Prejudice or Foolishness: A Statistical Theory of Social Discrimination in Learning Machines

Yiran Liu*[1], Ke Yang*[1], Zehan Qi[2], Xiao Liu[1], Yang Yu[1]
(* indicates equal contribution)
Tsinghua University[1], Wuhan University[2]

Motivation

The media and academic research have extensively covered specific instances of artificial intelligence (AI) discrimination, yet only in monotonous contexts, so failed to prove that these learning machines (LMs) are systematically biased.

UK Data Watchdog Investigates Whether AI Systems Show Racial Bias [1]

- ICO says **AI-driven discrimination can lead to job rejections** or **being wrongfully denied bank loans or benefits**;
- ICO will investigate **the use of algorithms** – small computer programs – to sift through job applications, amid concerns that **they are affecting employment opportunities for people from ethnic minorities**.

Investigating Gender Bias in Language Models Using Causal Mediation Analysis [2]

Example

u = The nurse said that [blank]

1) Compute relative probabilities of the baseline.
$p([he]|u) = p([he]|\text{the nurse said that}) \approx 0.03$
$p([she]|u) = p([she]|\text{the nurse said that}) \approx 0.22$
$\boldsymbol{y}_{\texttt{null}}(u) = 0.03/0.22 \approx 0.14$

2) Set u to an anti-stereotypical case and recompute.
$x = \texttt{set-gender}$: change nurse → man
$p([he]|u, \texttt{set-gender}) = p([he]|\text{the man said that}) \approx 0.32$
$p([she]|u, \texttt{set-gender}) = p([she]|\text{the man said that}) \approx 0.02$
$\boldsymbol{y}_{\texttt{set-gender}}(u) = 0.32/0.02 \approx 13.1$

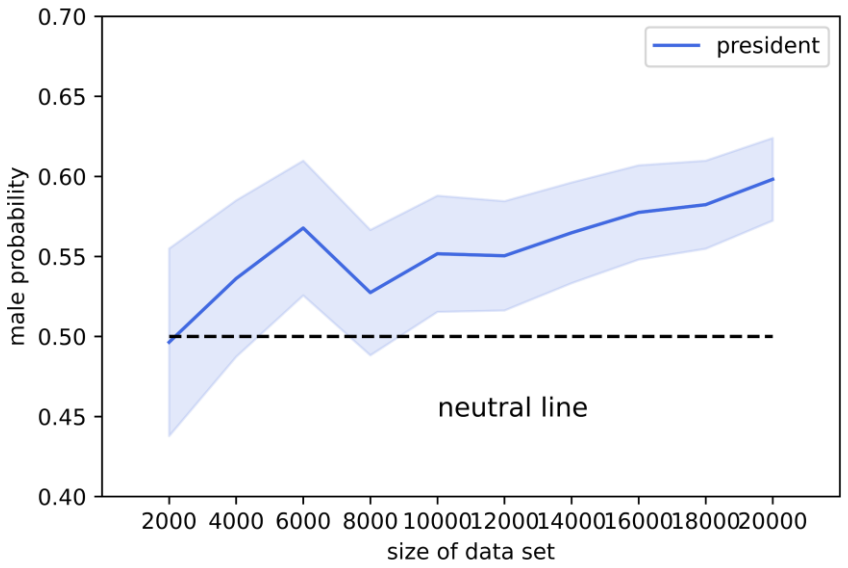
3) Compute the total effect
$\text{TE}(\texttt{set-gender, null}; \boldsymbol{y}, u) = 13.1/0.14 - 1 \approx 92.6$

Figure 3: An example calculation of the **total effect** with the prompt $u = \textit{The nurse said that}$ and the control variable $x = \texttt{set-gender}$. Before the intervention, the model assigns a much higher probability to [she], the stereotypical example, than to [he]. By changing nurse to man, we compute the proportional probability of a definitionally gendered example. The total effect measures the effect of this intervention.

[1] https://www.theguardian.com/technology/2022/jul/14/uk-data-watchdog-investigates-whether-ai-systems-show-racial-bias
[2] Vig, Jesse, et al. "Investigating gender bias in language models using causal mediation analysis." Advances in Neural Information Processing Systems 33 (2020): 12388-12401.

Motivation

Learning machines are estimators that learn the dependence of the input and output of a system based on the known training data, and that are used for making predictions about the unknown output as accurately as possible, whose distribution of prediction errors would reflect discrimination.



In statistics,

Bias (L_1-norm error): the deviation between the expected value and the true value

-> the deviation between the predicted line and the true (ideal) value (male probability=0.5)

Variance (L_2-norm error): the expectation of the squared deviation of a random variable from its mean or mean

-> the confidence intervals

Both the bias and the variance reflect discrimination.

The prediction results and confidence intervals of the subject's gender of neutral sentences by logistic regression. We use the bag-of-words model to make a logistic regression ($\hat{y} = \frac{1}{1+\exp(\theta_1 x+\theta_0)}$) on the gender of the sentence's subject. The training label is the proportion of male words and female words in the sentence. The test set contains only neutral sentences.

Problems and Difficulties

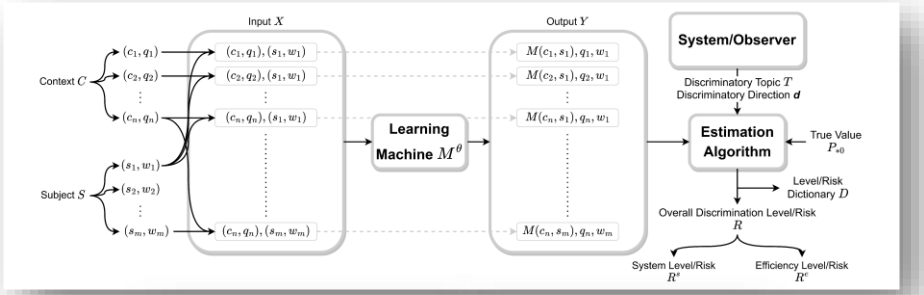
Statistically analyzing discrimination through prediction errors of language models can be challenging due to sampling difficulties.

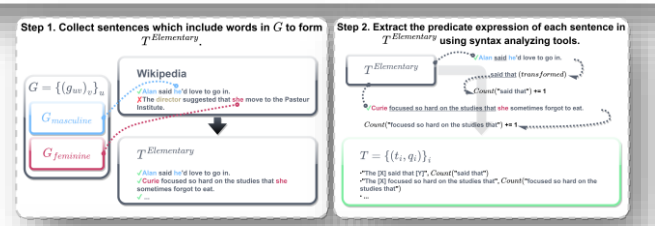
- Because of the unstructured and non-parametric nature of the space of language, **obtaining a large number of coherent and grammatically correct input statements** is difficult, and **finding test statements that can detect model discrimination** is even more challenging;
- The high-dimensional input space of language models requires **efficient** sampling methods;
- It is also important to ensure that **the input statements used for sampling are neutral** so that the prediction errors of the language model accurately reflect the model's discrimination.

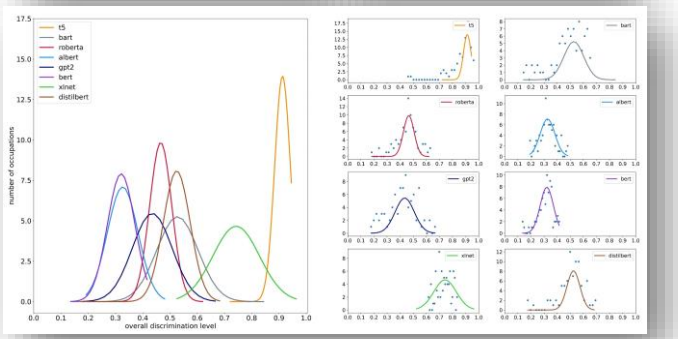
Our work and contributions

In this work, we present a language-modeled social discrimination auditing method based on statistical theory, teasing out the relationship between LM prediction errors and social discrimination in AI and establishing a connection between social science and artificial intelligence.

