedings of the 1st Workshop on Data Management for End-to-End Machine Learning*, pages 1–4, 2017.

[A167] S Samuel Yang, Wayne Wenzhong Xu, Mesfin Tesfaye, JoAnn FS Lamb, Hans-Joachim G Jung, Kathryn A VandenBosch, Carroll P Vance, and John W Gronwald. Transcript profiling of two alfalfa genotypes with contrasting cell wall com-

position in stems using a cross-species platform: optimizing analysis by masking biased probes. *BMC genomics*, 11(1):1–18, 2010.

[A168] Nassim Shahbazi, Sajad Bagheri, and Gevork B Gharehpetian. Transformer differential protection scheme based on self-adaptive biased characteristic curve and considering cross-country faults. *International Journal of Circuit Theory and Applications*, 51(3):1110–1131, 2023.

[A169] Laurent Sartran, Samuel Barrett, Adhiguna Kuncoro, Miloš Stanojević, Phil Blunsom, and Chris Dyer. Transformer grammars: Augmenting transformer language models with syntactic inductive biases at scale. *Transactions of the Association for Computational Linguistics*, 10:1423–1439, 2022.

[A170] José Antonio García-Díaz, Miguel Ángel Rodríguez-García, Francisco García-Sánchez, and Rafael Valencia-García. Umuteam at irostereo: Profiling irony and stereotype spreaders on twitter combining linguistic features with transformers. 2022.

[A171] Manuel Colavincenzo, Pierluigi Monaco, Emiliano Sefusatti, and Stefano Borgani. Uncertainty in the visibility mask of a survey and its effects on the clustering of biased tracers. *Journal of Cosmology and Astroparticle Physics*, 2017(03):052, 2017.

[A172] Andriy Nikolov and Mathieu d'Aquin. Uncovering semantic bias in neural network models using a knowledge graph. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1175–1184, 2020.

[A173] Chunjie Zhang, Yifan Zhang, Shuhui Wang, Junbiao Pang, Chao Liang, Qingming Huang, and Qi Tian. Undo the codebook bias by linear transformation for visual applications. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 533–536, 2013.

[A174] Dirk Schweim, Erik Hemberg, Dominik Sobania, Una-May O'Reilly, and Franz Rothlauf. Using knowledge of human-generated code to bias the search in program synthesis with grammatical evolution. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 331–332, 2021.

[A175] Jeffrey Matayoshi and Shamya Karumbaiah. Using marginal models to adjust for statistical bias in the analysis of state transitions. In *LAK21: 11th International learning analytics and knowledge conference*, pages 449–455, 2021.

[A176] Mark W Hauschild, Martin Pelikan, Kumara Sastry, and David E Goldberg. Using previous models to bias structural learning in the hierarchical boa. In

Proceedings of the 10th annual conference on Genetic and evolutionary computation, pages 415–422, 2008.

[A177] Michael T Reilly, Stewart A Robertson, Colin R Parker, Doris Kang, Mircea V Dusa, Susan S MacDonald, and Craig A West. Verification of the effect of mask bias on the mask error enhancement factor of contact holes. In *21st Annual BACUS Symposium on Photomask Technology*, volume 4562, pages 948–953. SPIE, 2002.

[A178] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34:28522–28535, 2021.

[A179] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *International Journal of Computer Vision*, pages 1–22, 2023.

[A180] Kankan Zhou, Yibin LAI, and Jing Jiang. Vlstereoset: A study of stereotypical bias in pre-trained vision-language models. Association for Computational Linguistics, 2022.

[A181] Roma Patel and Ellie Pavlick. "was it "stated" or was it "claimed"?: How linguistic bias affects generative language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10080–10095, 2021.

[A182] Yiwei Zhou, Elena Demidova, and Alexandra I Cristea. Who likes me more? analysing entity-centric language-specific bias in multilingual wikipedia. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 750–757, 2016.

[A183] Guillaume Bouchard. Bias-variance tradeoff in hybrid generative-discriminative models. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pages 124–129. IEEE, 2007.

