# Assessing Social and Intersectional Biases in Contextualized Word Representations

Yi Chern Tan, L. Elisa Celis
Yale University
{yichern.tan, elisa.celis}@yale.edu

Yale

# Social Bias in Contextual Word Models

Key Objectives:

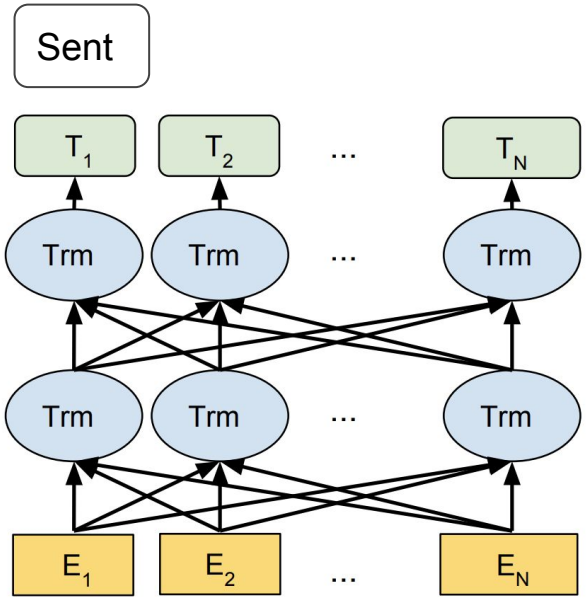
- Do embedding association tests demonstrate social bias on contextual word encodings in the test sentence?
- Can we develop more comprehensive tests for gender, race and intersectional identities?

# Extension to Contextual Word Level

Sentence encoding level

Contextual word level

Context free word level



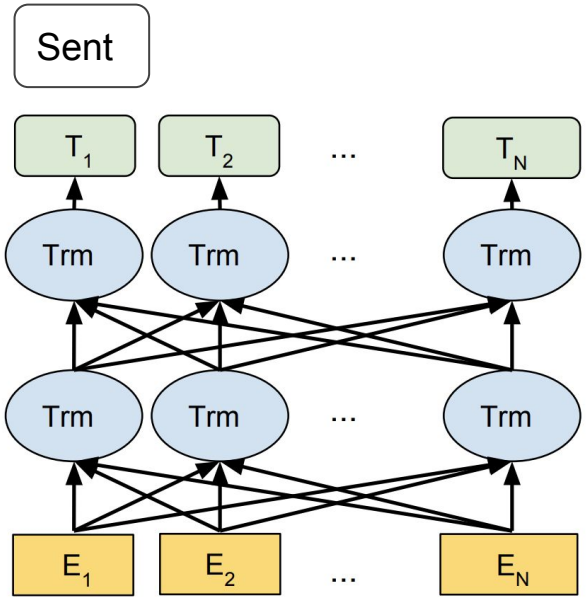
# Extension to Contextual Word Level

The nurse is here.

Sentence encoding level

Contextual word level

Context free word level



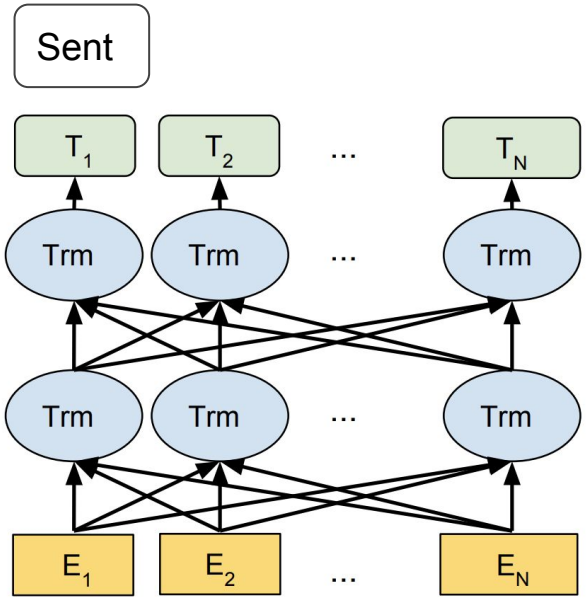
# Extension to Contextual Word Level

The nurse is here.

