Master's Thesis:

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

Chaewon Yun, CSS, RWTH Aachen

2023-04-26

Overview

Table of Contents

Introduction

Structural literature review

Empirical analysis

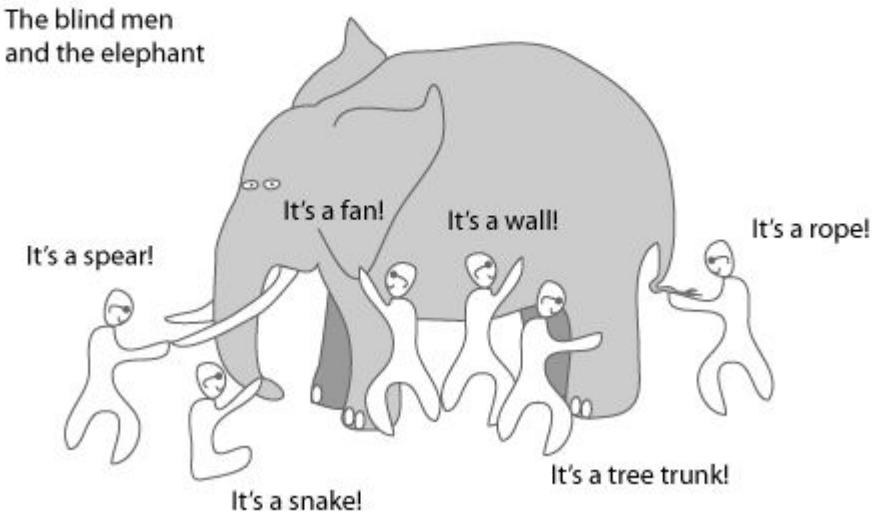
Normative analysis

Framework

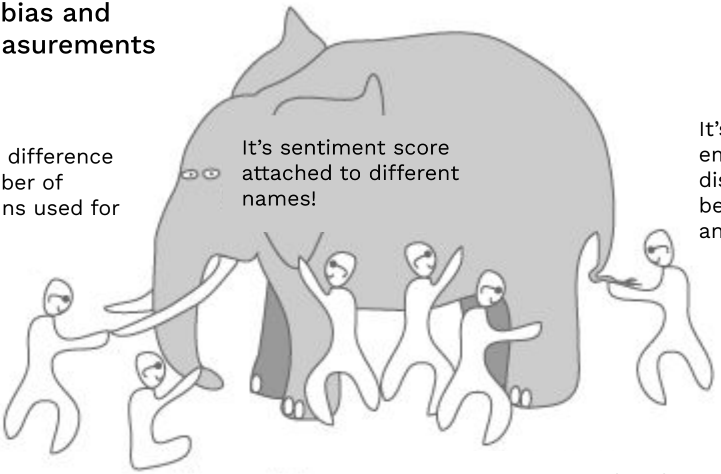
Conclusion

Gender Bias and Bias Measurements

Drawing an elephant blind-folded



Gender bias and Bias measurements



It's the difference in number of pronouns used for jobs!

It's sentiment score attached to different names!

It's an embedding distance between 'he' and 'she'!

It's the frequency of gendered pronouns!

It's topics related to character's genders!

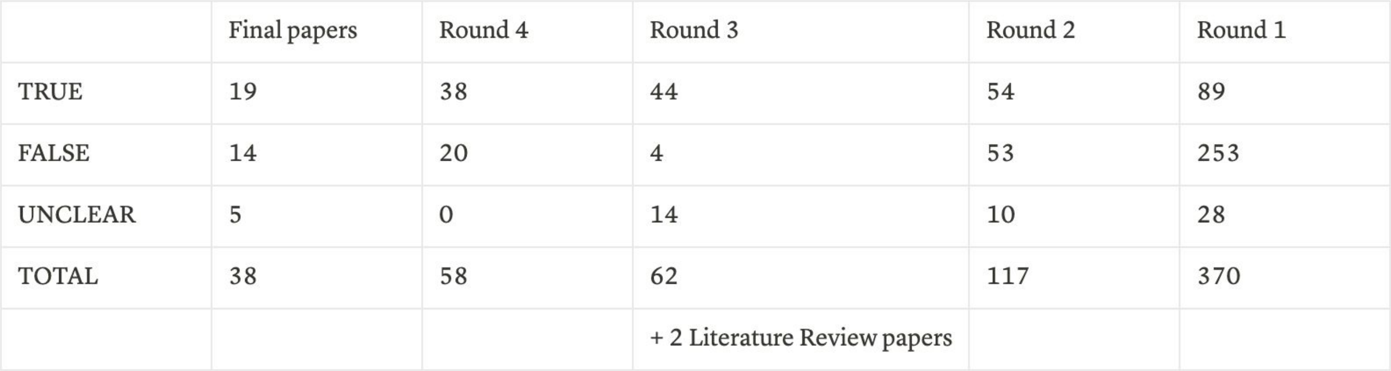
Original image credit: https://www.johnzada.com/blog/the-problem-with-paradigms/

“Language model = biased = bad”?

- It is unclear what it means to be *biased* for language models
    - 1) because bias is a complex concept
    - 2) and bias has been operationalized poorly
- Why bias in language models undesirable? Because the model’s output is
    - 1) descriptively inaccurate
    - 2) morally wrong
- Existing measures are limited both on empirical and normative fronts
- How can we identify, measure, and mitigate bias in a meaningful way?
    - Framework for comprehensive evaluation of bias in language models

Systematic literature review

- Review existing gender bias measurements in language models
- 19 papers suggesting novel metrics to measure gender bias
- Method
    - Codebook, Query, Database, Screening process
- Result



	Final papers	Round 4	Round 3	Round 2	Round 1
TRUE	19	38	44	54	89
FALSE	14	20	4	53	253
UNCLEAR	5	0	14	10	28
TOTAL	38	58	62	117	370
			+ 2 Literature Review papers		

Conceptual Framework to Evaluate Bias Measurements

- Levels of Conceptualization
    - Background concept, Systematized concept, Indicator (Adcock and Collier, 2001)
- Operationalization: How to measure gender bias?
    - Definition - Data - Metric alignment
- Measurement modeling (Adcock and Collier, 2001, Jacobs and Wallach 2021)
    - Face validity, Content validity, Convergent/Discriminant validity

Analysis

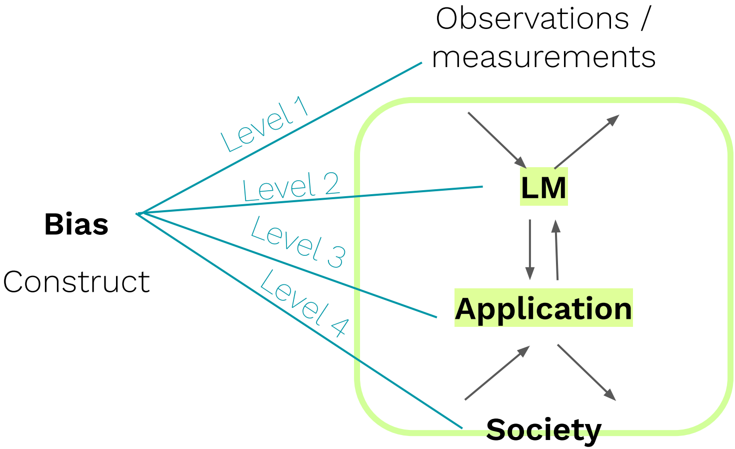
1. Underspecified conceptualization of ‘gender bias’
    1. Lack of explicit definition of construct being measured
    2. Heterogeneity of definitions makes metrics incomparable
2. Limited inventory of gender bias operationalization
    1. Stereotypes (63%, 13 papers), especially occupational stereotypes (29%, 6 papers), and sentiment scores (19%, 4 papers) account for majority of operationalization (17 out of 19 papers, out of 21 metric types)
    2. Suggest alternative dimensions of gender bias to measure, inspired by psychological scales (Samory et al. 2021)
3. Lack of validation for measurements
    1. Only few (27%, 4 papers) measurements refer previous works in related fields such as psychology, sociology, and linguistics to justify their operationalizations (Content validity)
    2. Some(32%, 6 papers) do not validate their measurement in any ways (Discriminant, Convergent validity)

Normative Analysis

1. Underspecified normative motivation
   - Despite motivating bias measurement on normative grounds, papers do not engage in normative reasoning
2. Objectives of existing bias measurements
   - Trade-off between *Descriptive Accuracy* and *Normative Correctness*
   - Beyond the trade-off view
3. Towards constructive bias critique of language models
   - Beyond outcome-based fairness and trade-off
   - Acknowledge structural injustice in language models

Framework for a Comprehensive Bias Measurement

- What should be considered to identify and measure bias in LM
- Consists of 4 levels
  1. Conceptualization of the construct, Alignment of construct and measurements
  2. Intrinsic and extrinsic evaluation
  3. Application-level evaluation
  4. Structural evaluation

