How Gender Debiasing Affects Internal Model Representations, and Why It Matters

Hadas Orgad, Seraphina Goldfarb-Tarrant, Yonatan Belinkov

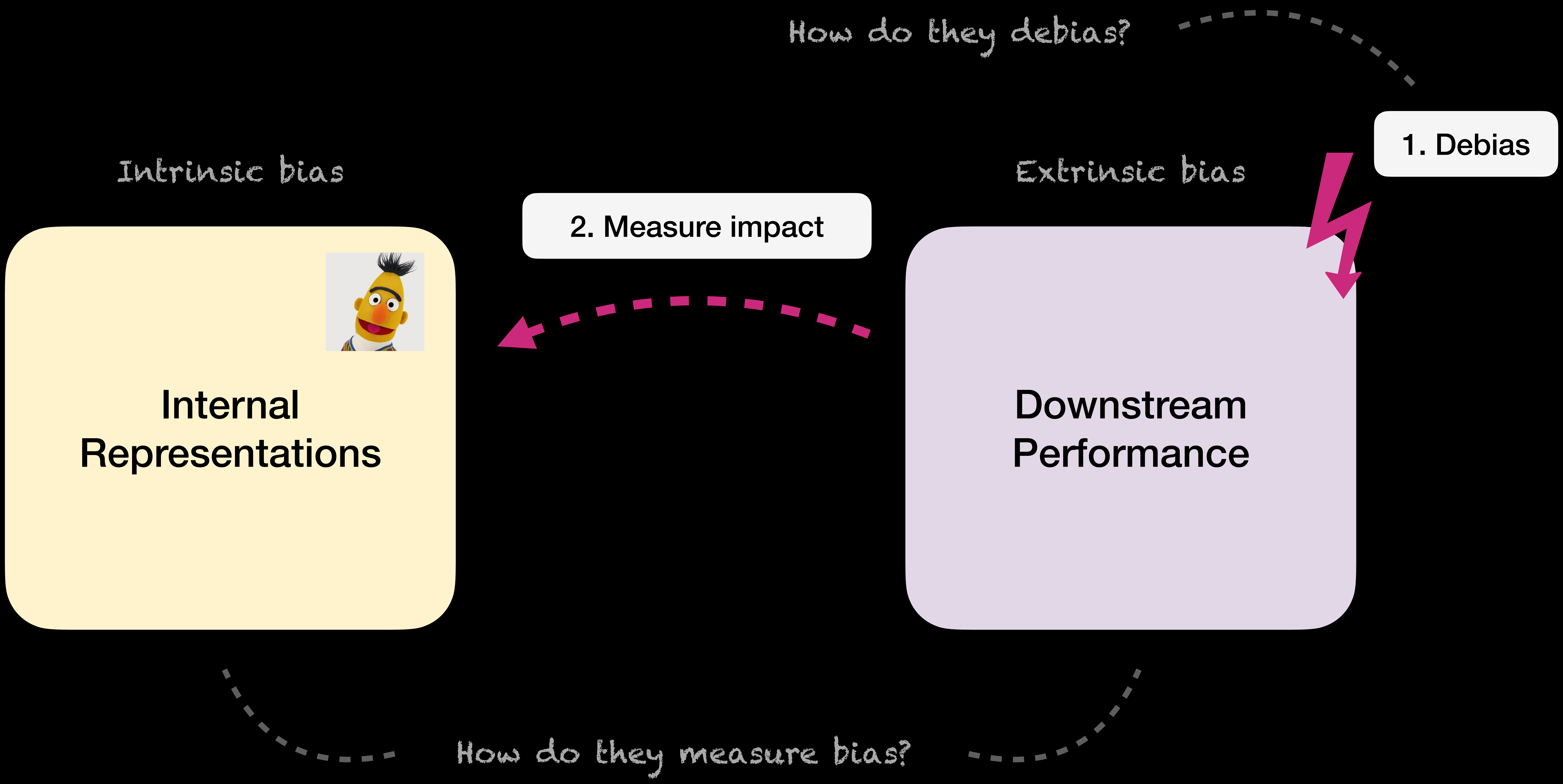
Presented by
Giuseppe Attanasio

Date
October 13, 2022

Extrinsic vs. Intrinsic Gender Bias

- Intrinsic bias
  - Internal representations (WEAT and co.)
- Extrinsic bias
  - Downstream performance (Group parity and co.)