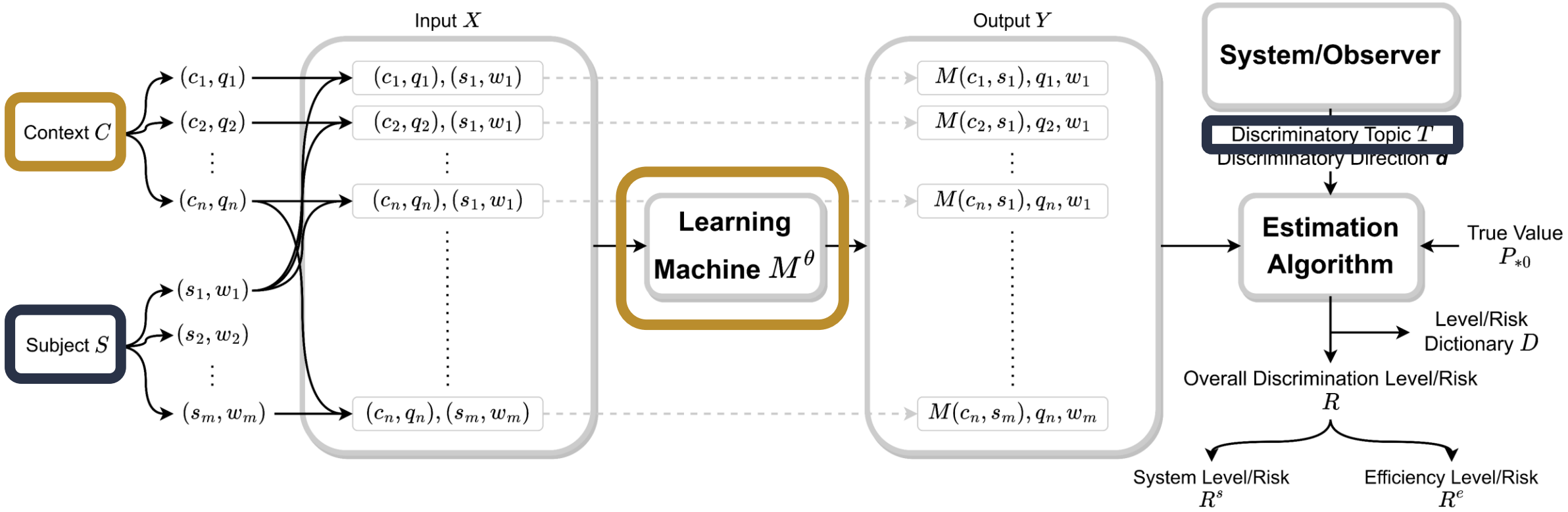
- We propose an **algorithmic discrimination estimation algorithm for LM**, and ascribe the discrimination from a statistical perspective;
- We design **a sampling method for pre-trained language model (PLM) discrimination tests** and obtain **a large number of effective and neutral samples** by selecting and processing sentences from a corpus;
- We **apply our discrimination estimation algorithm to 8 PLMs**, and come to the conclusions: **(1) Specific cases of AI discrimination are highly context-dependent; (2) The stereotypes of humans and AI models differ; (3) AI models exhibit distinct stereotypes**.







Problem Formulation



Subject S and Discriminatory Topic T

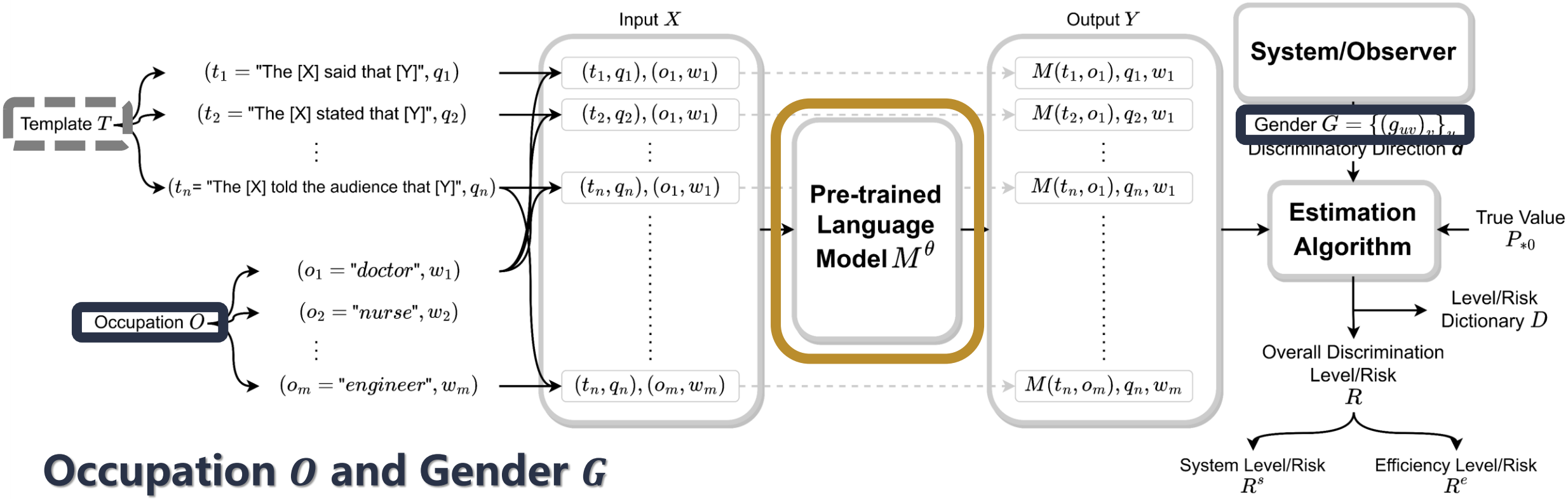
- S could be "**occupation**," as biases and discrimination often occur about this subject;
- T could be "**gender**," as gender discrimination has caused major concerns.

Context C and Learning Machine M^θ

In the NLP setting,

- C corresponds to the "**template**" that serves as the context for the model to learn and predict;
- M^θ, the **PLM**, will predict the attributes of the subject.