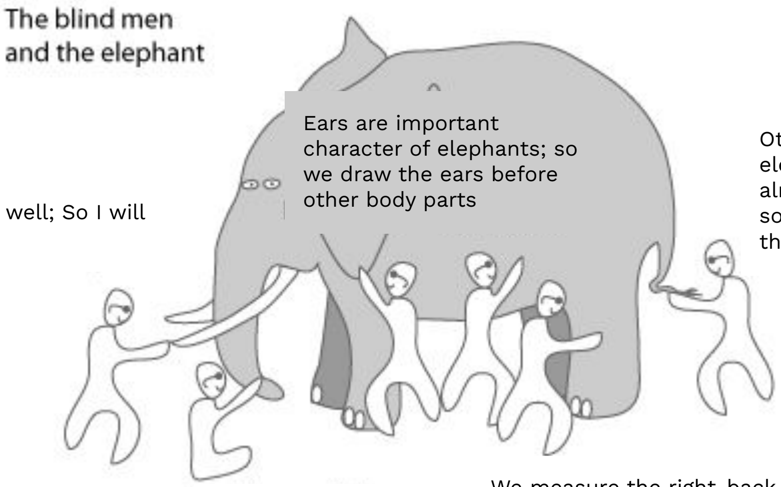


Level 1	Measurement modeling / validation
Level 2	Intrinsic / Extrinsic evaluation
Level 3	Application evaluation
Level 4	Structural evaluation

Empirical Analysis

Gender Bias and Bias Measurements

Drawing the elephant *better*:



Ears are important character of elephants; so we draw the ears before other body parts

I can touch the ivory well; So I will draw the ivory

Other parts of elephant is already drawn so I will draw the tail

We have a long ruler: So we are measuring the length of the trunk

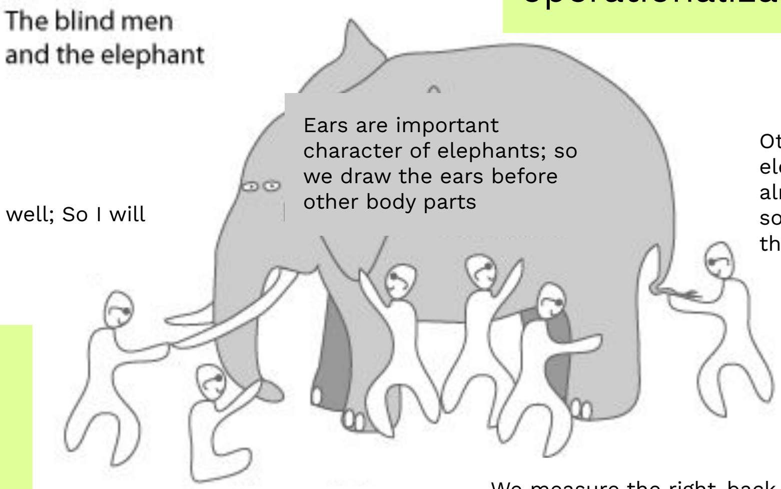
We measure the right-back leg and approximate others, because we cannot access all the legs.. But front legs might be different

Original image credit: https://www.johnzada.com/blog/the-problem-with-paradigms/

Gender Bias and Bias Measurements

Drawing the elephant *better*:

Theory-ground approaches for all-rounded operationalization



Ears are important character of elephants; so we draw the ears before other body parts

I can touch the ivory well; So I will draw the ivory

Other parts of elephant is already drawn so I will draw the tail

We have a long ruler: So we are measuring the length of the trunk

We measure the right-back leg because we cannot access all the legs.. But front legs might be different

Based on available tools - is it the best tool accessible? Is there other tool to measure this?

Built upon (comparable) existing works

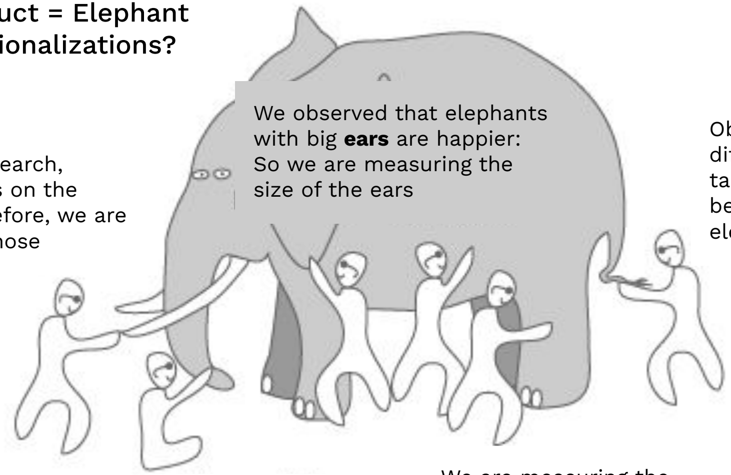
Acknowledge the limitation of the measurement

Original image credit: https://www.johnzada.com/blog/the-problem-with-paradigms/

Gender Bias and Bias Measurements

Measuring the elephant properly:

Construct = Elephant
Operationalizations?



According to previous research, elephant's **life span** relies on the length of the trunk: Therefore, we are measuring the length of nose

We observed that elephants with big **ears** are happier: So we are measuring the size of the ears

Observe differences in tail shapes between elephants

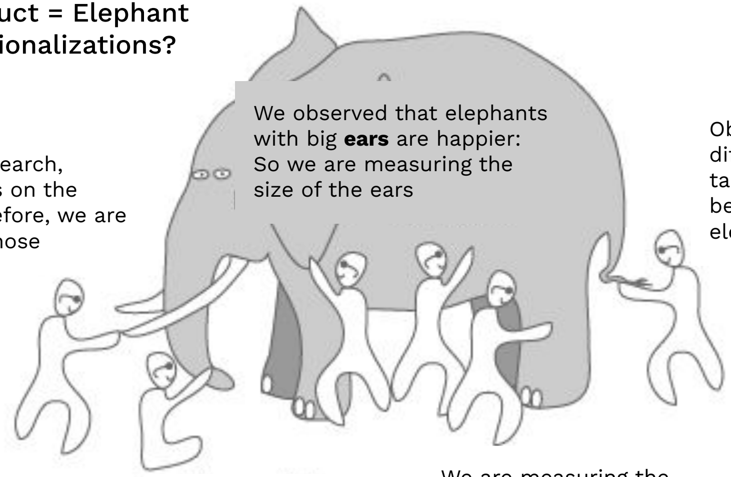
We want to make elephant **shoes**: So we are measuring the size of elephants **feet**

We are measuring the circumference of a **right-back leg,** based on previous work showing tall elephants run fast

Original image credit: https://www.johnzada.com/blog/the-problem-with-paradigms/

Why? How should we?

Measuring the elephant properly:



Construct = Elephant
Operationalizations?

According to previous research, elephant’s **life span** relies on the length of the trunk: Therefore, we are measuring the length of nose

We observed that elephants with big **ears** are happier: So we are measuring the size of the ears

Observe differences in tail shapes between elephants

We want to make elephant **shoes**: So we are measuring the size of elephants **feet**

We are measuring the circumference of a **right-back leg,** based on previous work showing tall elephants run fast

Original image credit: https://www.johnzada.com/blog/the-problem-with-paradigms/

Empirical analysis

1. Underspecified conceptualization of 'gender bias'
    a. Lack of explicit definition of construct being measured
    b. Heterogeneity of chosen background and systematized definitions makes metrics incomparable
2. Limited inventory of gender bias operationalization
    a. Stereotypes (63%, 13 papers), especially occupational stereotypes (29%, 6 papers), and sentiment scores (19%, 4 papers) accounts for majority of operationalization (out of 19 papers)
    b. Alternative dimensions of gender bias to measure, inspired by psychological scales (Samory et al. 2021)
3. Lack of validation for measurements
    a. Only few metrics refer previous works in related fields such as psychology, sociology, and linguistics to justify their operationalizations (Content validity)
    b. Not enough validation of metrics with existing, if not established, measurements to measure the same or related construct (Discriminant, Convergent validity)

1. Underspecified conceptualization of bias

Ideally, the papers should suggest definitions in all levels

- Background concept: Provide a broad overview of definition
- Systematized concept: Explains how authors formulate the concept
- Indicator: Metric

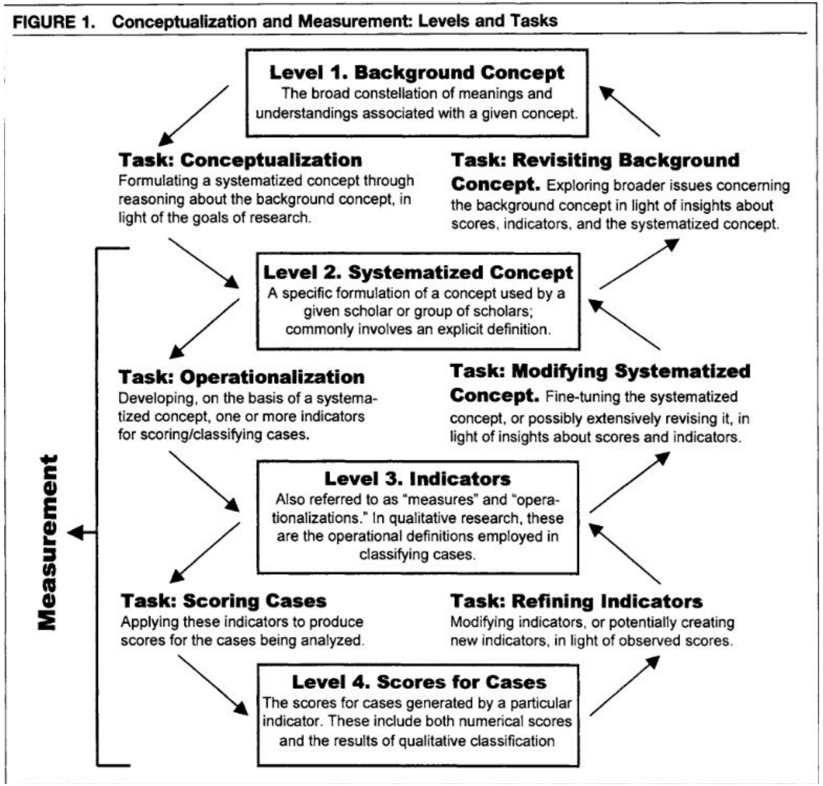


FIGURE 1. Conceptualization and Measurement: Levels and Tasks

Level 1. Background Concept The broad constellation of meanings and understandings associated with a given concept.