[A184] Alex Fine, Austin F Frank, T Florian Jaeger, and Benjamin Van Durme. Biases in predicting the human language model. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (Volume 2: Short papers)*, pages 7–12, 2014.

[A185] Karthik Kuber and Chilukuri K Mohan. Biasing evolving generations in learning classifier systems using information theoretic measures. In *Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers*, pages 2077–2080, 2009.

[A186] Ramya Srinivasan and Kanji Uchino. Biases in generative art: A causal look from the lens of art history. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 41–51, 2021.

[A187] Hai-Tao Yu, Adam Jatowt, Roi Blanco, Joemon M Jose, and Ke Zhou. A rank-biased neural network model for click modeling. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 183–191, 2019.

[A188] Yan Zhang, Xue Jiang, Siqi Liu, Bo Hu, and Xinbo Gao. Boundary-aware bias loss for transformer-based aerial image segmentation model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3528–3532. IEEE, 2022.

[A189] Jesús Andrés-Ferrer, Dario Albesano, Puming Zhan, and Paul Vozila. Contextual density ratio for language model biasing of sequence to sequence asr systems. *arXiv preprint arXiv:2206.14623*, 2022.

[A190] Agathe Balayn, Jie Yang, Zoltan Szlavik, and Alessandro Bozzon. Automatic identification of harmful, aggressive, abusive, and offensive language on the web: a survey of technical biases informed by psychology literature. *ACM Transactions on Social Computing (TSC)*, 4(3):1–56, 2021.

[A191] Lorena Leal Bando, Falk Scholer, and Andrew Turpin. Constructing query-biased summaries: a comparison of human and system generated snippets. In *Proceedings of the third symposium on Information interaction in context*, pages 195–204, 2010.

[A192] K Lauren Barnes, Gena Dunivan, Andrew L Sussman, Lauren McGuire, and Rohini McKee. Behind the mask: an exploratory assessment of female surgeons' experiences of gender bias. *Academic Medicine*, 95(10):1529–1538, 2020.

[A193] Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[A194] Philipp E Bayer, David Edwards, and Jacqueline Batley. Bias in resistance gene prediction due to repeat masking. *Nature Plants*, 4(10):762–765, 2018.

[A195] Amine Benamara, Jean-Claude Martin, Elise Prigent, Laurence Chaby, Mohamed Chetouani, Jean Zagdoun, Hélène Vanderstichel, Sébastien Dacunha, and Brian Ravenet. Copalz: A computational model of pathological appraisal biases for an interactive virtual alzheimer patient. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 72–81, 2022.

[A196] Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat, Wonsuk Yang, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. *arXiv preprint arXiv:2203.11933*, 2022.

[A197] Stefano Bocconi and Frank Nack. Automatic generation of biased video sequences. In *Proceedings of the 1st ACM workshop on Story representation, mechanism and context*, pages 9–16, 2004.

[A198] Haojie Cao, Zhongyuan Han, Zhenwei Mo, Zengyao Li, Ziwei Xiao, Zijian Li, and Leilei Kong. A multi-model voting ensemble classifier based on bert for profiling irony and stereotype spreaders on twitter.

[A199] Aïna Chalabaev, Philippe Sarrazin, David Trouilloud, and Lee Jussim. Can sex-undifferentiated teacher expectations mask an influence of sex stereotypes? alternative forms of sex bias in teacher expectations 1. *Journal of applied social psychology*, 39(10):2469–2498, 2009.

[A200] Lucian Chan, Rajendra Kumar, Marcel Verdonk, and Carl Poelking. A multilevel generative framework with hierarchical self-contrasting for bias control and transparency in structure-based ligand design. *Nature Machine Intelligence*, pages 1–13, 2022.

[A201] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*, pages 87–94, 2008.

[A202] Christine SM Currie and Russell CH Cheng. Balancing bias and variance in the optimization of simulation models. In *Proceedings of the Winter Simulation Conference, 2005.*, pages 6–pp. IEEE, 2005.

[A203] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021.