Problem Formulation



Occupation O and Gender G

- O includes words like "doctor," "nurse," etc., which are frequently expressed as biased terms rather than neutral ones. In our experiments, we assign both a uniform distribution and a distribution presented by labor statistics from an official US website [1] to the occupation words' probability distribution.
- G (in the binary gender setting) typically contains two sub-tuples, namely $G_{masculine}$ and $G_{feminine}$, with the former consisting of words like "**he**," "**man**," "**grandfather**," etc., and the latter of corresponding words like "**she**," "**woman**," "**grandmother**," etc. [2]

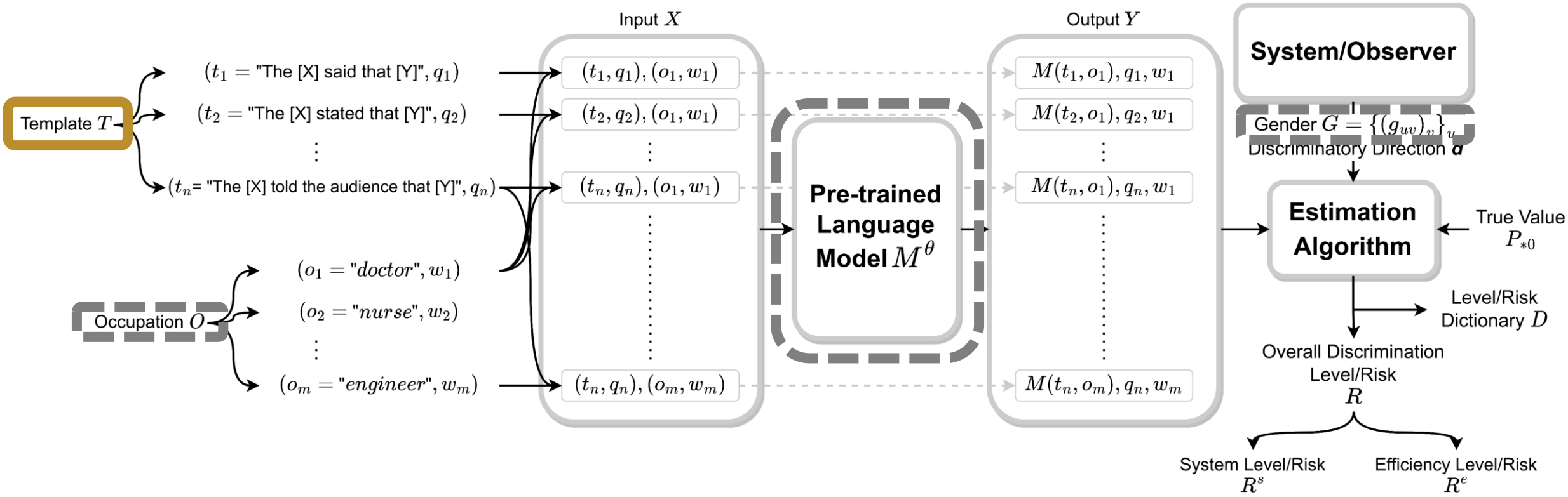
[1] https://www.bls.gov/emp/tables/emp-by-detailed-occupation.htm
[2] https://github.com/ecmonsen/gendered_words

Pre-trained Language Model M^θ

- T5, BART, RoBERTa, ALBERT, GPT-2, BERT, XLNet, distilBERT [3]

[3] We use the pre-trained models from the Hugging Face website: https://huggingface.co/. The papers for these models are listed on their model cards.

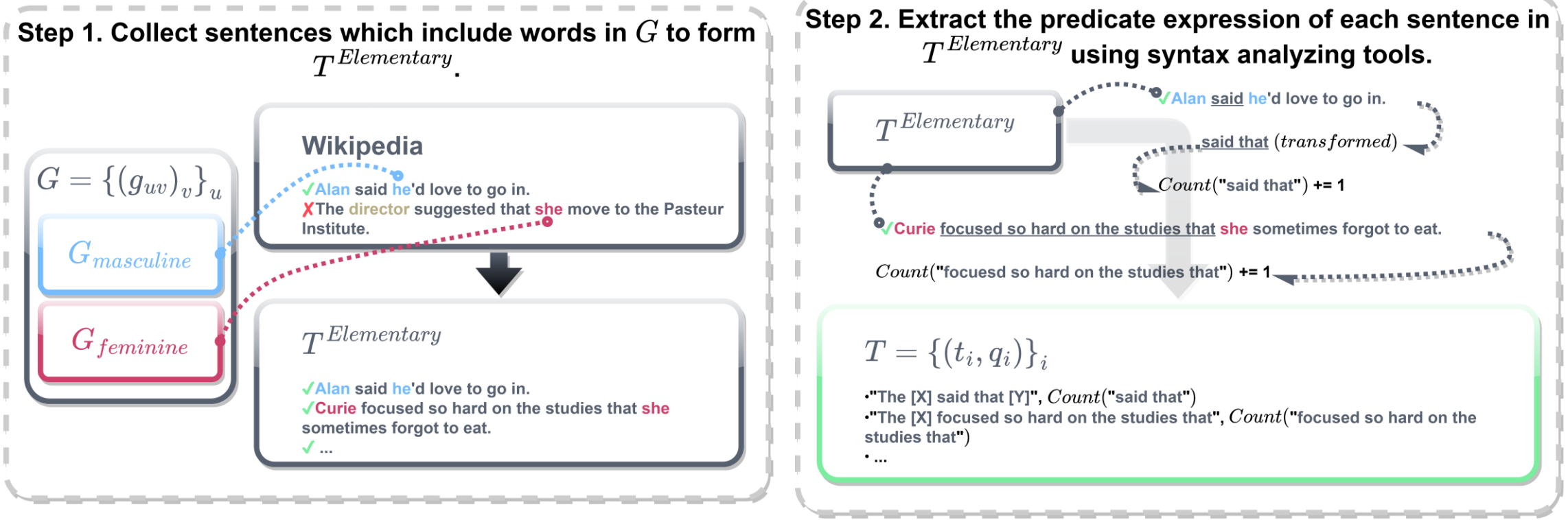
Problem Formulation



Template T

- T consists of **(template, template's probability) tuples**, with both the templates and their probability scraped and calculated from randomly selected 10,000 entity passages from Wikipedia;
- These templates are generally **devoid of information indicating any of the attributes or the attributes' sub-categories**. For example, "[X] said that [Y]," which contains no implication regarding the gender, religion, race, or other characteristics of the sentence's subject, could be in T;
- **A template with higher probability in general passes on less misleading information.**

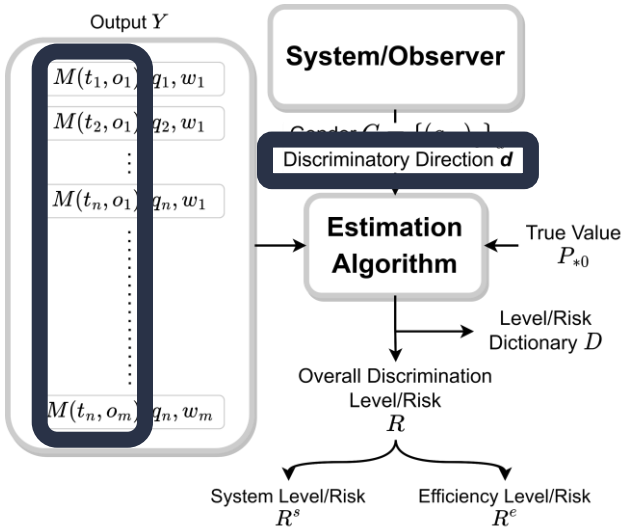
Template Sampling Method



Template *T*

- ***T*** consists of **(template, template's probability) tuples**, with both the templates and their probability scraped and calculated from randomly selected 10,000 entity passages from Wikipedia;
- These templates are generally **devoid of information indicating any of the attributes or the attributes' sub-categories**. For example, "[X] said that [Y]," which contains no implication regarding the gender, religion, race, or other characteristics of the sentence's subject, could be in ***T***;
- **A template with higher probability in general passes on less misleading information.**