"Our goal is [...] understanding the
relationship between a model's internal representations and its extrinsic gender bias
by examining the effects of various debiasing methods on the model's representations"


How do they debias?
1. Debias
Intrinsic bias
Extrinsic bias
2. Measure impact
Internal Representations
Downstream Performance
How do they measure bias?

Intrinsic bias

Internal Representations

- Contextualised Embedding Association Test (CEAT)

  Gun and Caliskan, 2020

- Compression
  - Predicting gender from model's representations
  - Minimum Description Length (MDL) probe

    Voita and Titov, 2020

Extrinsic bias

1. Debias

Downstream Performance

Intrinsic bias

Extrinsic bias

1. Debias

Internal Representations

Downstream Performance

Occupation Classification:
(TPR(teacher|men) -
TPR(teacher|women)).abs()

- Contextualised Embedding Association Test (CEAT)

  Gun and Caliskan, 2020

- Compression ↓
  - Predicting gender from model's representations
  - Minimum Description Length (MDL) probe

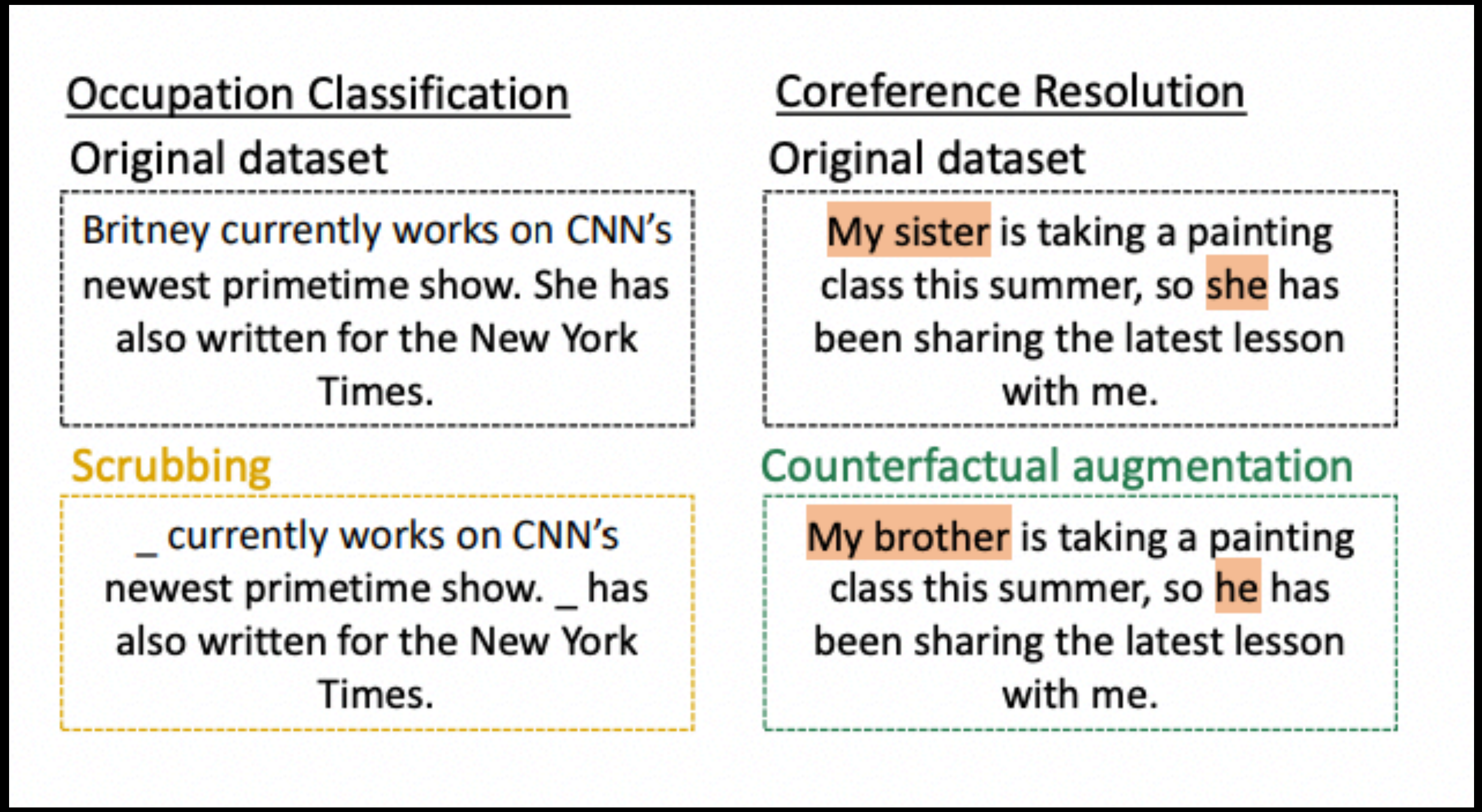
    Voita and Titov, 2020

- TPR and FPR gaps ↓
  - 1) sum(gaps)
  - 2) Pearson(class gap, women employment) (from labour statistics)
- Independence ↓ $KL(P(\mathrm{r}\,|\,\mathrm{z}=z), P(\mathrm{r}))\forall z \in \{M, F\}$
- Separation ↓ $KL(P(\mathrm{r}\,|\,\mathrm{y}=y, \mathrm{z}=z), P(\mathrm{r}\,|\,\mathrm{y}=y))\forall z \forall y$
- Sufficiency ↓ $Wass(P(\mathrm{y}\,|\,\mathrm{r}=r, \mathrm{z}=z), P(\mathrm{y}\,|\,\mathrm{r}=r))$

Extrinsic bias

1. Debias

Examples



- Scrubbing
  - Remove “he”, “she”, “husband”, etc.
- Balancing (over- or sub-sampling genders)
  - Stratified on class labels
- Anonymization (remove named entities)
- Counterfactual Augmentation

Setup

- Occupation Classification
  - Bias in Bios