[A204] Udo Dannlowski, Patricia Ohrmann, Jochen Bauer, Jürgen Deckert, Christa Hohoff, Harald Kugel, Volker Arolt, Walter Heindel, Anette Kersting, Bernhard T Baune, et al. 5-httlpr biases amygdala activity in response to masked facial expressions in major depression. *Neuropsychopharmacology*, 33(2):418–424, 2008.

[A205] Udo Dannlowski, Patricia Ohrmann, Jochen Bauer, Harald Kugel, Volker Arolt, Walter Heindel, Anette Kersting, Bernhard T Baune, and Thomas Suslow.

Amygdala reactivity to masked negative faces is associated with automatic judgmental bias in major depression: a 3 t fmri study. *Journal of Psychiatry and Neuroscience*, 32(6):423–429, 2007.

[A206] Hiroshi Fuketa, Masanori Hashimoto, Yukio Mitsuyama, and Takao Onoye. Correlation verification between transistor variability model with body biasing and ring oscillation frequency in 90nm subthreshold circuits. In *Proceedings of the 2008 international symposium on Low Power Electronics & Design*, pages 3–8, 2008.

[A207] Aditya Grover, Jiaming Song, Ashish Kapoor, Kenneth Tran, Alekh Agarwal, Eric J Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. *Advances in neural information processing systems*, 32, 2019.

[A208] Graziano Guella, Ines Mancini, Aysel Öztunç, and Francesco Pietra. Conformational bias in macrocyclic ethers and observation of high solvolytic reactivity at a masked furfuryl (= 2-furylmethyl) c-atom. *Helvetica Chimica Acta*, 83(2):336–348, 2000.

[A209] L Gui, W Merzkirch, and R Fei. A digital mask technique for reducing the bias error of the correlation-based piv interrogation algorithm. *Experiments in fluids*, 29(1):30–35, 2000.

[A210] Lb Gui, J Longo, and F Stern. Biases of piv measurement of turbulent flow and the masked correlation-based interrogation algorithm. *Experiments in Fluids*, 30(1):27–35, 2001.

[A211] Saket Gupta and Sachin S Sapatnekar. Current source modeling in the presence of body bias. In *2010 15th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 199–204. IEEE, 2010.

[A212] Shaohua Han, Xinqiang Wan, Yong Lai, Lei Ge, and Jinian Pang. Dc bias diagnosis of power transformer with soundprint features. In *Proceedings of the 7th International Conference on Cyber Security and Information Engineering*, pages 213–217, 2022.

[A213] Brian T Hutsel, Scott D Kovaleski, Emily A Baxter, and Jae Wan Kwon. Charged-particle emission and self-biasing of a piezoelectric transformer plasma source. *IEEE Transactions on Plasma Science*, 41(1):99–105, 2012.

[A214] Dietmar Jannach, Iman Kamehkhosh, and Geoffray Bonnin. Biases in automated music playlist generation: A comparison of next-track recommending techniques. In *Proceedings of the 2016 conference on user modeling adaptation and personalization*, pages 281–285, 2016.

[A215] Young-Chang Kim, Geert Vandenberghe, and Kurt G Ronse. Attpsm cd control: mask bias and flare effects. In *Optical Microlithography XV*, volume 4691, pages 1041–1053. SPIE, 2002.

[A216] Dean Knox, Will Lowe, and Jonathan Mummolo. Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637, 2020.

[A217] Guneet Kohli, Prabsimran Kaur, and Jatin Bedi. Arguably at comma@ icon: Detection of multilingual aggressive, gender biased, and communally charged tweets using ensemble and fine-tuned indicbert. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 46–52, 2021.

[A218] David P Kreil and Christos A Ouzounis. Comparison of sequence masking algorithms and the detection of biased protein sequence regions. *Bioinformatics*, 19(13):1672–1681, 2003.

[A219] Jan-David Krieger, Timo Spinde, Terry Ruas, Juhi Kulshrestha, and Bela Gipp. A domain-adaptive pre-training approach for language bias detection in news. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–7, 2022.

[A220] Onkar Krishna and Kiyoharu Aizawa. Age-adapted saliency model with depth bias. In *Proceedings of the ACM Symposium on Applied Perception*, pages 1–8, 2017.