Sentence encoding level

Contextual word level

Context free word level



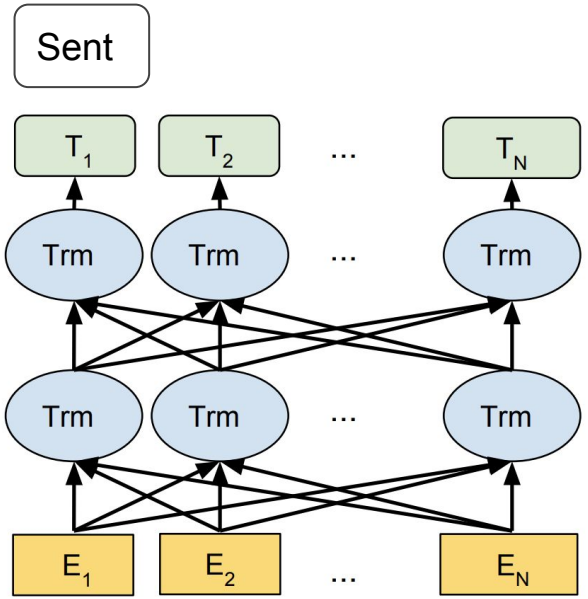
# Extension to Contextual Word Level

The nurse is here.

Sentence encoding level

Contextual word level

Context free word level



# Embedding Association Tests

How related is concept X with attribute A, and concept Y with attribute B? As opposed to X with B, and Y with A?

| Concept | Attribute |
| --- | --- |
| X: Male names<br><br>*E.g., “This is Paul.”* | A: Stereotypically Female Occupations<br><br>*E.g., “The nurse is here”* |
| Y: Female names<br><br>*E.g., “This is Emily”* | B: Stereotypically Male Occupations<br><br>*E.g., “The doctor is there”* |

# Methods

| Concept | Attribute |
| --- | --- |
| Gender | • Stereotypical Occupations<br>• Pleasant/Unpleasant<br>• Career/Family<br>• Science/Arts<br>• Likable/Unlikable<br>• Competent/Incompetent |
| Race | • Pleasant/Unpleasant<br>• Career/Family<br>• Science/Arts<br>• Likable/Unlikable<br>• Competent/Incompetent |

Models:

- CBoW (GloVe)
- ELMo
- BERT (bbc, blc)
- GPT
- GPT-2 (117M, 345M)

# Analysis

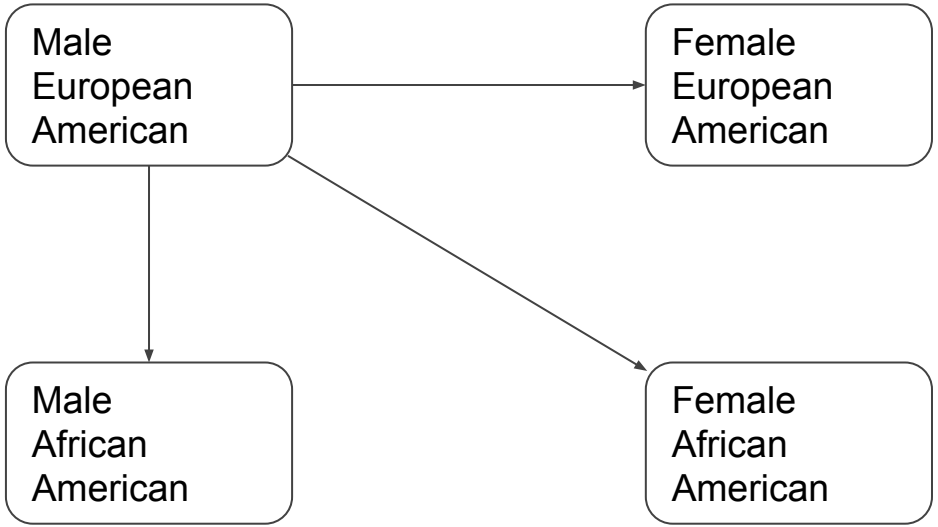
| Test | CBoW | ELMo | BERT (bbc) | BERT (blc) | GPT | GPT-2 (117M) | GPT-2 (345M) |
|---|---|---|---|---|---|---|---|
| gender | 0.73 | 0.03 | 0.32 | 0.12 | 0.35 | 0.24 | 0.15 |
| race | 0.60 | 0.10 | 0.58 | 0.58 | 0.39 | 0.42 | 0.42 |
| intersectional | 0.29 | 0.10 | 0.71 | 0.38 | 0.33 | 0.29 | 0.10 |
| disability, age | 0.75 | 0.17 | 0.00 | 0.00 | 0.17 | 0.33 | 0.17 |
| **Overall** | 0.58 | 0.08 | 0.48 | 0.33 | 0.35 | 0.32 | 0.23 |

- All instances of significant effects had positive effect sizes.