  - Probe: [CLS], gender from bio
- Coreference Resolution
  - FT: Ontonotes 5.0, T: Winobias
  - Probe: profession word, stereotypical gender
- RoBERTa, DeBERTa

The doctor called the nurse because he/she needed help

Compression!

- Compression captures variations on debiasing
- CEAT in CR shows no bias for unbiased models
- Superficial debiasing: effects on extrinsic don't match intrinsic

Debiasing Strategy	Intrinsic		Extrinsic							
			Before				After			
	Compression	CEAT	TPR (P)	FPR (S)	Sep	Suff	TPR (P)	FPR (S)	Sep	Suff
Random	5.61*	0.12†	-	-	-	-	-	-	-	-
Pre-trained	10.12	0.49*	-	-	-	-	-	-	-	-
None	4.12	0.22	0.76	0.08	0.33	9.45	0.78	0.073	0.33	9.70
Oversampling	8.52*	0.29	0.73	0.09*	0.31	8.32*	0.81*	0.068*	0.33	10.91*
Subsampling	3.57	0.22	**0.32***	**0.03***	**0.20***	**1.22***	**0.70***	0.08*	0.30*	1.32*
Scrubbing	**1.70***	0.23	0.70*	0.06*	0.30	4.93*	0.71*	**0.06***	**2.56***	**0.81***

(a) Occupation classification: Comparison of intrinsic and extrinsic metrics before and after retraining of classification layer, over 10 seeds per fine-tuned model and per retrained classification model.

Debiasing Strategy	Intrinsic		Extrinsic							
			Before				After			
	Compression	CEAT	F1 diff	FPR (S)	Sep	Suff	F1 diff	FPR (S)	Sep	Suff
Random	0.83*	0.12†	-	-	-	-	-	-	-	-
Pre-trained	0.96	0.49*	-	-	-	-	-	-	-	-
None	1.98	0.35	6.63	0.12	1.25	8.69	6.07	0.11	1.19	7.35
Anon	2.07*	0.31*	7.26	0.13	1.34	8.82	7.42*	0.13*	1.34*	8.66*
CA	**1.50***	0.27*	**2.30***	0.05*	**0.54***	1.67*	3.67*	0.06*	0.67*	2.40*
Anon + CA	1.54*	**0.25***	2.42*	**0.049***	0.56*	**1.56***	**2.86***	**0.05***	**0.59***	**1.65***

(b) Coreference resolution: Comparison of intrinsic and extrinsic metrics before and after retraining of classification layer, over 10 seeds per fine-tuned model and 5 seeds per retrained classification model.