Task: Conceptualization Formulating a systematized concept through reasoning about the background concept, in light of the goals of research.

Task: Revisiting Background Concept. Exploring broader issues concerning the background concept in light of insights about scores, indicators, and the systematized concept.

Level 2. Systematized Concept A specific formulation of a concept used by a given scholar or group of scholars; commonly involves an explicit definition.

Task: Operationalization Developing, on the basis of a systematized concept, one or more indicators for scoring/classifying cases.

Task: Modifying Systematized Concept. Fine-tuning the systematized concept, or possibly extensively revising it, in light of insights about scores and indicators.

Level 3. Indicators Also referred to as "measures" and "operationalizations." In qualitative research, these are the operational definitions employed in classifying cases.

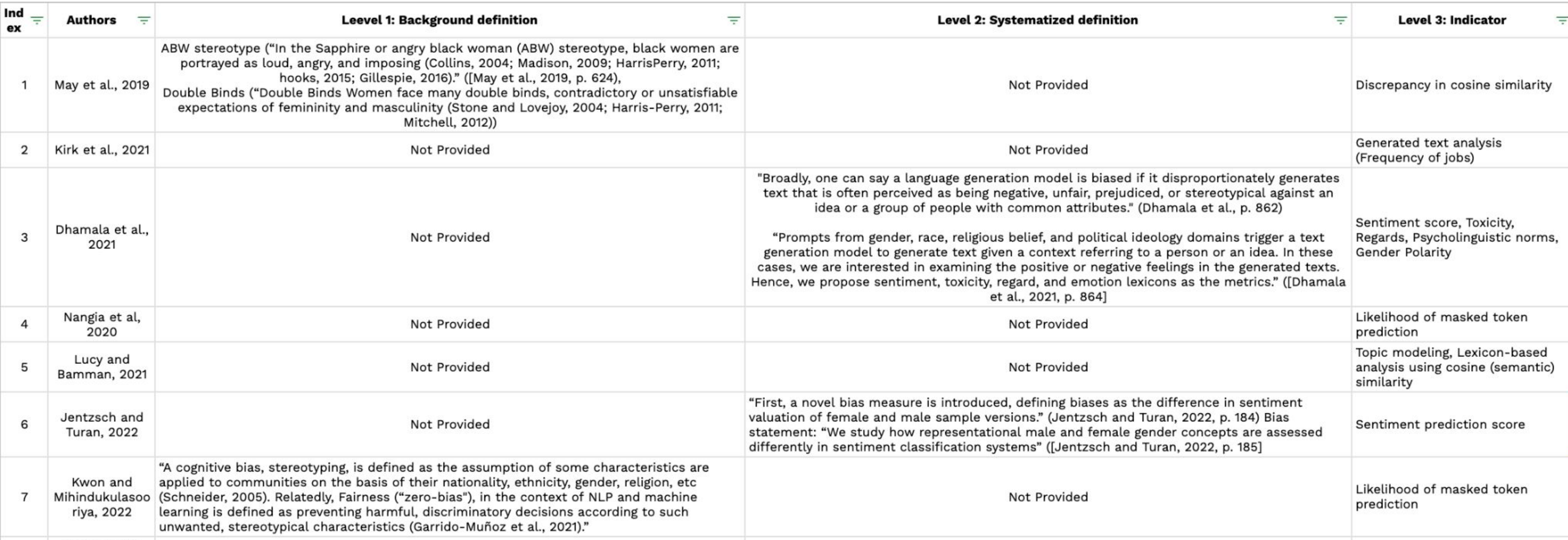
Task: Scoring Cases Applying these indicators to produce scores for the cases being analyzed.

Task: Refining Indicators Modifying indicators, or potentially creating new indicators, in light of observed scores.

Level 4. Scores for Cases The scores for cases generated by a particular indicator. These include both numerical scores and the results of qualitative classification.

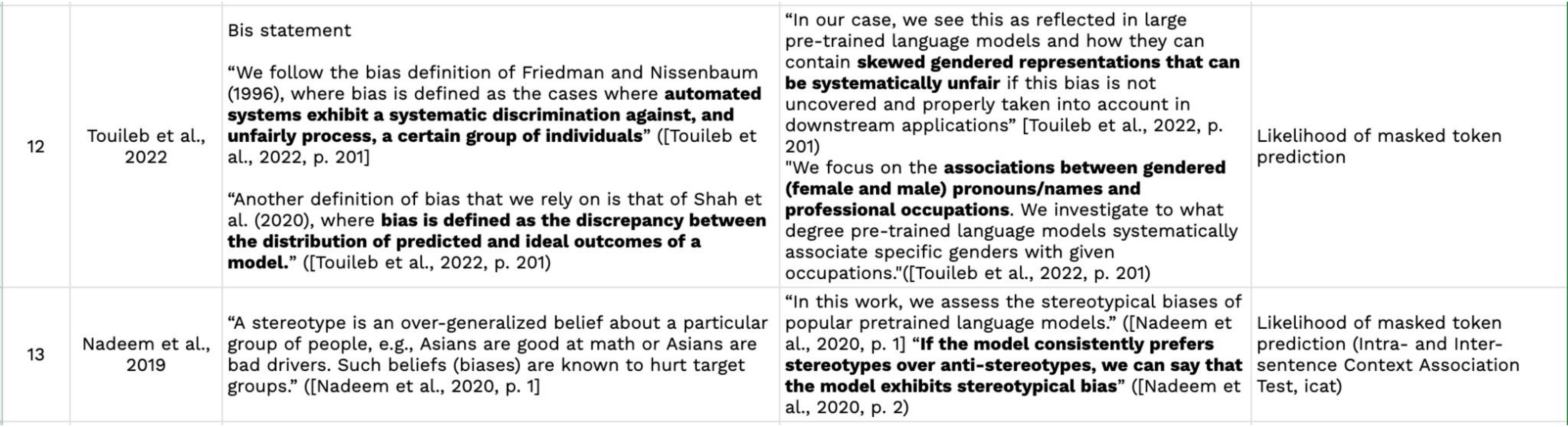
(Adcock and Collier, 2001, p. 529)