[A221] Sanjay V Kumar, Chris H Kim, and Sachin S Sapatnekar. An analytical model for negative bias temperature instability. In *Proceedings of the 2006 IEEE/ACM international conference on Computer-aided design*, pages 493–496, 2006.

[A222] Jennifer YF Lau and Essi M Viding. Anxiety-related biases in children's avoidant responses to a masked angry face. *Behaviour research and therapy*, 45(7):1639–1645, 2007.

[A223] Margherita Lembo, Giulio Fabbian, Julien Carron, and Antony Lewis. Cmb lensing reconstruction biases from masking extragalactic sources. *Physical Review D*, 106(2):023525, 2022.

[A224] Xingming Liang, Quanhua Liu, Banghua Yan, and Ninghai Sun. A deep learning trained clear-sky mask algorithm for viirs radiometric bias assessment. *Remote Sensing*, 12(1):78, 2019.

[A225] Hou-An Lin, Yuan-Chang Wang, and Hwann-Kaeo Chiou. A 5-11 ghz cmos power amplifier using guanella-type transmission line transformer and adaptive bias circuit. *Microwave and Optical Technology Letters*, 61(1):267–270, 2019.

[A226] Wenbin Lin, Zhongyuan Han, Jinxi Zhang, Zengyao Li, Guiyuan Cao, Jianhong Yu, and Leilei Kong. A bert-based model for profiling irony and stereotype spreaders on twitter. In *CLEF*, pages 1613–0073, 2022.

[A227] Rolf Lohaus, Nicholas L Geard, Janet Wiles, and Ricardo BR Azevedo. A generative bias towards average complexity in artificial cell lineages. *Proceedings of the Royal Society B: Biological Sciences*, 274(1619):1741–1751, 2007.

[A228] Michael EJ Masson. Bias in masked word identification: Unconscious influences of repetition priming. *Psychonomic bulletin & review*, 9(4):773–779, 2002.

[A229] Hai Min, Wei Jia, and Yang Zhao. A region-bias fitting model based level set for segmenting images with intensity inhomogeneity. In *Proceedings of the 2nd International Symposium on Image Computing and Digital Medicine*, pages 83–87, 2018.

[A230] Mkhuseli Ngxande, Jules-Raymond Tapamo, and Michael Burke. Bias remediation in driver drowsiness detection systems using generative adversarial networks. *IEEE Access*, 8:55592–55601, 2020.

[A231] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *26th International Conference on Intelligent User Interfaces*, pages 340–350, 2021.

[A232] Daniel Parres and Claudia Gomez. Bert sentence embeddings in different machine learning and deep learning models for author profiling applied to irony and stereotype spreaders on twitter. 2022.

[A233] William Pencak. Canada as a semiotic society: Harold innis, roberta kevelson, and the bias of legal communications. *International Journal for the Semiotics of Law*, 18(2):207–215, 2005.

[A234] Frederick J Riggins and David M Weber. A model of peer-to-peer (p2p) social lending in the presence of identification bias. In *Proceedings of the 13th International Conference on Electronic Commerce*, pages 1–8, 2011.

[A235] Namsik Ryu, Seunghyun Jang, Kwang Chun Lee, and Yongchae Jeong. Cmos doherty amplifier with variable balun transformer and adaptive bias control for wireless lan application. *IEEE Journal of Solid-State Circuits*, 49(6):1356–1365, 2014.

[A236] Joni Salminen, Soon-gyo Jung, Shammur Chowdhury, and Bernard J Jansen. Analyzing demographic bias in artificially generated facial pictures. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2020.

[A237] Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. Bert is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, 2019.

[A238] Yukiko Umemoto, Koji Nii, Jiro Ishikawa, Kazuyoshi Okamoto, Kazutaka Mori, and Kazumasa Yanagisawa. A 28 nm 50% power reduced 2t mask rom with 0.72 ns read access time using column source bias. In *2011 IEEE Custom Integrated Circuits Conference (CICC)*, pages 1–4. IEEE, 2011.