- 93 instances where a test has a significant effect on either contextual word level (c-word) or sentence (sent) encoding
  - 36.6% (34) observed only with c-word encoding
  - 25.8% (24) observed only with sent encoding
  - 37.6% (35) observed on both encoding types

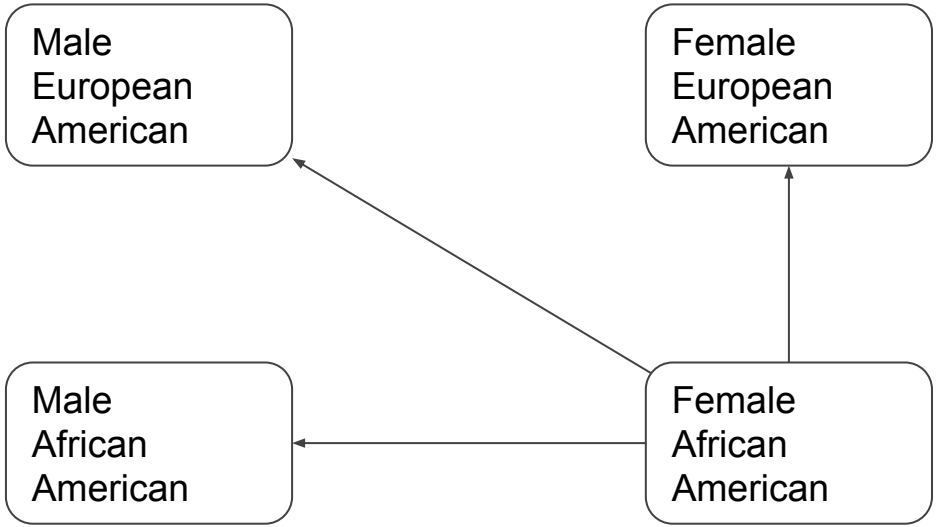
# Intersectional Identities

“The experiences of women of color are frequently the product of intersecting patterns of racism and sexism.” - Kimberlé Crenshaw



# Intersectional Identities

“The experiences of women of color are frequently the product of intersecting patterns of racism and sexism.” - Kimberlé Crenshaw



# Analysis: Intersectionality

| Test | Encoding | CBoW | ELMo | BERT (bbc) | BERT (blc) | GPT | GPT-2 (117M) | GPT-2 (345M) |
|---|---|---|---|---|---|---|---|---|
| +I1: (F) EA/AA Names, P/U | word | +1.19 | -0.01 | +1.13 | +1.43 | -1.16 | +1.07 | +0.65 |
| +I1: (F) EA/AA Names, P/U | sent | +0.15 | +0.04 | +1.35 | 0.00 | +0.44 | -0.75 | -0.75 |
| +I1: (F) EA/AA Names, P/U | c-word | NA | +0.04 | +0.98 ⚠ | -0.12 | +1.45 ⚠ | +0.09 | +0.41 ✓ |
| +I2: (AA) M/F Names, P/U | word | -0.63 | +0.64 | +0.96 | +1.07 | -0.78 | +0.70 | -0.49 |
| +I2: (AA) M/F Names, P/U | sent | -0.94 | +0.02 | +0.89 ✗ | 0.00 | -0.80 | -0.66 | -0.88 |
| +I2: (AA) M/F Names, P/U | c-word | NA | +0.07 | -0.43 | -0.10 | +0.20 | +0.31 ✓ | -0.23 |
| +I3: (M) EA/AA Names, P/U | word | +1.06 | -0.31 | +0.37 | +0.37 | -0.93 | +1.43 | +0.98 |
| +I3: (M) EA/AA Names, P/U | sent | +0.28 | -0.44 | +0.94 | +1.05 | +0.79 | +0.17 | +0.21 |
| +I3: (M) EA/AA Names, P/U | c-word | NA | -0.02 | +0.85 ⚠ | +0.43 ⚠ | +1.11 ⚠ | -0.56 | -0.49 |
| +I4: (EA) M/F Names, P/U | word | -0.22 | +0.36 | -0.42 | -0.39 | -0.48 | +1.06 | +0.21 |
| +I4: (EA) M/F Names, P/U | sent | -0.23 | -0.58 | +0.14 | -0.05 | -0.45 | +0.28 | -0.07 |
| +I4: (EA) M/F Names, P/U | c-word | NA | +0.02 | -0.59 | +0.50 ✓ | -0.27 | -0.31 | -0.03 |
| +I5: EA M/AA F Names, P/U | word | +0.48 | +0.48 | +1.19 | +1.26 | -1.15 | +1.64 | +0.77 |
| +I5: EA M/AA F Names, P/U | sent | -0.10 | -0.42 | +1.48 | +1.68 ✗ | -0.06 | -0.56 | -0.78 |
| +I5: EA M/AA F Names, P/U | c-word | NA | +0.07 | +0.42 ⚠ | +0.26 | +1.26 ✓ | -0.43 | +0.16 |
| ABW Stereotype Names | word | +1.10 | +0.53 | +1.23 | +1.69 | -0.79 | +0.87 | +0.21 |
| ABW Stereotype Names | sent | +0.62 | +0.52 ✗ | +1.62 | 0.00 | -0.82 | -0.70 | -0.92 |
| ABW Stereotype Names | c-word | NA | +0.19 | +1.34 ⚠ | +0.08 | +1.04 ✓ | +0.15 | -0.28 |