Preliminaries

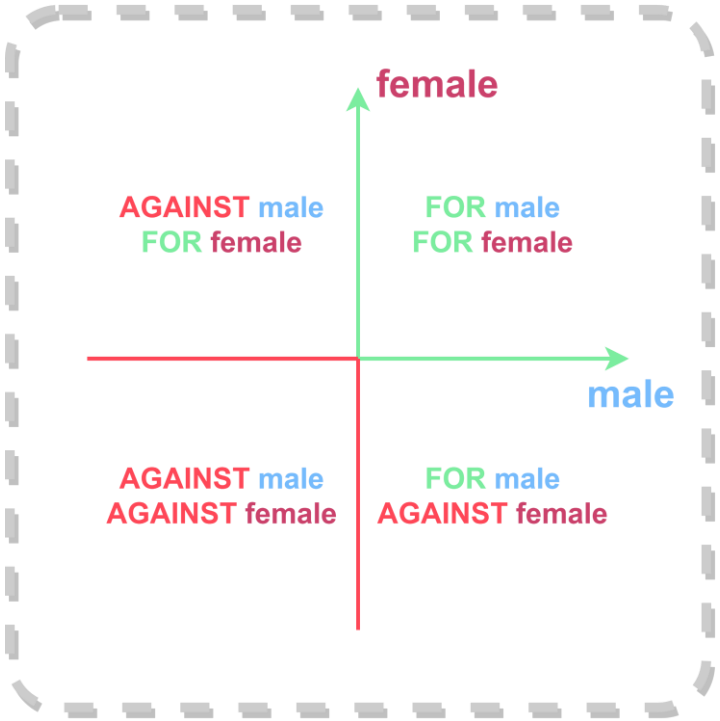


PLM's prediction p_{ij}^{θ}

- $p_{ij}^{\theta} = M^{\theta}(c_i, s_j)$, where p_{ij}^{θ} is a vector with the same dimension as the number of attribute categories, and represents **the probability distribution of the subject s_j belonging to each attribute category in the context c_i**;
- p_{ij}^{θ} reflects the model's stereotype and discrimination of subject s_j under context c_i.

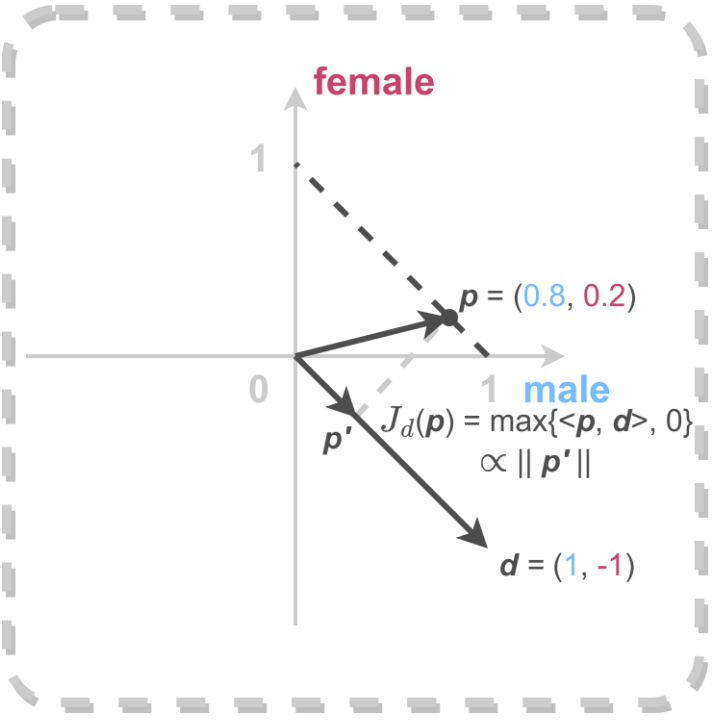
Discriminatory Direction d

- d is determined by the System/Observer, which might be hidden in the data or provided by curious researchers;
- d has the same dimensionality as p_{ij}, and each dimension corresponds to one attribute category.



Discrimination Level Criterion J_d

- J_d describes the degree of overlap between model prediction and social discrimination;
- $J_d(p_{ij}) = \max\{< p_{ij}, d >, 0\}$, where $< p_{ij}, d >$ is the inner product of probability distribution prediction p_{ij} and discriminatory direction d.



From AI's prediction to AI's discrimination

Definition 1

AI model's stereotype r_j is the expected value of J_d over the distribution of Context C under the topic of sub-subject s_j:

$$r_j = \int J_d(p_{*j})dc \quad \cdots\cdots \quad (1)$$

where p_{*j} implies that we fix the s_j in $p_{ij} = M(c_i, s_j)$ and c could be any.

As we cannot enumerate all sub-context c, we approximate the integral in Equation (1) with the summation over all discrete c_i:

$$r_j = \sum_i \{J_d(p_{*j}) \cdot q_i\} \quad \cdots\cdots \quad (2)$$

We calculate the Overall Discrimination Risk R according to the stereotype of the model on different sub-subjects s_j.

Definition 2

Overall Discrimination Risk R We derive R from calculating the expected values of r_j over the distribution of Subject S:

$$R = \int rds \quad \cdots\cdots \quad (3)$$

We approximate these integrals with the summations over all discrete s_j:

$$R = \sum_j \{r_j \cdot w_j\} \quad \cdots\cdots \quad (4)$$

Discrimination Risk Ascription

Definition 3

System Risk r_j^s Under the Topic of Sub-Subject s_j

In statistics, an estimator's bias refers to the separation between the estimator's expected value and the true value of the parameter being estimated. P_{*0} denotes the true value of P_{*j}. When the learning machines make the prediction P_{*j}, the system risk caused by bias is:

$$r_j^s = J_d\left(E\left(P_{*j}\right) - P_{*0}\right) \quad \cdots\cdots \quad (5)$$

where $E(P_{*j})$ is the expected value of P_{*j}. Ideally, $E(P_{*j}) = \int p_{*j} dc$, or discretely, $E(P_{*j}) = \sum_i \{p_{ij} \cdot q_i\}$.

In the experiments, we might assign a uniform distribution with its dimensionality equalling the number of the attribute's categories to P_{*0}. For example, in the binary gender setting, we could set $P_{*0} = (0.5, 0.5)$. When a learning machine predicts $(0{,}5, 0.5)$ precisely, it is unbiased in both the statistical and the discriminatory sense.

Definition 4

Efficiency Risk r_j^e Under the Topic of Sub-Subject s_j

In statistics, an efficient estimator has the minimum possible variance, suggesting a minimal deviation between the estimated value and the true value in the L_2 norm sense. In the context of discrimination estimation, $J(E(P_{*j}))$ is the discrimination risk of an efficient learning machine, as the variance of $E(P_{*j})$ equals 0. We define r_j^e as the overall discrimination risk r_j minus the risk of a most efficient learning machine:

$$r_j^e = R - J(E(P_{*j})) \quad \cdots\cdots \quad (6)$$

Discrimination Risk Ascription

Definition 5

System Risk R^s and Efficiency Risk R^e

Similarly with the definition of R in Definition 2, We derive R^s and R^e from calculating the expected values of r_j^s and r_j^e over the distribution of Subject S:

$$R^s = \int r^s ds \quad \cdots\cdots \quad (7)$$

$$R^e = \int r^e ds \quad \cdots\cdots \quad (8)$$

We approximate these integrals with the summations over all discrete s_j:

$$R^s = \sum_j \{r_j^s \cdot w_j\} \quad \cdots\cdots \quad (9)$$

$$R^e = \sum_j \{r_j^e \cdot w_j\} \quad \cdots\cdots \quad (10)$$

Interestingly, if we combine c_i and s_j, and treat it as a whole, then the computed values reveal the quality of a learning machine as a standard estimator. We can derive P from: $E(P) = \int pd(c,s)$ or $E(P) = \sum_{i,j}\{p_{ij} \cdot q_{ij}\}$, where q_{ij} is the probability of sample p_{ij}. The corresponding discrimination risks are: $\bar{R} = \int J(p)d(c,s), \overline{R^s} = J(E(P) - P_0), \overline{R^e} = \bar{R} - J(E(P))$, where the $\bar{R}$ could also be determined by: $\bar{R} = \sum_{i,j}\{J(p_{ij}) \cdot q_{ij}\}$.