[A239] Robbert Van Den Berg, Boris Skoric, and Vincent van der Leest. Bias-based modeling and entropy analysis of pufs. In *Proceedings of the 3rd international workshop on Trustworthy embedded devices*, pages 13–20, 2013.

[A240] Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1324–1332, 2022.

[A241] HK Verma and RP Maheshwari. Adaptive digital differential protection of transformer improvement over the fixed bias scheme. *Journal of the Institution of Engineers (India): Electrical Engineering Division*, 1996.

[A242] Sholom Wacholder, Jay H Lubin, Mustafa Dosemeci, and Mitchell H Gail. Bias despite masked assessment of clinical outcomes when an outcome is defined as one of several component events. *Controlled clinical trials*, 12(4):457–461, 1991.

[A243] Dongjiang Wang, Minda Hu, Junqing Zhou, Chenglong Zhang, Xinpeng Wang, and Haiyang Zhang. Cd bias loading control in metal hard mask open process. *ECS Transactions*, 52(1):317, 2013.

[A244] Zezhong Wang, Junling Wu, Lianguang Liu, Yuyan Li, Lu Chen, and Zhiguang Guo. Analysis of current characteristics of transformer bias caused by subway stray current based on measured data. In *Journal of Physics: Conference Series*, volume 2087, page 012084. IOP Publishing, 2021.

[A245] Yuan Xia, Dongfeng Liu, Jinkui Zhang, and Ke Li. Bert-based automated risk of bias assessment; [bert]. *Chinese Journal of Evidence-Based Medicine*, 21(2):204 – 209, 2021. Cited by: 0.

[A246] Lei Xu, Zhentao Liu, Peng Liu, and Liyan Cai. A low spectral bias generative adversarial model for image generation. In *Data Science: 8th International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2022, Chengdu, China, August 19–22, 2022, Proceedings, Part I*, pages 354–362. Springer, 2022.

[A247] Shuai Zhang, Yongliang Shen, Zeqi Tan, Yiquan Wu, and Weiming Lu. Debias for generative extraction in unified ner task. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 808–818, 2022.

[A248] Shengjia Zhao, Hongyu Ren, Arianna Yuan, Jiaming Song, Noah Goodman, and Stefano Ermon. Bias and generalization in deep generative models: An empirical study. *Advances in Neural Information Processing Systems*, 31, 2018.

[A249] Zhi Zheng, Zhaopeng Qiu, Tong Xu, Xian Wu, Xiangyu Zhao, Enhong Chen, and Hui Xiong. Cbr: Context bias aware recommendation for debiasing user modeling and click prediction. In *Proceedings of the ACM Web Conference 2022*, pages 2268–2276, 2022.

[A250] Jianghong Zhou, Sayyed M Zahiri, Simon Hughes, Khalifeh Al Jadda, Surya Kallumadi, and Eugene Agichtein. De-biased modeling of search click behavior with reinforcement learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1637–1641, 2021.

[A251] Dimion Asael, Zachary Ziegler, and Yonatan Belinkov. A generative approach for mitigating structural biases in natural language inference. *arXiv preprint arXiv:2108.14006*, 2021.

[A252] Martin Ragot, Nicolas Martin, and Salomé Cojean. Ai-generated vs. human artworks. a perception bias towards artificial intelligence? In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–10, 2020.

[A253] Annerieke Heuvelink, Michel CA Klein, and Jan Treur. An agent memory model enabling rational and biased reasoning. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 193–199. IEEE, 2008.

[A254] James Montgomery and Daniel Ashlock. Applying the biased form of the adaptive generative representation. In *2017 IEEE Congress on Evolutionary Computation (CEC)*, pages 1079–1086. IEEE, 2017.

[A255] Yanbo Fang, Zuohui Fu, Xin Luna Dong, Yongfeng Zhang, and Gerard de Melo. Assessing combinational generalization of language models in biased scenarios. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 392–397, 2022.

[A256] Courtney Mansfield, Amandalynne Paullada, and Kristen Howell. Behind the mask: Demographic bias in name detection for pii masking. *arXiv preprint arXiv:2205.04505*, 2022.