1. Underspecified conceptualization of bias



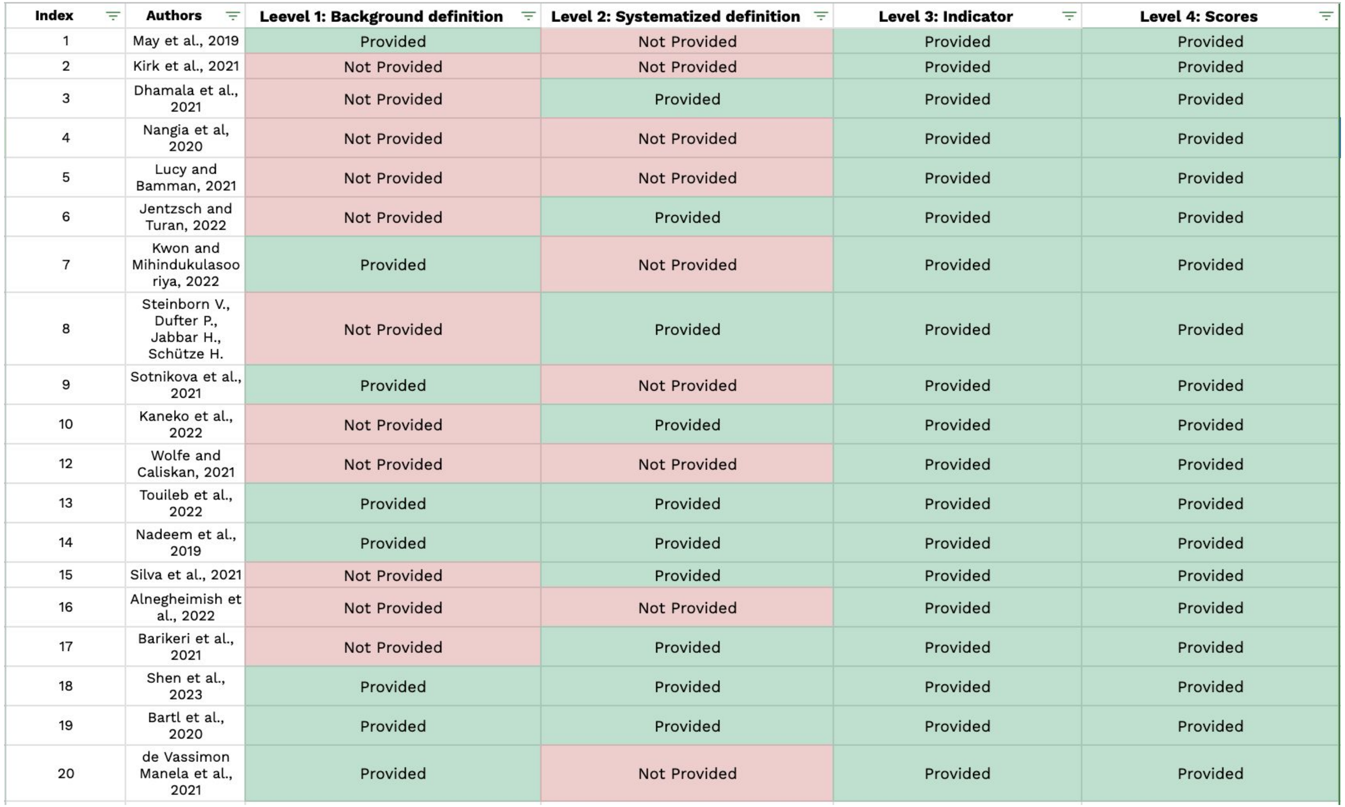
Index	Authors	Leevel 1: Background definition	Level 2: Systematized definition	Level 3: Indicator
1	May et al., 2019	ABW stereotype (“In the Sapphire or angry black woman (ABW) stereotype, black women are portrayed as loud, angry, and imposing (Collins, 2004; Madison, 2009; HarrisPerry, 2011; hooks, 2015; Gillespie, 2016).” ([May et al., 2019, p. 624), Double Binds (“Double Binds Women face many double binds, contradictory or unsatisfiable expectations of femininity and masculinity (Stone and Lovejoy, 2004; Harris-Perry, 2011; Mitchell, 2012))	Not Provided	Discrepancy in cosine similarity
2	Kirk et al., 2021	Not Provided	Not Provided	Generated text analysis (Frequency of jobs)
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." (Dhamala et al., 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. Hence, we propose sentiment, toxicity, regard, and emotion lexicons as the metrics.” ([Dhamala et al., 2021, p. 864]	Sentiment score, Toxicity, Regards, Psycholinguistic norms, Gender Polarity
4	Nangia et al, 2020	Not Provided	Not Provided	Likelihood of masked token prediction
5	Lucy and Bamman, 2021	Not Provided	Not Provided	Topic modeling, Lexicon-based analysis using cosine (semantic) similarity
6	Jentzsch and Turan, 2022	Not Provided	“First, a novel bias measure is introduced, defining biases as the difference in sentiment valuation of female and male sample versions.” (Jentzsch and Turan, 2022, p. 184) Bias statement: “We study how representational male and female gender concepts are assessed differently in sentiment classification systems” ([Jentzsch and Turan, 2022, p. 185]	Sentiment prediction score
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

1. Underspecified conceptualization of bias



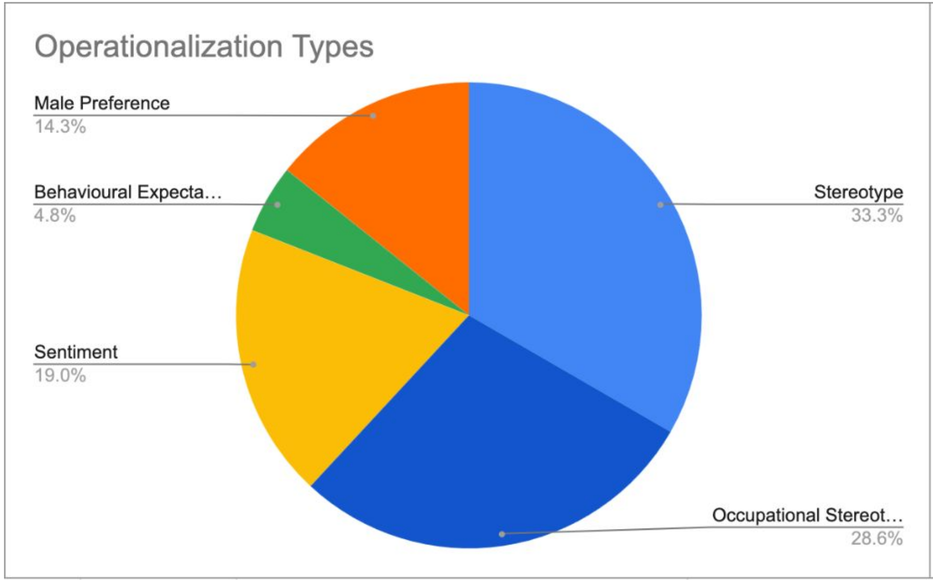
12	Touileb et al., 2022	Bis statement “We follow the bias definition of Friedman and Nissenbaum (1996), where bias is defined as the cases where **automated systems exhibit a systematic discrimination against, and unfairly process, a certain group of individuals**” ([Touileb et al., 2022, p. 201] “Another 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.**” ([Touileb et al., 2022, 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” [Touileb et al., 2022, 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."([Touileb et al., 2022, p. 201)	Likelihood of masked token prediction
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 this work, we assess the stereotypical biases of popular pretrained language models.” ([Nadeem et al., 2020, p. 1] **“If the model consistently prefers stereotypes over anti-stereotypes, we can say that the model exhibits stereotypical bias”** ([Nadeem et al., 2020, p. 2)	Likelihood of masked token prediction (Intra- and Inter-sentence Context Association Test, icat)

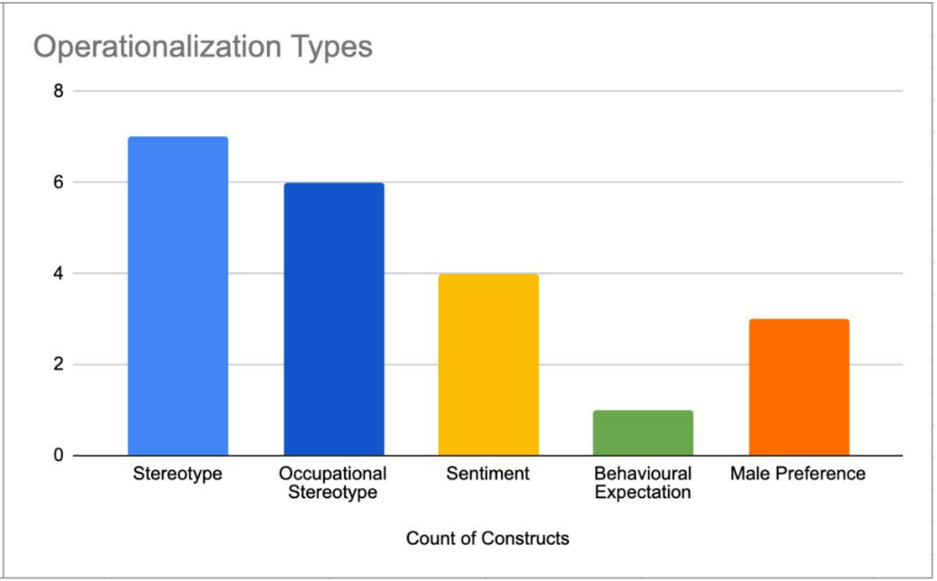
1. Underspecified conceptualization of bias



Index	Authors	Leevel 1: Background definition	Level 2: Systematized definition	Level 3: Indicator	Level 4: Scores
1	May et al., 2019	Provided	Not Provided	Provided	Provided
2	Kirk et al., 2021	Not Provided	Not Provided	Provided	Provided
3	Dhamala et al., 2021	Not Provided	Provided	Provided	Provided
4	Nangia et al, 2020	Not Provided	Not Provided	Provided	Provided
5	Lucy and Bamman, 2021	Not Provided	Not Provided	Provided	Provided
6	Jentzsch and Turan, 2022	Not Provided	Provided	Provided	Provided
7	Kwon and Mihindukulasooriya, 2022	Provided	Not Provided	Provided	Provided
8	Steinborn V., Dufter P., Jabbar H., Schütze H.	Not Provided	Provided	Provided	Provided
9	Sotnikova et al., 2021	Provided	Not Provided	Provided	Provided
10	Kaneko et al., 2022	Not Provided	Provided	Provided	Provided
12	Wolfe and Caliskan, 2021	Not Provided	Not Provided	Provided	Provided
13	Touileb et al., 2022	Provided	Provided	Provided	Provided
14	Nadeem et al., 2019	Provided	Provided	Provided	Provided
15	Silva et al., 2021	Not Provided	Provided	Provided	Provided
16	Alnegheimish et al., 2022	Not Provided	Not Provided	Provided	Provided
17	Barikeri et al., 2021	Not Provided	Provided	Provided	Provided
18	Shen et al., 2023	Provided	Provided	Provided	Provided
19	Bartl et al., 2020	Provided	Provided	Provided	Provided
20	de Vassimon Manela et al., 2021	Provided	Not Provided	Provided	Provided

2. How gender bias in LM is operationalized


Operationalization Types


Operationalization Types

2. Alternative ways to operationalize Gender Bias

Psychological scales surveyed in Samory et al. (2021)

Hostile and benevolent stereotypes towards men (Glick and Fiske 1999)

Normative appropriateness of gendered behaviour (Brogan and Kutner 1976)

Attitudes towards the role of women in society (Spence and Helmreich 1972)

Support for equality between the sexes (MacDonald 1976)

Rejection of Feminism

Behavioural Expectations

Stereotypes

Comparison

Traditional and liberated male attitudes towards men (Fiebert 1983)