Discrimination Risk Ascription

It is noteworthy that under some conditions, there is a definite equation of relationships between R, R^s, and R^e:

Theorem 1 If $< P_{*0}, d >= 0$, i.e. the true value P_{*0} is perpendicular to discrimination direction d, then $r_j = r_j^s + r_j^e$.

Corollary 1 If $< P_{*0}, d >= 0$ holds for any s_j, then $R = R^s + R^e$.

Empirically, an extremely stereotyped learning machine has a large R^s, while its R_e may equal 0; a randomly initialized learning machine without any learned knowledge has no R^s, but its R^e soars.

DISCRimination Estimation ALgorithm "DISCREAL"

Algorithm 1: DISCREAL-PLM-GEN&OCC

A gender DISCRimination Estimation ALgorithm (DISCREAL) for pre-trained language models (PLMs) on the subject of occupation.

Input: a PLM M^θ the context set *Template* $T = \{(t_i, q_i)\}_i$ the subject set *Occupation* $O = \{(o_j, w_j)\}_j$, the topic set *Gender* $G = \{(g_{uv})_v\}_u$, the discriminatory direction d, the true value P

Output: for each *occupation* o_j: overall gender discrimination risk r_j, system risk r_j^s, efficiency risk r_j^e; on the subject of *occupation*: overall gender discrimination risk R, system risk R^s, efficiency risk R^e

Pair the components in T and O to form the *Input* X;
Feed X into M^θ and get the *Output* Y;
Initialize $R = 0, R^s = 0, R^e = 0$;
Initialize an empty dict D, which stores r_j, r_j^s, r_j^e for each o_j;
for o_j *in* $\{o_1, ..., o_{max}\}$ **do**
    Initialize $E(P_{*j}) = 0, r_j = 0$;
    for t_i *in* $\{t_1, ..., t_{max}\}$ **do**
        $p_{ij} = [\frac{\sum_{g \in G_u} \hat{p}_{ij}(g)}{\sum_{g \in G} \hat{p}_{ij}(g)}, u \in \{1, ..., ||G||_0\}]$, where $\hat{p}_{ij}(\cdot) = M^\theta(t_i, o_j)[index(\cdot)]$;
        $E(P_{*j})$ += $p_{ij} \cdot q_i$;
        r_j += $J_d(p_{ij}) \cdot q_i$;
    end
    $r_j^s = J_d(E(P_{*j}) - P_{*0})$;
    $r_j^e = r_j - J_d(E(P_{*j}))$;
    Store $[r_j, r_j^s, r_j^e]$ in $D[o_j]$;
    R += $r_j \cdot w_j$;
    R^s += $r_j^s \cdot w_j$;
    R^e += $r_j^e \cdot w_j$;
end
return D, R, R^s, R_e;

Pre-trained Language Model M^θ

- T5, BART, RoBERTa, ALBERT, GPT-2, BERT, XLNet, distilBERT

Occupation O and Gender G

- O includes words like "doctor," "nurse," etc., which are frequently expressed as biased terms rather than neutral ones. In our experiments, we assign both a uniform distribution and a distribution presented by labor statistics from an official US website to the occupation words' probability distribution.
- G (in the binary gender setting) typically contains two sub-tuples, namely $G_{masculine}$ and $G_{feminine}$, with the former consisting of words like "**he**," "**man**," "**grandfather**," etc., and the latter of corresponding words like "**she**," "**woman**," "**grandmother**," etc.

Template T

- T consists of **(template, template's probability) tuples**, with both the templates and their probability scraped and calculated from randomly selected 10,000 entity passages from Wikipedia;
- These templates are generally **devoid of information indicating any of the attributes or the attributes' sub-categories**. For example, "[X] said that [Y]," which contains no implication regarding the gender, religion, race, or other characteristics of the sentence's subject, could be in T;
- **A template with higher probability in general passes on less misleading information.**

DISCRimination Estimation ALgorithm "DISCREAL"

Algorithm 1: DISCREAL-PLM-GEN&OCC

A gender DISCRimination Estimation ALgorithm (DISCREAL) for pre-trained language models (PLMs) on the subject of occupation.

Input: a PLM M^{θ}, the context set *Template* $T = \{(t_i, q_i)\}_i$, the subject set *Occupation* $O = \{(o_j, w_j)\}_j$, the topic set *Gender* $G = \{(g_{uv})_v\}_u$, the discriminatory direction d, the true value P

Output: for each *occupation* o_j: overall gender discrimination risk r_j, system risk r_j^s, efficiency risk r_j^e; on the subject of *occupation*: overall gender discrimination risk R, system risk R^s, efficiency risk R^e

Pair the components in T and O to form the *Input* X;
Feed X into M^{θ} and get the *Output* Y;
Initialize $R = 0, R^s = 0, R^e = 0$;
Initialize an empty dict D, which stores r_j, r_j^s, r_j^e for each o_j;
for o_j *in* $\{o_1, ..., o_{max}\}$ **do**
    Initialize $E(P_{*j}) = 0, r_j = 0$;
    for t_i *in* $\{t_1, ..., t_{max}\}$ **do**
        $p_{ij} = [\frac{\sum_{g \in G_u} \hat{p}_{ij}(g)}{\sum_{g \in G} \hat{p}_{ij}(g)}, u \in \{1, ..., ||G||_0\}]$, where $\hat{p}_{ij}(\cdot) = M^{\theta}(t_i, o_j)[index(\cdot)]$;
        $E(P_{*j})$ += $p_{ij} \cdot q_i$;
        r_j += $J_d(p_{ij}) \cdot q_i$;
    end
    $r_j^s = J_d(E(P_{*j}) - P_{*0})$;
    $r_j^e = r_j - J_d(E(P_{*j}))$;
    Store $[r_j, r_j^s, r_j^e]$ in $D[o_j]$;
    R += $r_j \cdot w_j$;
    R^s += $r_j^s \cdot w_j$;
    R^e += $r_j^e \cdot w_j$;
end
return D, R, R^s, R_e;