By anchoring the comparison on the most privileged group, models exhibit more bias for identities at an intersection of gender and race than constituent minority identities.

# Analysis: Gender

| Test | Encoding | CBoW | ELMo | BERT (bbc) | BERT (blc) | GPT | GPT-2 (117M) | GPT-2 (345M) |
|---|---|---|---|---|---|---|---|---|
| +C11: M/F Names, P/U | word | -1.31 | +0.34 | +0.69 | +0.83 | -0.43 | +0.82 | -0.10 |
| +C11: M/F Names, P/U | sent | -0.87 | +0.15 | +0.68 ✗ | +0.18 | -0.64 | +0.27 | -0.17 |
| +C11: M/F Names, P/U | c-word | NA | +0.14 | -0.44 | +0.27 | -0.35 | +0.46 ✓ | -0.13 |
| C6: M/F Names, Career/Family | word | +1.81 | -0.44 | -0.49 | -0.51 | -0.10 | -0.25 | -0.27 |
| C6: M/F Names, Career/Family | sent | +1.74 | -0.38 | -0.74 | -0.57 | +1.04 | +0.27 | +0.25 |
| C6: M/F Names, Career/Family | c-word | NA | -0.10 | +0.67 ✓ | -0.04 | +1.07 ! | +0.39 ✓ | -0.26 |
| C8: Science/Arts, M/F Terms | word | +1.24 | +0.24 | -0.23 | -0.15 | +0.25 | +0.51 | +0.87 |
| C8: Science/Arts, M/F Terms | sent | +1.01 | -0.30 | +0.11 | -0.16 | +0.89 | -0.15 | -0.15 |
| C8: Science/Arts, M/F Terms | c-word | NA | +0.16 | +1.02 ✓ | -0.08 | +1.03 ! | +0.64 ✓ | +0.67 ✓ |
| Double Bind M/F (Competent) | word | +1.62 | -0.34 | -0.35 | -0.26 | -0.66 | +1.00 | -0.04 |
| Double Bind M/F (Competent) | sent | +0.79 | -0.15 | -0.06 | 0.00 | +0.27 | +0.52 ✗ | +0.25 |
| Double Bind M/F (Competent) | c-word | NA | -0.07 | +0.42 ✓ | +0.02 | -0.02 | -0.94 | +0.57 ✓ |
| Double Bind M/F (Competent) | sent (u) | +0.84 | +0.21 | +0.39 | +0.60 | -0.76 | +1.26 ✗ | -0.59 |
| Double Bind M/F (Competent) | c-word (u) | NA | -0.48 | +0.46 | -0.37 | -0.36 | -0.72 | +0.56 |
| Double Bind M/F (Likable) | word | +1.29 | -0.61 | -1.37 | -0.64 | +0.15 | +0.83 | +0.02 |
| Double Bind M/F (Likable) | sent | +0.69 | -0.45 | -0.66 | -0.29 | -0.53 | -0.44 | -0.13 |
| Double Bind M/F (Likable) | c-word | NA | -0.38 | +0.64 ✓ | +0.13 | -0.03 | -0.68 | +0.50 ✓ |
| Double Bind M/F (Likable) | sent (u) | +0.51 | -0.92 | +0.74 | -0.97 | -1.57 | +0.25 | -1.01 |
| Double Bind M/F (Likable) | c-word (u) | NA | +0.20 | +1.29 ✓ | -0.78 | -1.22 | -0.98 | +0.39 |
| +Occ: M/F Names, Occ Terms | word | +1.59 | +0.63 | +0.55 | +0.65 | -0.38 | +0.76 | +0.46 |
| +Occ: M/F Names, Occ Terms | sent | +1.48 | +0.06 | +0.30 | +0.51 | +1.74 ✗ | -0.00 | -0.27 |
| +Occ: M/F Names, Occ Terms | c-word | NA | -0.27 | +0.98 ! | +0.67 ! | +0.10 | +0.27 ✓ | +0.43 ✓ |

Models trained on datasets with lower % of occupation associations overall exhibit smaller effect sizes at the contextual word level.