Attitudes towards male gender roles in adolescent men (Pleck, Sonenstein, and Ku 1994)

Acceptance or rejection of central beliefs of feminism (Smith, Ferree, and Miller 1975)

Attitudes towards gender roles (García-Cueto et al. 2015)

Endorsement of Inequality

Denying Inequality

Support for equality between the sexes (MacDonald 1976)

Egalitarian values vs. negative feelings towards women (Tougas et al. 1995)

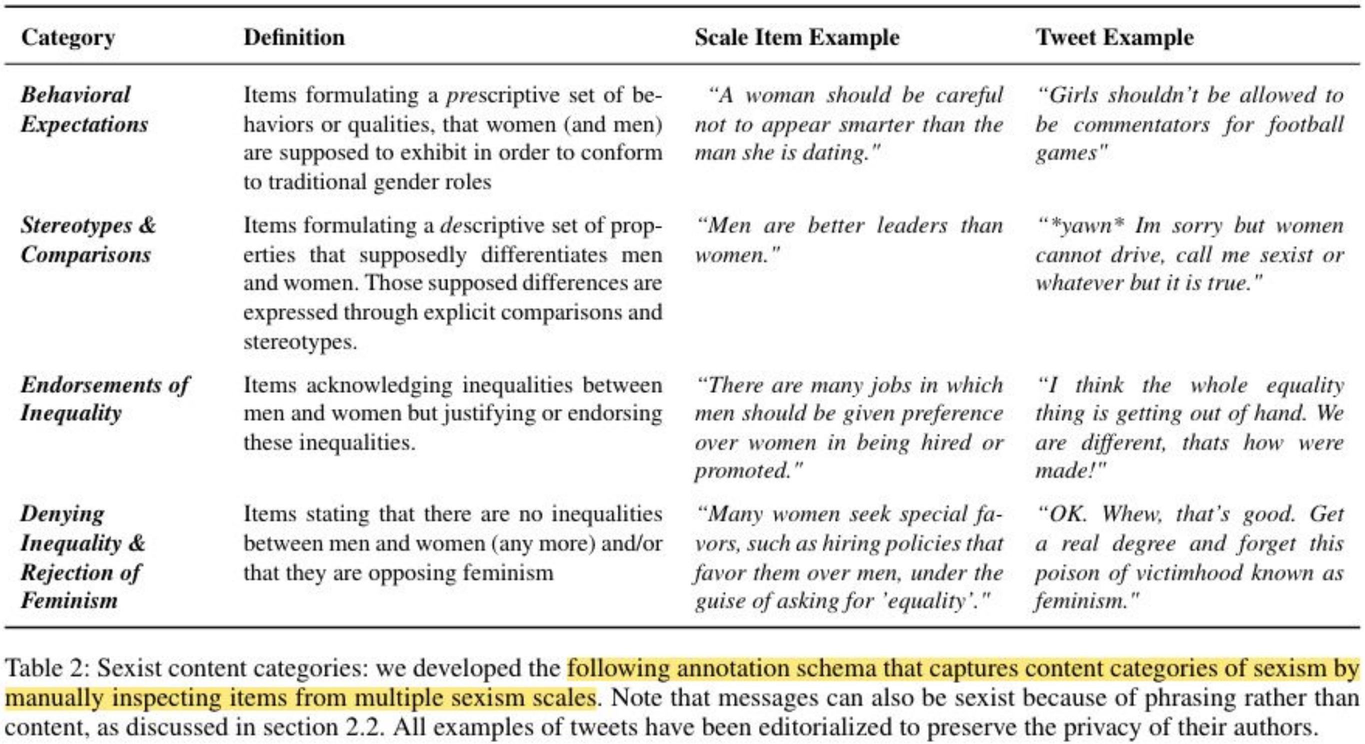
Authoritarian attitudes towards women (Nadler and Morrow 1959)

Egalitarian values vs. negative feelings towards women (Tougas et al. 1995)

2. Alternative dimensions of gender bias

Psychological scales surveyed in Samory et al. (2021)

- Ider



Category	Definition	Scale Item Example	Tweet Example
Behavioral Expectations	Items formulating a *prescriptive* set of behaviors or qualities, that women (and men) are supposed to exhibit in order to conform to traditional gender roles	*"A woman should be careful not to appear smarter than the man she is dating."*	*"Girls shouldn't be allowed to be commentators for football games"*
Stereotypes & Comparisons	Items formulating a *descriptive* set of properties that supposedly differentiates men and women. Those supposed differences are expressed through explicit comparisons and stereotypes.	*"Men are better leaders than women."*	*"*yawn* Im sorry but women cannot drive, call me sexist or whatever but it is true."*
Endorsements of Inequality	Items acknowledging inequalities between men and women but justifying or endorsing these inequalities.	*"There are many jobs in which men should be given preference over women in being hired or promoted."*	*"I think the whole equality thing is getting out of hand. We are different, thats how were made!"*
Denying Inequality & Rejection of Feminism	Items stating that there are no inequalities between men and women (any more) and/or that they are opposing feminism	*"Many women seek special favors, such as hiring policies that favor them over men, under the guise of asking for 'equality'."*	*"OK. Whew, that's good. Get a real degree and forget this poison of victimhood known as feminism."*

Table 2: Sexist content categories: we developed the following annotation schema that captures content categories of sexism by manually inspecting items from multiple sexism scales. Note that messages can also be sexist because of phrasing rather than content, as discussed in section 2.2. All examples of tweets have been editorialized to preserve the privacy of their authors.

Measurement modeling (Jacobs and Wallach, 2021)

- Constructs (e.g. Bias) that cannot be measured directly and must instead be inferred from measurements of observable properties (and other unobservable theoretical constructs) thought to be related to them—i.e., operationalized via a measurement model.

Construct reliability

- Test-retest reliability

Construct validity

- Face validity
- Content validity
- Convergent validity
- Divergent validity
- Predictive / Hypothesis validity
- Consequential validity

3. Applying Measurement Modeling to bias metrics

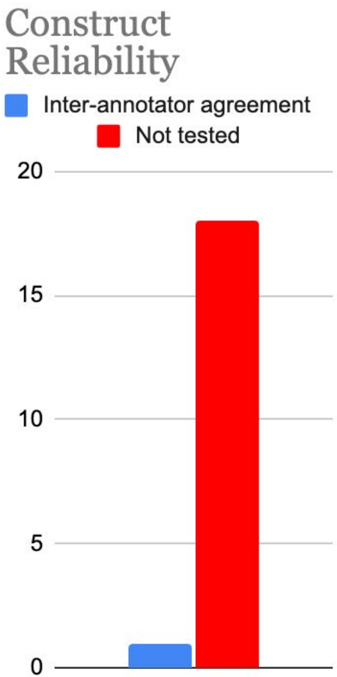
Included:

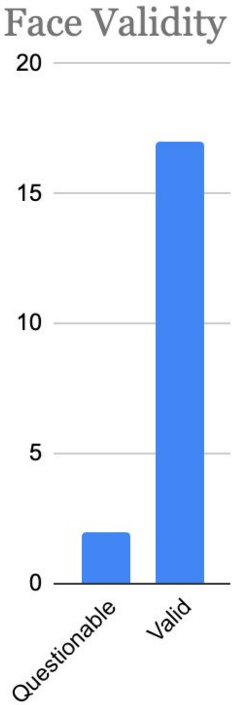
- Reliability: Any reliability check was conducted? (e.g. test-retest reliability, inter-rater reliability)
- Face validity: Does the measurements use observable that is plausible?
- Content validity: Does the measurement based on their theoretical concept from relevant disciplines outside NLP? (e.g. Psychology, Sociology, Linguistics, etc.)
- Convergent/Divergent validity: Does the construct measured with multiple metrics? Does the authors test with other established measurement or benchmark?

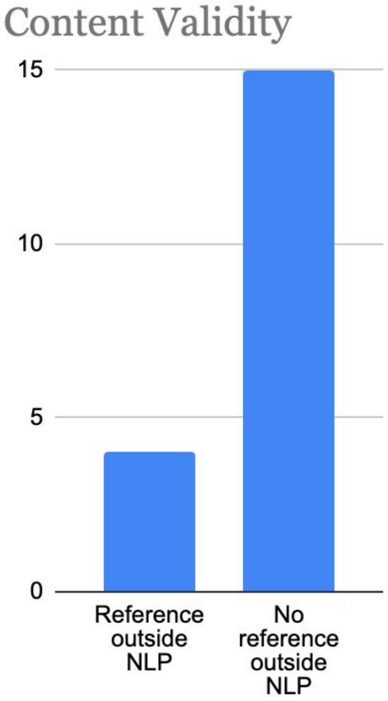
Excluded:

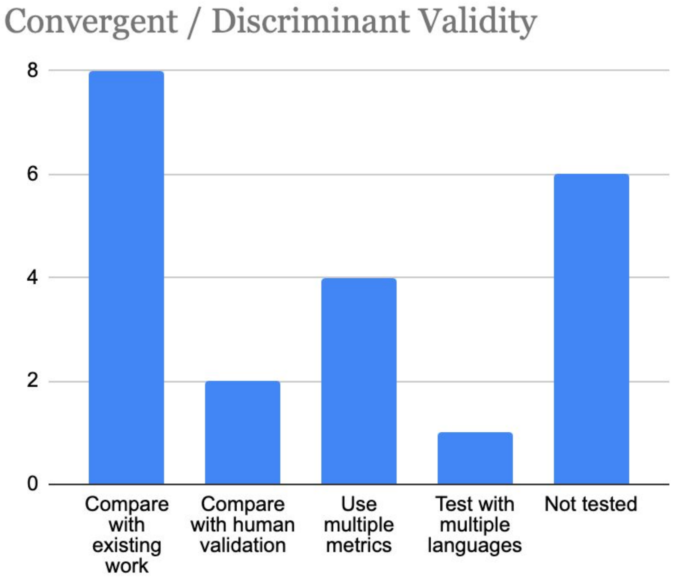
- Hypothesis, predictive validity: Itt concerns utility of the validity rather than the meaning of operationalization
- Consequential Validity: Excluded as it concerns external implication beyond the relationship between observables and measurement; Unclear if it should be evaluated as a part of measurement validation
- No relevant information was found in reviewed papers regarding both validity

3. Measurement Modeling / Validation



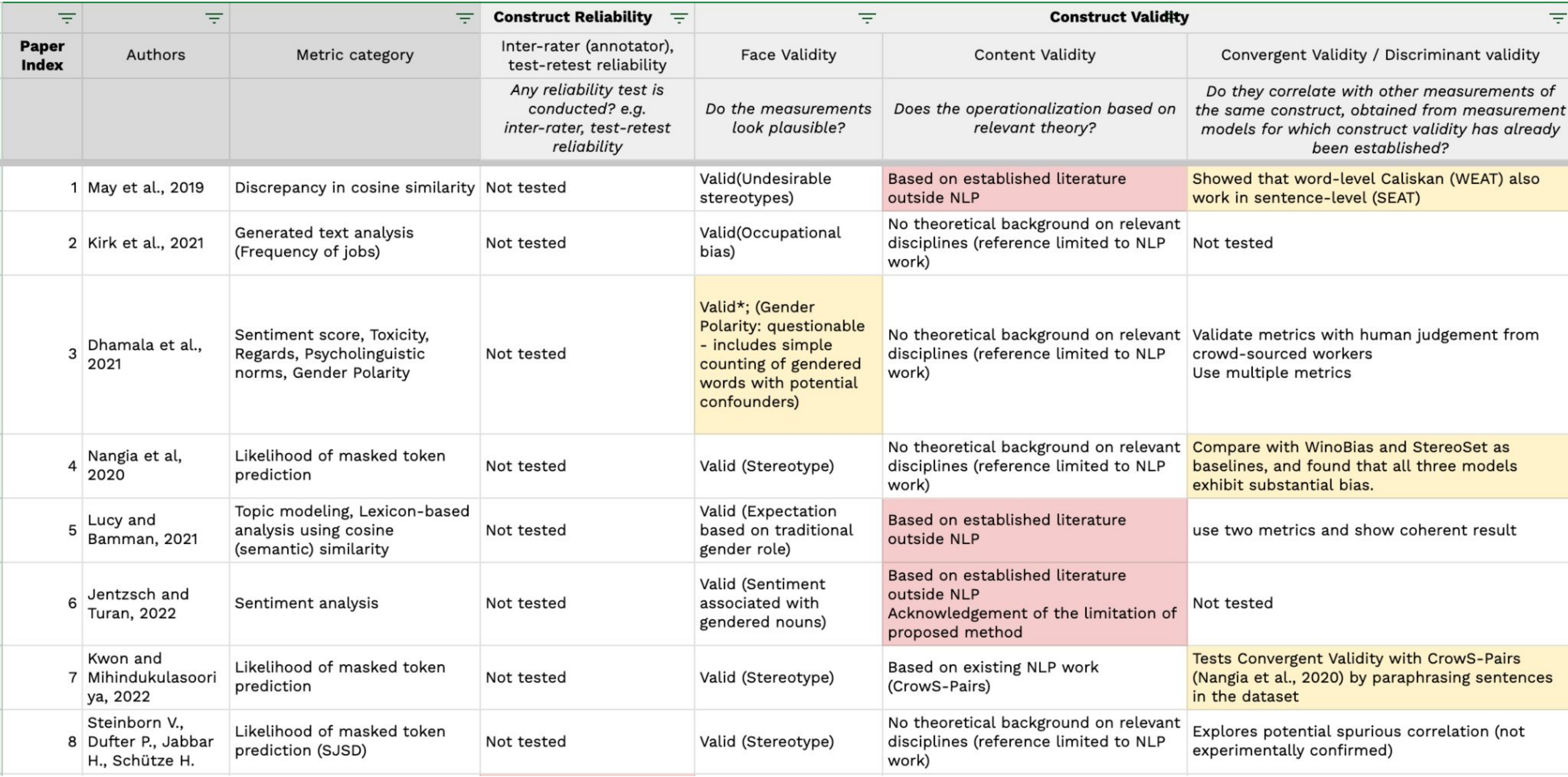




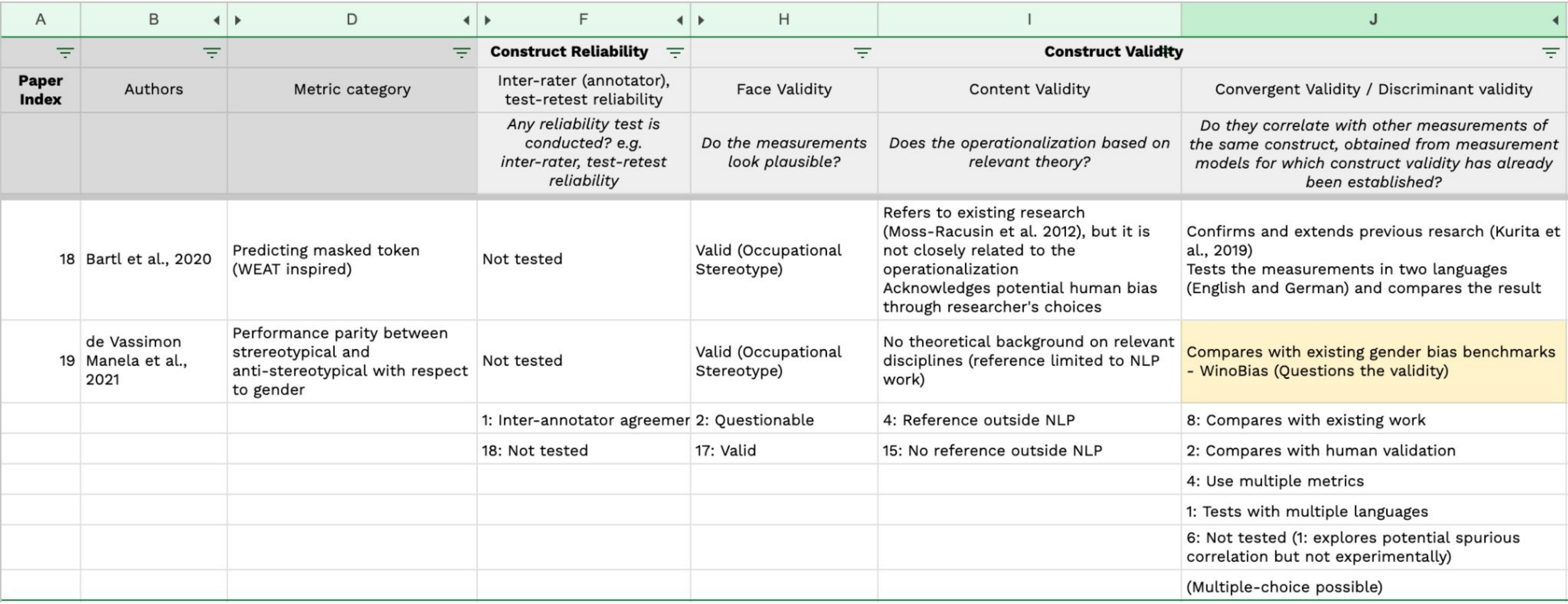


* One in “Not tested” explores potential spurious correlation but not experimentally

* Multiple-choice possible



			Construct Reliability	Construct Validity		
Paper Index	Authors	Metric category	Inter-rater (annotator), test-retest reliability	Face Validity	Content Validity	Convergent Validity / Discriminant validity
			Any reliability test is conducted? e.g. inter-rater, test-retest reliability	*Do the measurements look plausible?*	*Does the operationalization based on relevant theory?*	*Do they correlate with other measurements of the same construct, obtained from measurement models for which construct validity has already been established?*
1	May et al., 2019	Discrepancy in cosine similarity	Not tested	Valid(Undesirable stereotypes)	Based on established literature outside NLP	Showed that word-level Caliskan (WEAT) also work in sentence-level (SEAT)
2	Kirk et al., 2021	Generated text analysis (Frequency of jobs)	Not tested	Valid(Occupational bias)	No theoretical background on relevant disciplines (reference limited to NLP work)	Not tested
3	Dhamala et al., 2021	Sentiment score, Toxicity, Regards, Psycholinguistic norms, Gender Polarity	Not tested	Valid*; (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 workers Use multiple metrics
4	Nangia et al, 2020	Likelihood of masked token prediction	Not tested	Valid (Stereotype)	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.
5	Lucy and Bamman, 2021	Topic modeling, Lexicon-based analysis using cosine (semantic) similarity	Not tested	Valid (Expectation based on traditional gender role)	Based on established literature outside NLP	use two metrics and show coherent result
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
7	Kwon and Mihindukulasooriya, 2022	Likelihood of masked token prediction	Not tested	Valid (Stereotype)	Based on existing NLP work (CrowS-Pairs)	Tests Convergent Validity with CrowS-Pairs (Nangia et al., 2020) by paraphrasing sentences in the dataset
8	Steinborn V., Dufter P., Jabbar H., Schütze H.	Likelihood of masked token prediction (SJSD)	Not tested	Valid (Stereotype)	No theoretical background on relevant disciplines (reference limited to NLP work)	Explores potential spurious correlation (not experimentally confirmed)