$$R = \sum_j \{r_j \cdot w_j\} \cdots\cdots (4)$$

$$R^s = \sum_j \{r_j^s \cdot w_j\} \cdots\cdots (9)$$

$$R^e = \sum_j \{r_j^e \cdot w_j\} \cdots\cdots (10)$$

$$p_{ij}^{\theta} = M^{\theta}(c_i, s_j)$$

$$E(P_{*j}) = \sum_i \{p_{ij} \cdot q_i\}$$

$$J_d(p_{ij}) = \max\{< p_{ij}, d >, 0\}$$

$$r_j = \sum_i \{J_d(p_{*j}) \cdot q_i\} \cdots\cdots (2)$$

$$r_j^s = J_d(E(P_{*j}) - P_{*0}) \cdots\cdots (5)$$

$$r_j^e = R - J(E(P_{*j})) \cdots\cdots (6)$$

Results

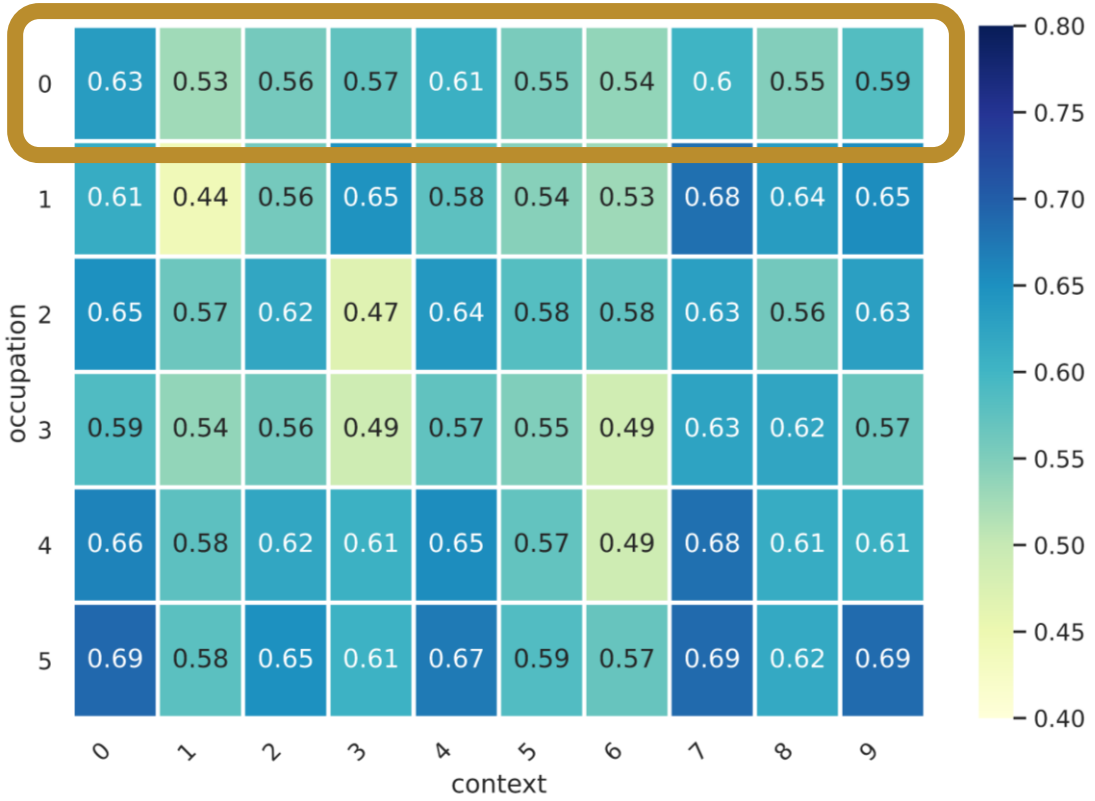
If we assign **a uniform distribution** to both Template T and Occupation O, then R, R^s, and R^e measure the "**discrimination level**" of PLMs as standard estimators, emphasizing their statistical properties more. When assigning **distributions aligning with the real world** to T and O, we stick to using the term "**discrimination risk**" to describe the potential social risk faced by the PLMs when they are applied to real-world tasks.

	Gender						Race					
	Discrimination Level (e-3)			Discrimination Risk (e-3)			Discrimination Level (e-3)			Discrimination Risk (e-3)		
	R	R^s	R^e	R	R^s	R^e	R	R^s	R^e	R	R^s	R^e
ideal	0	0	0	0	0	0	0	0	0	0	0	0
stereotyped	1000	1000	0	1000	1000	0	1000	1000	0	1000	1000	0
random	1000	0	1000	1000	0	1000	1000	0	1000	1000	0	1000
T5	**870.32**	**869.14**	*1.18*	**837.02**	**836.00**	*1.02*	662.15	640.67	21.48	650.97	628.65	22.31
BART	484.57	467.66	16.91	420.95	406.04	14.91	*417.81*	*297.08*	**120.73**	430.07	*322.22*	**107.85**
RoBERTa	453.50	417.08	36.42	369.19	334.94	34.25	789.18	754.85	34.33	793.85	766.09	27.76
ALBERT	328.73	*253.10*	**75.63**	358.56	*272.76*	**85.81**	431.43	419.69	11.74	*426.12*	414.31	11.80
GPT-2	415.67	395.63	20.04	361.31	346.86	14.45	492.67	459.80	32.87	510.42	476.80	33.62
BERT	*310.07*	304.74	5.33	*304.91*	301.83	3.09	617.40	611.75	5.65	635.73	631.42	*4.31*
XLNet	734.25	717.68	16.57	790.57	779.32	11.25	**933.29**	**912.02**	21.27	**930.31**	**913.25**	17.07
distilBERT	507.86	491.39	16.47	471.78	462.22	9.56	561.35	557.19	*4.17*	602.33	600.55	1.78

Table 1: Discrimination level and discrimination risk calculated with DISCREAL, with the worst performance in **bold**, the best performance in *italic*, and both of them highlighted with underline for each column. Row 1-3 list the evaluation results from the ideal model (which always makes an unbiased prediction given any probe), the stereotyped model (which sticks to a biased prediction), and the randomly initialized model respectively. We evaluate 8 popular pre-trained language models including T5, BART, RoBERTa, ALBERT, GPT-2, BERT, XLNet, and distilBERT (from row 4 to row 11). As for the *Discriminatory Topic*, we test *Gender* (from column 1 to column 6) and *Race* (from colum 7 to column 12).

We could compare the discrimination level/risk of various PLMs under different topics with DISCREAL.

Results



BERT's gender prediction for the **(0) nurse, (1) stylist, (2) receptionist, (3) doctor, (4) programmer, and (5) captain** in different texts. The 10 templates are: **(0) "The [X] explained that [Y]," (1) "The [X] confirmed that [Y]," (2) "The [X] told that [Y]," (3) "The [X] found that [Y]," (4) "The [X] mentioned that [Y]," (5) "The [X] was old [Y]," (6) "The [X] was one [Y]," (7) "The [X] asked if [Y]," (8) "The [X] was aware that [Y]," (9) "The [X] alleged that [Y]."** The values in the figure represent the probability that the model believes that the occupation in the corresponding sentence is **male**.

AI would exhibit biased or inconsistent outputs, depending largely on the contexts.

AI models and humans may have different stereotypes about the same profession.