[A257] Xiao Li, Min Fang, and Haikun Li. Bias alleviating generative adversarial network for generalized zero-shot classification. *Image and Vision Computing*, 105:104077, 2021.

[A258] Wietske van Osch, Michel Avital, and Orr Mendelson. Biases in usefulness assessment: the realized value of generative support systems. 2011.

[A259] Ling Liu and Mans Hulden. Can a transformer pass the wug test? tuning copying bias in neural morphological inflection models. *arXiv preprint arXiv:2104.06483*, 2021.

[A260] Daniel McDuff, Shuang Ma, Yale Song, and Ashish Kapoor. Characterizing bias in classifiers using generative models. *Advances in neural information processing systems*, 32, 2019.

[A261] Yi Zhang, Junyang Wang, and Jitao Sang. Counterfactually measuring and eliminating social bias in vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4996–5004, 2022.

[A262] Jahna Otterbacher. Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1955–1964, 2015.

[A263] Navid Rekabsaz and Markus Schedl. Do neural ranking models intensify gender bias? In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2065–2068, 2020.

[A264] Catherine M Tabor. Examining language bias in computer science. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pages 1413–1413, 2020.

[A265] Deepa Muralidhar. Examining religion bias in ai text generators. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 273–274, 2021.

[A266] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 349–358, 2019.

[A267] Yan Chen, Christopher Mahoney, Isabella Grasso, Esma Wali, Abigail Matthews, Thomas Middleton, Mariama Njie, and Jeanna Matthews. Gender bias and

under-representation in natural language processing across human languages. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 24–34, 2021.

[A268] Adel Abusitta, Esma Aïmeur, and Omar Abdel Wahab. Generative adversarial networks for mitigating biases in machine learning systems. *arXiv preprint arXiv:1905.09972*, 2019.

[A269] Steven Phillips and Miroslav Dudík. Generative and discriminative learning with unknown labeling bias. *Advances in Neural information Processing Systems*, 21, 2008.

[A270] Ulrich von Hecker, Grzegorz Sedek, Kinga Piber-Dabrowska, and Sylwia Bedynska. Generative reasoning as influenced by depression, aging, stereotype threat, and prejudice. 2005.

[A271] Yizhi Li, Ge Zhang, Bohao Yang, Chenghua Lin, Shi Wang, Anton Ragni, and Jie Fu. Herb: Measuring hierarchical regional bias in pre-trained language models. *arXiv preprint arXiv:2211.02882*, 2022.

[A272] Periklis Perikleous, Andreas Kafkalias, Zenonas Theodosiou, Pinar Barlas, Evgenia Christoforou, Jahna Otterbacher, Gianluca Demartini, and Andreas Lanitis. How does the crowd impact the model? a tool for raising awareness of social bias in crowdsourced training data. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4951–4954, 2022.

[A273] Muhammad Asim, Max Daniels, Oscar Leong, Ali Ahmed, and Paul Hand. Invertible generative models for inverse problems: mitigating representation error and dataset bias. In *International Conference on Machine Learning*, pages 399–409. PMLR, 2020.

[A274] Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. Iterative adversarial removal of gender bias in pretrained word embeddings. In *Proceedings of the 37th ACM/SIGAPP Symposium On Applied Computing*, pages 829–836, 2022.

[A275] Nina Markl. Language variation and algorithmic bias: understanding algorithmic bias in british english automatic speech recognition. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 521–534, 2022.

[A276] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268, 2022.

[A277] Zujie Liang, Haifeng Hu, and Jiaying Zhu. Lpf: a language-prior feedback objective function for de-biased visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1955–1959, 2021.

[A278] Said Farooq Shah, Salman Arif Cheema, Zawar Hussain, and Ejaz Ali Shah. Masking data: a solution to social desirability bias in paired comparison experiments. *Communications in Statistics-Simulation and Computation*, 51(6):3149–3167, 2022.