Table 1: Results on both tasks. * marks significant reduction or increase in bias ($p < 0.05$ on Pitman's permutation test), compared to the non-debiased model (debiasing strategy None). The lowest bias score in each column is marked with **bold**. P = Pearson; S = Sum. † was computed only on 3 out of 10 tests for which CEAT's $p < 0.05$.

Compression!

"After": fine-tune, freeze RoBERTa, fine-tune CLS head

- Compression captures variations on debiasing
- CEAT in CR shows no bias for unbiased models
- Superficial debiasing: effects on extrinsic don't match intrinsic
- Strength of bias restoration is predicted by compression

Debiasing Strategy	Intrinsic		Extrinsic							
			Before				After			
	Compression	CEAT	TPR (P)	FPR (S)	Sep	Suff	TPR (P)	FPR (S)	Sep	Suff
Random	5.61*	0.12†	-	-	-	-	-	-	-	-
Pre-trained	10.12	0.49*	-	-	-	-	-	-	-	-
None	4.12	0.22	0.76	0.08	0.33	9.45	0.78	0.073	0.33	9.70
Oversampling	8.52*	0.29	0.73	0.09*	0.31	8.32*	0.81*	0.068*	0.33	10.91*
Subsampling	3.57	0.22	**0.32***	**0.03***	**0.20***	**1.22***	**0.70***	0.08*	0.30*	1.32*
Scrubbing	**1.70***	0.23	0.70*	0.06*	0.30	4.93*	0.71*	**0.06***	**2.56***	**0.81***

(a) Occupation classification: Comparison of intrinsic and extrinsic metrics before and after retraining of classification layer, over 10 seeds per fine-tuned model and per retrained classification model.

Debiasing Strategy	Intrinsic		Extrinsic							
			Before				After			
	Compression	CEAT	F1 diff	FPR (S)	Sep	Suff	F1 diff	FPR (S)	Sep	Suff
Random	0.83*	0.12†	-	-	-	-	-	-	-	-
Pre-trained	0.96	0.49*	-	-	-	-	-	-	-	-
None	1.98	0.35	6.63	0.12	1.25	8.69	6.07	0.11	1.19	7.35
Anon	2.07*	0.31*	7.26	0.13	1.34	8.82	7.42*	0.13*	1.34*	8.66*
CA	**1.50***	0.27*	**2.30***	0.05*	**0.54***	1.67*	3.67*	0.06*	0.67*	2.40*
Anon + CA	1.54*	**0.25***	2.42*	**0.049***	0.56*	**1.56***	**2.86***	**0.05***	**0.59***	**1.65***

(b) Coreference resolution: Comparison of intrinsic and extrinsic metrics before and after retraining of classification layer, over 10 seeds per fine-tuned model and 5 seeds per retrained classification model.

Table 1: Results on both tasks. * marks significant reduction or increase in bias ($p < 0.05$ on Pitman's permutation test), compared to the non-debiased model (debiasing strategy None). The lowest bias score in each column is marked with **bold**. P = Pearson; S = Sum. † was computed only on 3 out of 10 tests for which CEAT's $p < 0.05$.

Correlation between Intrinsic and Extrinsic

- OC: correlations appears with Compression after retraining
- CR: correlation is high “before” and decreases “after”
- CEAT has low correlation