- Social discrimination may lead to the belief that nurses are predominantly female, while **AI models such as T5, BERT, RoBERTA, and XLNet may classify nurses as more likely to be male.**

Results

	Gender	
	Discrimination Level	**Discrimination Risk**
T5	electrician dentist architect economist diver	electrician dentist economist architect surgeon
BART	electrician builder architect machinist welder	electrician builder architect promoter machinist
RoBERTa	captain builder electrician promoter chemist	captain builder electrician promoter chemist
ALBERT	stylist receptionist waitress reporter designer	broker journalist economist stylist secretary
GPT-2	coach captain economist developer programmer	coach economist captain developer programmer
BERT	engineer promoter architect surveyor underwriter	promoter engineer architect underwriter manager
XLNet	promoter developer doctor inspector representative	doctor broker chemist inspector promoter
distilBERT	engineer architect surveyor builder officer	coach engineer architect manager surveyor

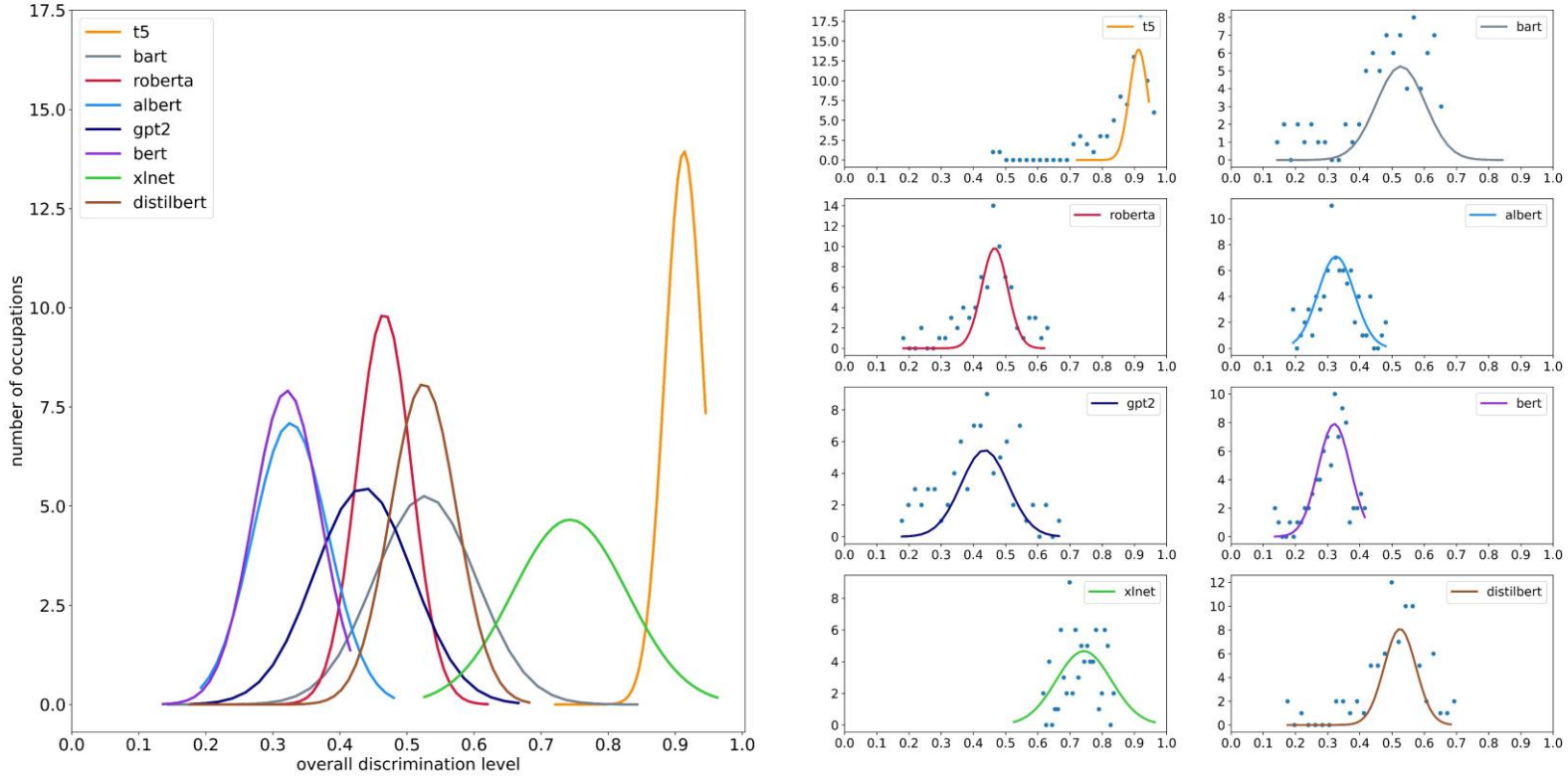
Table 2: Top 5 models with the highest discrimination level/risk under the topic of *Gender*.

	T5	**BART**	**RoBERTa**	**ALBERT**	**GPT-2**	**BERT**	**XLNet**	**distilBERT**
T5	1	0.6127	0.5533	-0.2331	0.5255	0.4066	0.4071	0.4166
BART	0.6127	1	0.8424	-0.6144	0.7318	0.6406	0.4421	0.7367
RoBERTa	0.5533	0.8424	1	-0.5994	0.7200	0.6002	0.4181	0.6637
ALBERT	-0.2331	-0.6144	-0.5994	1	-0.3960	-0.4823	-0.3983	-0.5258
GPT-2	0.5255	0.7318	0.7200	-0.3960	1	0.6244	0.4466	0.6398
BERT	0.4066	0.6406	0.6002	-0.4823	0.6244	1	0.5814	0.8177
XLNet	0.4071	0.4421	0.4181	-0.3983	0.4466	0.5814	1	0.4991
distilBERT	0.4166	0.7367	0.6637	-0.5258	0.6398	0.8177	0.4991	1

Table 3: Spearman's rank correlation coefficient (SRCC) between the ranking of pre-trained language models' occupation discrimination level. In general, an SRCC with an absolute value closer to 1 indicates a tighter correlation; conversely, an SRCC approaching 0 reflects the irrelevance of the two rankings.

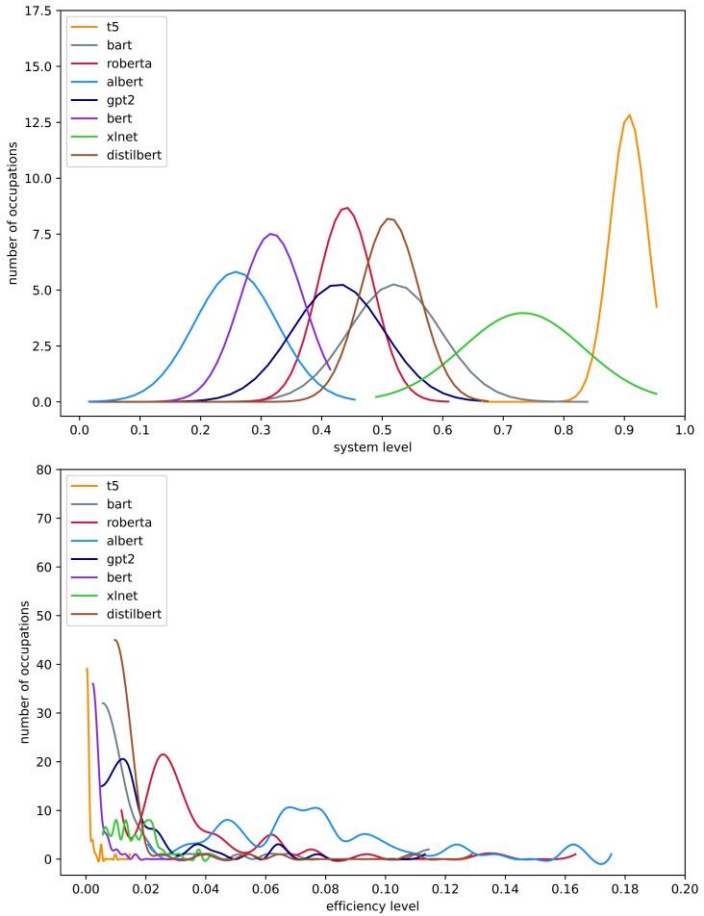
AI models differ significantly in their inherent discrimination level, and there is no distinct discrimination pattern as could be observed in human society.