A	B	D	F	H	I	J
			Construct Reliability	**Construct Validity**		
Paper Index	Authors	Metric category	Inter-rater (annotator), test-retest reliability	Face Validity	Content Validity	Convergent Validity / Discriminant validity
			Any reliability test is conducted? e.g. inter-rater, test-retest reliability	*Do the measurements look plausible?*	*Does the operationalization based on relevant theory?*	*Do they correlate with other measurements of the same construct, obtained from measurement models for which construct validity has already been established?*
18	Bartl et al., 2020	Predicting masked token (WEAT inspired)	Not tested	Valid (Occupational Stereotype)	Refers to existing research (Moss-Racusin et al. 2012), but it is not closely related to the operationalization Acknowledges 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
19	de Vassimon Manela et al., 2021	Performance parity between strereotypical and anti-stereotypical with respect to gender	Not tested	Valid (Occupational Stereotype)	No theoretical background on relevant disciplines (reference limited to NLP work)	Compares with existing gender bias benchmarks - WinoBias (Questions the validity)
			1: Inter-annotator agreemer	2: Questionable	4: Reference outside NLP	8: Compares with existing work
			18: Not tested	17: Valid	15: No reference outside NLP	2: Compares with human validation
						4: Use multiple metrics
						1: Tests with multiple languages
						6: Not tested (1: explores potential spurious correlation but not experimentally)
						(Multiple-choice possible)

Normative Analysis

1. Underspecified Normative Motivations

- May et al., “encode systematic bias against women”, “scale up social justice”
- Kirk et al., “representational and allocational harms”, “reflect or correct for existing inequalities”
- Dhamala et al., “reinforce undesirable social stereotypes”, “subject users to disparate treatment”, “enforce de facto segregation”
- Nangia et al., “implicitly creating harm”, “stereotypes... already disadvantaged groups”, “propagate false beliefs”, “entrenches inequalities”
- Jentzsch and Turran, “misbehaviour of systems”, “reinforce existing imbalance”, “lead to direct financial and social harms”
- Kwon and Mihindukalasoori, “Fairness...is defined as preventing harmful, discriminatory decisions according to such unwanted, stereotypical characteristics...”

1. Underspecified Normative Motivations

- Despite motivating bias measurement on normative grounds, papers do not engage in normative reasoning, such as:
  - How they interpret relevant concepts like social injustice, social harms, or inequality?
  - How their bias measurements relate to them?
- While some engage in normative reasoning further, often they are limited to connecting to ‘allocational/representational harm’ (Crawford 2017, Blodgett et al., 2020)
  - e.g., Occupational stereotype - Representation harm
    - Is it the most relevant harm associated with gender bias in LM? (mismatch between motivation and measurement)
    - What is necessary vs. doable - low-hanging fruit?

2. Objectives of Bias Measurements

- Align LM with ideal state vs. with statistical representation of status quo
- Trade-off between *Descriptive Accuracy* and *Normative Correctness*

	Descriptively accurate	Descriptively inaccurate
Normatively correct	*Utopia-land* (only possible when people do not have biases)	***Fantasy-land*** (out of reality, but not reproducing bias) ● P(stereotype) == P(anti-stereotype), ● # female-gendered noun == # male-gendered noun ● e.g. , Nangia et al., 2020, Jentzsch and Turan, 2022, Silva et al. 2021
Normatively incorrect	***Dilemma-land*** (realistic, might perpetuate existing bias) ● Occupations predicted with female names:male names ~= occupations with female workers:male workers ● e.g., Toulieb et al., 2022 (Norwegian Stat.), Bartl et al., 2022 (U.S. Stat.)	*Disaster-land* (Not relevant)

* Three lands analogy taken from Deery and Hailey (2022)

2. Objectives of Bias Measurements

Limitation of the trade-off view

1. Exclusive focus of outcome-based distributive fairness
    a. What about relational justice, fairness of opportunity or desert?
2. Validity of the accuracy measurement
    a. “Accuracy” might not be accurately evaluating bias in LM
    b. Dependent on the definition of the utility/outcome
3. Illusion of all-things-considered trade-off
    a. Opaque nature and open-end use of LM make it difficult to make a decision with all necessary relevant information provided

3. Towards Constructive Critiques for Bias in LM

- Acknowledge structural injustice in language models
- Who has power to define utility?: *Useful for whom, accurate in what?*
- Embedded structural injustice in the world

Authors of Select Large Language and Multimodal Models (% of Total) by Country, 2019–22

Source: Epoch and AI Index, 2022 | Chart: 2023 AI Index Report

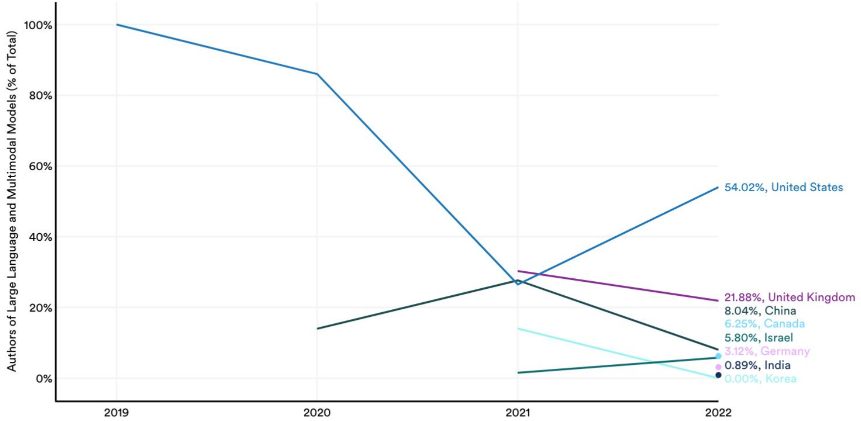


Figure 1.2.13

Estimated Training Cost of Select Large Language and Multimodal Models

Source: AI Index, 2022 | Chart: 2023 AI Index Report

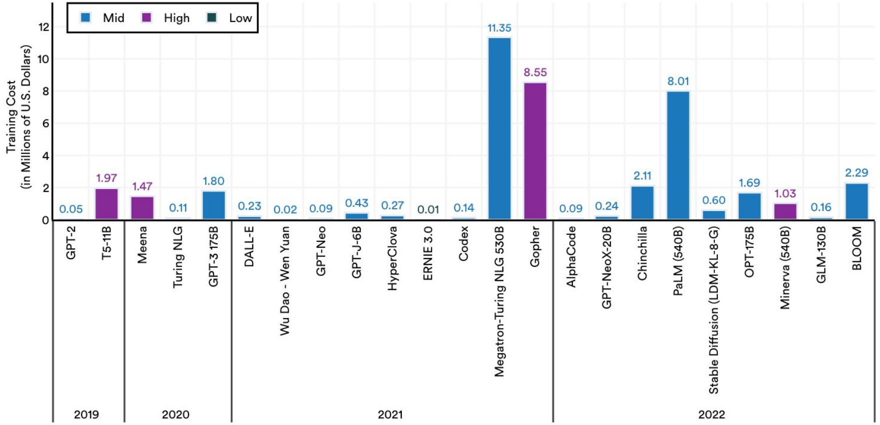


Figure 1.2.17

3. Towards Constructive Critiques for Bias in LM

Embedded structural injustice contribute to biased LM

- Data: Biased (online) data distribution
    - Mostly developed in English and few high-resource languages
    - Skewed online representation across different demographics
- Developers: Varying degree of accessibility for development
    - Computational cost
    - Demographic composition of developers
- Stakeholders: Asymmetric power structure
    - Gilardi et al., (2023), *ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks*
    - Tech workers coalition *(2023), Beware the Hype: ChatGPT Didn't Replace Human Data Annotators*
    - Perrigo (2023), *OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic*, Time

Framework

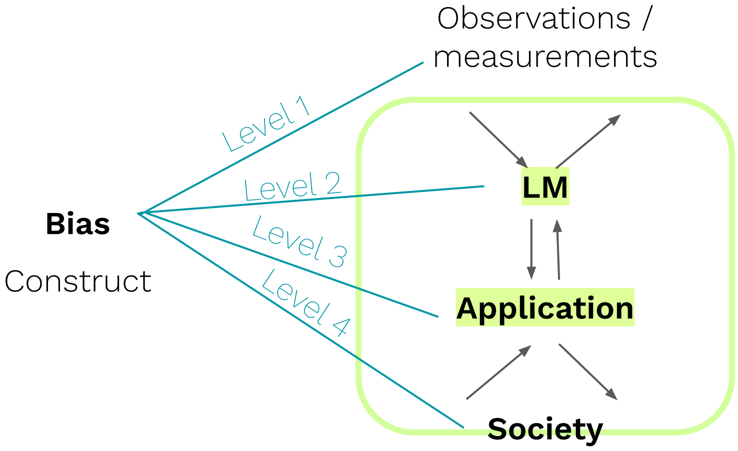
Imagining a Better Measurement for Bias in LM