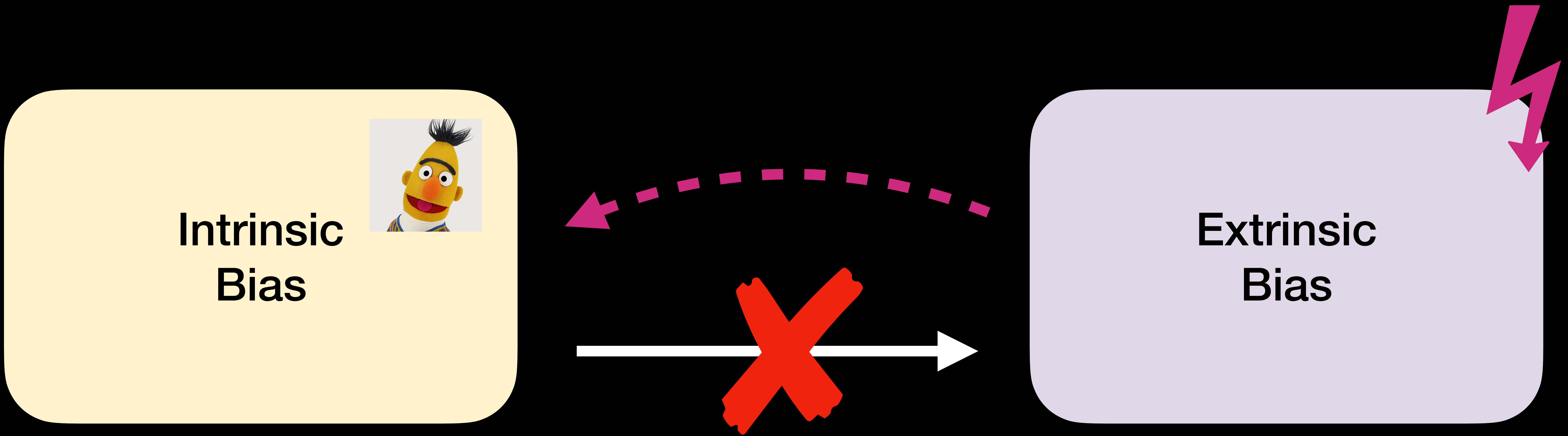
	Occupation Classification				Coreference Resolution			
	R^2 Compression		R^2 CEAT		R^2 Compression		R^2 CEAT	
Metric	Before	After	Before	After	Before	After	Before	After
F1 diff (*pro* − *anti*)	-	-	-	-	0.821	0.709	0.246	0.005
TPR gap (P)	0.046	0.304	0.042	0.049	0.222	0.006	0.008	0.012
TPR gap (S)	0.049	0.449	0.022	0.036	0.817	0.752	0.297	0.003
FPR gap (P)	0.001	0.120	0.008	0.002	0.021	0.054	0.002	0.000
FPR gap (S)	0.353	0.046	0.079	0.001	0.844	0.773	0.263	0.004
Precision gap (P)	0.032	0.173	0.000	0.000	0.068	0.038	0.019	0.000
Precision gap (S)	0.174	0.529	0.000	0.021	0.849	0.774	0.268	0.006
Independence gap (S)	0.251	0.382	0.050	0.005	0.778	0.732	0.355	0.001
Separation gap (S)	0.066	0.165	0.046	0.009	0.835	0.776	0.261	0.005
Sufficiency gap (S)	0.202	0.567	0.040	0.034	0.825	0.753	0.287	0.002

Table 2: Coefficient determination of the regression line taken on the compression rate or CEAT and each extrinsic metric, before and after retraining of the classification layer. P = Pearson; S = Sum.

Authors' take

- Compression (gender extractability) is a better indicator than CEAT for gender bias in NLP models
- High gender extractability and low extrinsic bias metrics means superficial debiasing
  - Bias is still "there", retraining restores it
- OP and CR have tell different stories

My take



But authors assume it all along...

while not always in CEAT. Thus, gender extractability is a more reliable indicator of gender bias in NLP models.

ing. Hence, the debiasing was primarily cosmetic, and the representations within the LM were not debiased. The model fine-tuned on oversampled

sic metrics. However, compression as an intrinsic bias metric can indicate meaningful debiasing of internal model representations even when not all

My take

Intrinsic Bias

Extrinsic Bias

We should prove this!

But authors assume it all along...

while not always in CEAT. Thus, gender extractability is a more reliable indicator of gender bias in NLP models.

ing. Hence, the debiasing was primarily cosmetic, and the representations within the LM were not debiased. The model fine-tuned on oversampled

sic metrics. However, compression as an intrinsic bias metric can indicate meaningful debiasing of internal model representations even when not all

An experimental laboratory, dark pink

Generated by the author with Stable Diffusion

A crowd of researchers attending a conference in the middle of the desert

Generated by the author with Stable Diffusion