[A279] Shiva Omrani Sabbaghi and Aylin Caliskan. Measuring gender bias in word embeddings of gendered languages requires disentangling grammatical gender signals. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 518–531, 2022.

[A280] Osman Aka, Ken Burke, Alex Bauerle, Christina Greer, and Margaret Mitchell. Measuring model biases in the absence of ground truth. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 327–335, 2021.

[A281] Jaimeen Ahn and Alice Oh. Mitigating language-dependent ethnic bias in bert. *arXiv preprint arXiv:2109.05704*, 2021.

[A282] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866, 2021.

[A283] Matan Halevy, Camille Harris, Amy Bruckman, Diyi Yang, and Ayanna Howard. Mitigating racial biases in toxic language detection with an equity-based ensemble framework. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–11. 2021.

[A284] Nikhil Shah, Apoorve Singhal, Chinmay Singh, and Yash Khandelwal. Model agnostic information biasing for vqa. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, pages 419–419, 2021.

[A285] Marisa Vasconcelos, Carlos Cardonha, and Bernardo Gonçalves. Modeling epistemological principles for bias mitigation in ai systems: an illustration in hiring decisions. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 323–329, 2018.

[A286] Christoph Hube and Besnik Fetahu. Neural based statement classification for biased language. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 195–203, 2019.

[A287] Annette Rios, Chantal Amrhein, Noëmi Aepli, and Rico Sennrich. On biasing transformer attention towards monotonicity. *arXiv preprint arXiv:2104.03945*, 2021.

[A288] Luoqiu Li, Xiang Chen, Hongbin Ye, Zhen Bi, Shumin Deng, Ningyu Zhang, and Huajun Chen. On robustness and bias analysis of bert-based relation extraction. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction: 6th China Conference, CCKS 2021, Guangzhou, China, November 4-7, 2021, Proceedings 6*, pages 43–59. Springer, 2021.

[A289] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306, 2021.

[A290] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654, 2022.

[A291] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*, 2019.

[A292] Masashi Takeshita, Rafal Rzepka, and Kenji Araki. Speciesist language and nonhuman animal bias in english masked language models. *Information Processing & Management*, 59(5):103050, 2022.

[A293] Paul Soulos, Sudha Rao, Caitlin Smith, Eric Rosen, Asli Celikyilmaz, R Thomas McCoy, Yichen Jiang, Coleman Haley, Roland Fernandez, Hamid Palangi, et al. Structural biases for improving transformers on translation into morphologically rich languages. *arXiv preprint arXiv:2208.06061*, 2022.

[A294] Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 2022.

[A295] Abigail Z Jacobs, Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. The meaning and measurement of bias: lessons from natural language processing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 706–706, 2020.

[A296] Andrew W Lo and Ruixun Zhang. The wisdom of crowds versus the madness of mobs: An evolutionary model of bias, polarization, and other challenges to collective intelligence. *Collective Intelligence*, 1(1):26339137221104785, 2022.

[A297] Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. The world of an octopus: How reporting bias influences a language model's perception of color. *arXiv preprint arXiv:2110.08182*, 2021.

[A298] Vihari Piratla, Sunita Sarawagi, and Soumen Chakrabarti. Topic sensitive attention on generic corpora corrects sense bias in pretrained embeddings. *arXiv preprint arXiv:1906.02688*, 2019.

[A299] Zhuoyan Li, Zhuoran Lu, and Ming Yin. Towards better detection of biased language with scarce, noisy, and biased annotations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 411–423, 2022.

[A300] Halim Acosta, Nathan Henderson, Jonathan Rowe, Wookhee Min, James Minogue, and James Lester. What's fair is fair: Detecting and mitigating encoded bias in multimodal models of museum visitor attention. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 258–267, 2021.

[A301] Tejas Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models. *arXiv preprint arXiv:2104.08666*, 2021.

[A302] Jiawei Chen, Anbang Xu, Zhe Liu, Yufan Guo, Xiaotong Liu, Yingbei Tong, Rama Akkiraju, and John M Carroll. A general methodology to quantify biases in natural language data. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2020.