Results



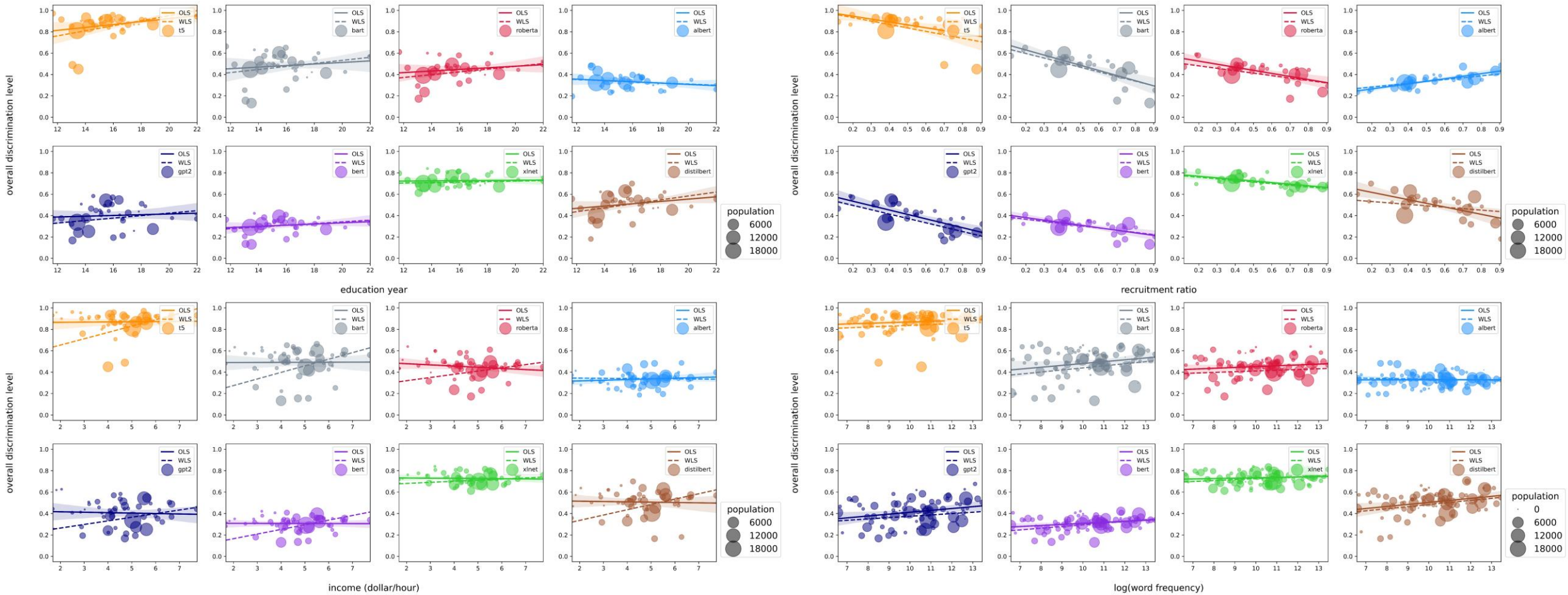
The overall discrimination level under the topic of Gender. We first fit the overall discrimination distribution with normal distribution (right), and then map the distribution shape of each pre-trained language models to the same coordinate system (left).

- **The expected value of R^s is generally larger than that of R^e, suggesting that the dominant reason for the high discrimination level of PLMs is their built-in prejudice.**
- **The distribution of R^s is typically bell-shaped, while the distribution of R^e often follows a long-tail pattern.**



The detailed discrimination decomposition under the topic of Gender. We fit the system level distribution with normal distribution. To better demonstrate the amorphous distribution of efficiency level, we perform interpolation on the calculated values and plot the interpolated lines.

The Correlation Between Social and Economic Factors and the PLMs' Discrimination Level

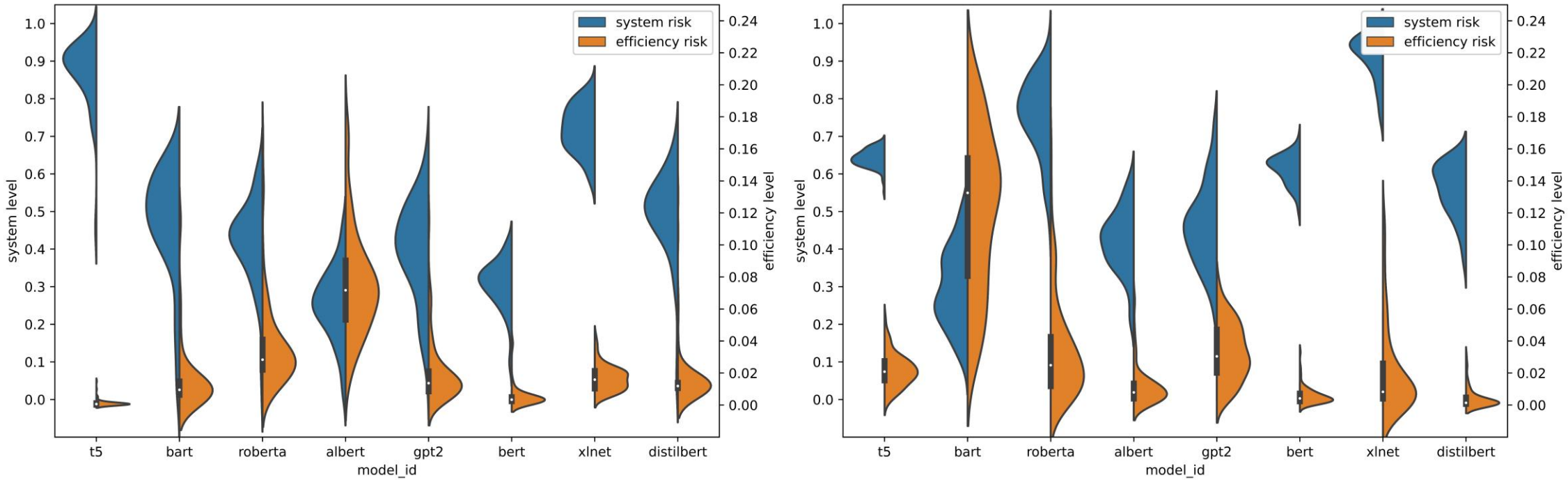


The regressions between social and economic factors (i.e., education year [1], recruitment ratio [2], salary [3], and word frequency [4]) and Gender discrimination level.
In every sub-plot, each point denotes an occupation, with its size indicating the population of that occupation. We present the result determined by the ordinary least squares (OLS) principle with solid lines, and by the weighted least squares (WLS) principle, where the weights are derived from the labor statistics by occupation, with dashed lines.

PLMs are more likely to make prejudiced predictions for occupations with a higher proportion of women.

[1] https://www.bls.gov/emp/tables/educational-attainment.htm [2] https://www.bls.gov/cps/cpsaat11.htm [3] https://www.bls.gov/mwe/tables.htm
[4] https://github.com/IlyaSemenov/wikipedia-word-frequency/blob/master/results/enwiki-2022-08-29.txt

Managing discrimination in AI



(a) Gender (b) Race

The violin plot for comparing the range and shape of system level distribution and efficiency level distribution under the topic of Gender and Race. Both system distribution and efficiency distribution are comparable across the topic and the discrimination criterion along the y-stick.

To effectively address AI discrimination, we should take a two-pronged approach that focuses on both bias and efficiency.

Future Work

- **Template Generalizability**
- **How Learning Machines Itself Might Effect Its Inborn Discrimination Level**
- **Learning Machines Other Than PLMs**

Thank you
for listening.