# Analysis: Race

| Test | Encoding | CBoW | ELMo | BERT (bbc) | BERT (blc) | GPT | GPT-2 (117M) | GPT-2 (345M) |
|---|---|---|---|---|---|---|---|---|
| C3: EA/AA Names, P/U | word | +1.41 | -0.41 | +0.38 | +0.63 | -1.06 | +1.34 | +0.54 |
| C3: EA/AA Names, P/U | sent | +0.52 | -0.38 | +0.73 | +1.04 | +0.65 | -0.14 | -0.30 |
| C3: EA/AA Names, P/U | c-word | NA | -0.02 | +0.93 ⚠ | +0.21 ⚠ | +1.05 ⚠ | +0.63 ✅ | +1.22 ✅ |
| +C12: EA/AA Names, Career/Family | word | -0.15 | -0.24 | -0.58 | -0.37 | -0.95 | -1.34 | -0.87 |
| +C12: EA/AA Names, Career/Family | sent | 0.00 | -0.18 | -0.50 | -0.66 | -0.69 | -0.17 | +0.30 |
| +C12: EA/AA Names, Career/Family | c-word | NA | -0.03 | -0.09 | -0.32 | -1.09 | +0.47 ✅ | +0.51 ⚠ |
| +C13: EA/AA Names, Science/Arts | word | -0.51 | -0.36 | -0.08 | +0.10 | +0.48 | +0.60 | +0.61 |
| +C13: EA/AA Names, Science/Arts | sent | +0.14 | -0.35 | +0.39 | -0.03 | -0.11 | +0.31 | -0.13 |
| +C13: EA/AA Names, Science/Arts | c-word | NA | +0.02 | +0.90 ✅ | -0.25 | +0.18 | +0.03 | -0.06 |
| +Double Bind EA/AA (Competent) | word | +1.49 | +0.22 | +0.90 | +1.20 | -0.66 | +1.21 | +0.09 |
| +Double Bind EA/AA (Competent) | sent | +1.03 | +0.14 | +1.19 | +1.05 | +0.35 | -0.30 | +0.42 ❌ |
| +Double Bind EA/AA (Competent) | c-word | NA | +0.10 | +0.91 ⚠ | +0.31 ⚠ | +0.77 ⚠ | -0.81 | -0.01 |
| +Double Bind EA/AA (Competent) | sent (u) | +1.15 | -0.33 | +1.23 | +1.03 | +1.17 | -0.78 | +0.44 |
| +Double Bind EA/AA (Competent) | c-word (u) | NA | +0.06 | +1.01 ⚠ | +0.70 ⚠ | +0.78 ⚠ | -0.70 | +0.59 ✅ |
| +Double Bind EA/AA (Likable) | word | +1.62 | +0.38 | +0.79 | +0.60 | -0.56 | +1.33 | +0.06 |
| +Double Bind EA/AA (Likable) | sent | +1.24 | +0.28 | +1.14 | +0.90 | -0.04 | +0.38 ❌ | -0.48 |
| +Double Bind EA/AA (Likable) | c-word | NA | +0.22 ⚠ | +0.61 ⚠ | +0.21 ⚠ | +0.66 ✅ | -0.79 | -0.07 |
| +Double Bind EA/AA (Likable) | sent (u) | +1.29 | +0.42 | +1.30 ❌ | +1.02 | +0.51 | -0.53 | +0.51 |
| +Double Bind EA/AA (Likable) | c-word (u) | NA | -0.17 | -0.34 | +0.87 ⚠ | -0.42 | -0.76 | -0.90 |

Models exhibit more significant effect sizes on tests relating to pleasantness, competence, likability, than on tests relating to career/family or science/art.

## Contributions

1. Either sentence encoding or contextual word representations can uncover latent social bias that the other cannot.
2. Models exhibit more bias for identities at an intersection of race and gender than constituent minorities.

## Limitations

1. No significant positive associations ⇏ no social bias
2. Assumption of binary gender

Thank You!

Poster:
10:45 AM -- 12:45 PM
@ East Exhibition Hall B + C
#73