[A303] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. Societal biases in retrieved contents: Measurement framework and adversarial mitigation of bert rankers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 306–316, 2021.

[A304] Sina Zarrieß, Hannes Gröner, Torgrim Solstad, and Oliver Bott. This isn't the bias you're looking for: Implicit causality, names and gender in german language models. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 129–134, 2022.

[A305] Ángel Pavón Pérez. Bias in artificial intelligence models in financial services. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 908–908, 2022.

[A306] Shikha Bordia and Samuel R Bowman. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*, 2019.

[A307] Serge Dolgikh. Fairness and bias in learning systems: a generative perspective.

[A308] Przemyslaw Joniak and Akiko Aizawa. Gender biases and where to find them: Exploring gender bias in pre-trained transformer-based language models using movement pruning. *arXiv preprint arXiv:2207.02463*, 2022.

[A309] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

[A310] Debora Nozza, Federcio Bianchi, Dirk Hovy, et al. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics, 2022.

[A311] Max Hort, Jie M Zhang, Federica Sarro, and Mark Harman. Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 994–1006, 2021.

[A312] Ismael Garrido-Muñoz. Analysis, detection and mitigation of biases in deep learning language models. 2022.

[A313] Magnus Sahlgren and Fredrik Olsson. Gender bias in pretrained swedish embeddings. In *Proceedings of the 22nd Nordic Conference on computational linguistics*, pages 35–43, 2019.

[A314] Lizhen Liang and Daniel E Acuna. Artificial mental phenomena: Psychophysics as a framework to detect perception biases in ai models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 403–412, 2020.

[A315] Jishun Zhao, Bingjie Du, Shucheng Zhu, and Pengyuan Liu.
(construction of chinese sentence-level gender-unbiased data set and evaluation of gender bias in pre-training language). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 564–575, 2021.

[A316] Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. French crows-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, 2022.

[A317] Michele Dusi, Nicola Arici, Alfonso E Gerevini, Luca Putelli, and Ivan Serina. Graphical identification of gender bias in bert with a weakly supervised approach.

In *Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022) co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2022)*, 2022.

[A318] Giuseppe Samo, Caterina Bonan, and Fuzhen SIa. Health-related content in transformer-based deep neural network language models: Exploring cross-linguistic syntactic bias. *Stud Health Technol Inform*, pages 221–225, 2022.

[A319] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. Investigating gender bias in bert. *Cognitive Computation*, 13(4):1008–1018, 2021.

[A320] Juliana Shihadeh, Margareta Ackerman, Ashley Troske, Nicole Lawson, and Edith Gonzalez. Brilliance bias in gpt-3. In *2022 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 62–69. IEEE, 2022.

[A321] Oskar Van Der Wal, Jaap Jumelet, Katrin Schulz, and Willem Zuidema. The birth of bias: A case study on the evolution of gender bias in an english language model. *arXiv preprint arXiv:2207.10245*, 2022.

[A322] Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. Theory-grounded measurement of us social stereotypes in english language models. *arXiv preprint arXiv:2206.11684*, 2022.

[A323] Sijing Zhang and Ping Li. Unmasking the stereotypes: Evaluating social biases in chinese bert. In *2022 4th International Conference on Natural Language Processing (ICNLP)*, pages 324–330. IEEE, 2022.

[A324] Alfa Yohannis and Dimitris Kolovos. Towards model-based bias mitigation in machine learning. In *Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems*, pages 143–153, 2022.

[A325] Karën Fort, Aurélie Névéol, Yoann Dupont, and Julien Bezançon. Use of a citizen science platform for the creation of a language resource to study bias in language models for french: a case study. In *2nd LREC Workshop on Novel Incentives in Data Collection from People*, 2022.

[A326] Alexander Kwako, Yixin Wan, Jieyu Zhao, Kai-Wei Chang, Li Cai, and Mark Hansen. Using item response theory to measure gender and racial bias of a bert-based automated english speech assessment system. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 1–7, 2022.

[A327] Saurabh Gupta, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Viable threat on news reading: Generating biased news using natural language models. *arXiv preprint arXiv:2010.02150*, 2020.