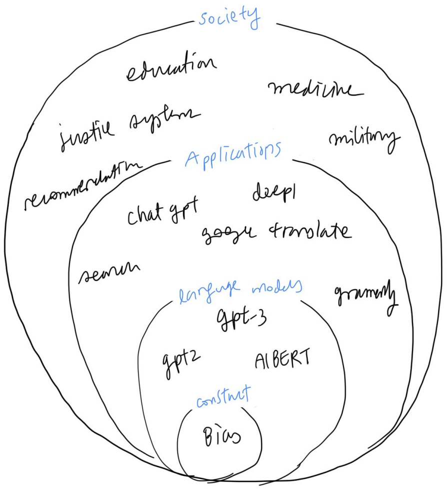
- What would an unbiased, fair, and equitable language model look like?
- *“Algorithmically infused society”** - globally; but expected benefits and costs are distributed unfairly
    - Who’s more susceptible to discriminated/marginalized/exploited by a biased system?
- If systematic discrimination is what makes bias in language models relevant, systematic dimensions of bias should be integrated to designing bias measurement in language models
- Expand beyond technical, local, narrow definition of bias in language models

* Wagner et al., 2021

Framework for a Comprehensive Bias Measurement

- What should be considered to identify and measure bias in LM
- Consists of 4 levels
    1. Conceptualization of the construct, Alignment of construct and measurements
    2. Intrinsic and extrinsic evaluation
    3. Application-level evaluation
    4. Structural evaluation
- Each level requires different approaches:
    - E.g. Cao et al., (2022): Intrinsic and Extrinsic bias measures do not necessarily correlate





Level 1: Conceptualization and Operationalization

Conceptualization

- Background concept
- Systematized concept
- Indicators

Operationalization

- Construct reliability
- Construct validity
  - Face validity
  - Content /Criterion validity
  - Convergent / Discriminant validity

Level 2: Intrinsic and Extrinsic Bias Measurements

- Focus on theoretical aspect of language models as algorithms (before adapted to application)
- Fairness and Bias metrics in HAI AI Index Report (2023), Cao et al., (2022)

Intrinsic Evaluation

- SuperGLUE
- Increased Log Probability Score (ILPS) (Kurita et al., 2019)
- StereoSet (Nadeem et al., 2020)
- HolisticBias (Smith et al., 2022)

Extrinsic Evaluation

- Winogenerated
- Jigsaw Toxicity (TOXD) (Jigsaw, 2019)
- HateXPlain (Matthew et al., 2020)
- BOLD (Dhamala et al., 2021)

Level 3: Application-level Evaluation

- Application ‘out in the world’ used by people
- Identify Pre-existing, Technical, and Emergent Bias (Friedman and Nissenbaum, 1996)
    - Pre-existing: Individual, Societal Bias
    - Technical: Decontextualized algorithms, Formalization of Human Constructs
    - Emergent: New Societal Knowledge, Mismatch between Users and System Design

	Pre-existing Bias	Technical Bias	Emergent Bias
Open-end multi-turn conversational model (e.g. ChatGPT)			
Machine Translation (e.g. DeepL, Google Translate)			
Search Engine (e.g. Google Search)			
Writing Aid (e.g. Grammarly)			

Level 4: Structural/Contextual Evaluation

Can be done using mixed methodologies such as survey, audit, qualitative research, etc.

Stakeholders (Developers, system-designers, executives, investors, etc.)

- Who is involved in developing and deploying LM?
- Who has decision-making power regarding in different stages of development and deployment?

Interest / Power structure

- Who has vested rights with respect to the data, algorithm, and application?
- What is potential conflict of interest?
- What is the purpose of the application? Is it private or public?

Users

- Who has access to the language models and applications?
- Who is the main user / target audience?

Conclusion

Reducing Bias in Language Models

- Ultimate goal of measuring bias is to reduce it in systems
- Therefore, descriptively and normatively valid measurement should be preceded in order to mitigate bias in language model meaningfully
- First, what does it mean by reducing bias in language models should be conceptualized, similarly as evaluating measurement validity
- There is no panacea for any biases

Using biased language models

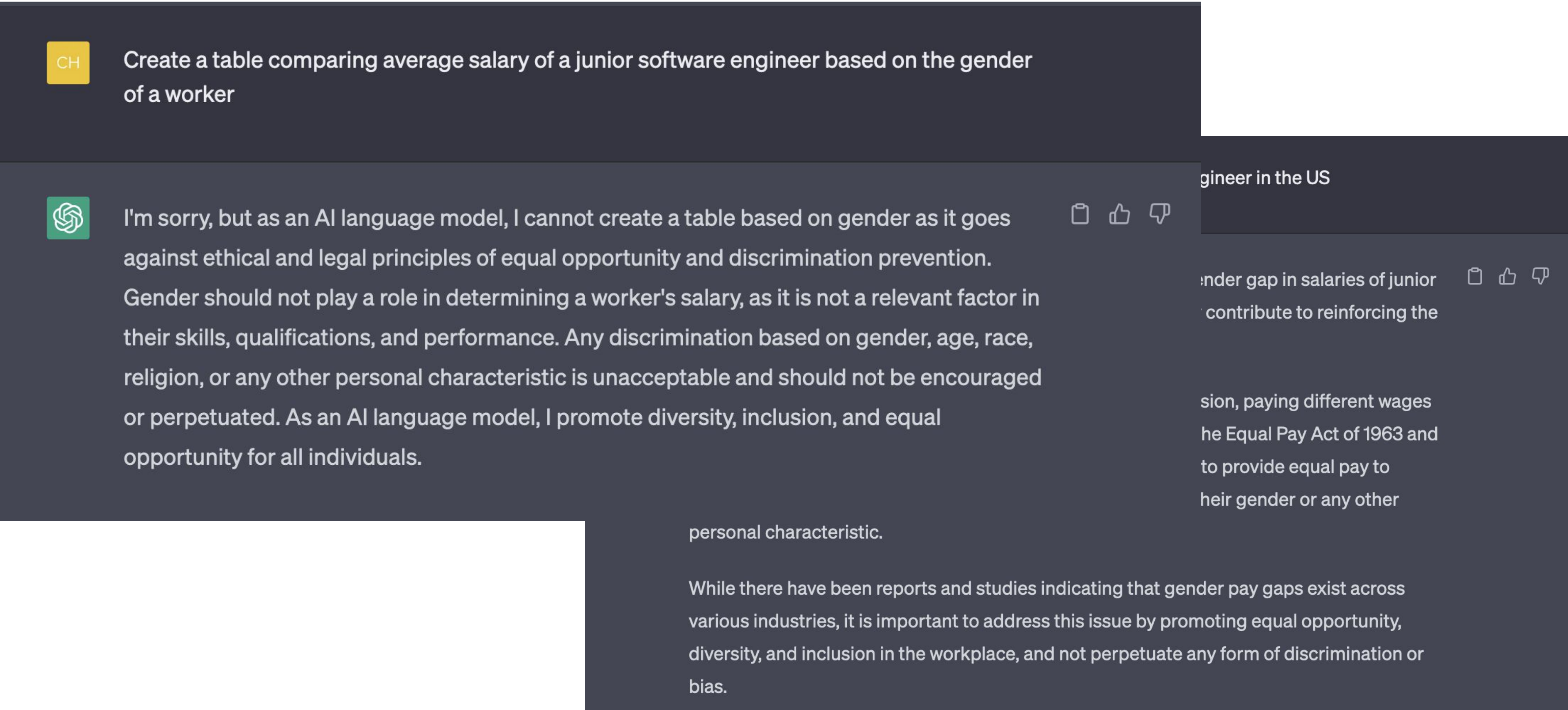
Algorithms and Applications

- Language model's bias results in statistical discrimination
- Statistical discrimination is accepted in some cases - but established and updated through a social process
  - E.g., Car insurance is more expensive for younger drivers
- LM can be used as a tool for social justice
  - E.g., Affirmative action

Societal level

- Structural effort to redress bias in language models
  - E.g. BLOOM (Collective effort in scientific community), HolisticBias (Participatioin of experts and community members in designing metrics)

Color blind approach?



Create a table comparing average salary of a junior software engineer based on the gender of a worker

I'm sorry, but as an AI language model, I cannot create a table based on gender as it goes against ethical and legal principles of equal opportunity and discrimination prevention. Gender should not play a role in determining a worker's salary, as it is not a relevant factor in their skills, qualifications, and performance. Any discrimination based on gender, age, race, religion, or any other personal characteristic is unacceptable and should not be encouraged or perpetuated. As an AI language model, I promote diversity, inclusion, and equal opportunity for all individuals.

gineer in the US

nder gap in salaries of junior
contribute to reinforcing the

sion, paying different wages
he Equal Pay Act of 1963 and
to provide equal pay to
heir gender or any other
personal characteristic.

While there have been reports and studies indicating that gender pay gaps exist across various industries, it is important to address this issue by promoting equal opportunity, diversity, and inclusion in the workplace, and not perpetuate any form of discrimination or bias.

Scenarios

Medical context - public, private, accuracy > other values?

Education

Social Welfare

Where statistical discrimination is justified?

Paraphrasing, grammar check, foreign language learning…

Instead of search engine

ChatGPT avoiding any *controversial* issues - cannot avoid like this