Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases

Ryan Steed [1] Aylin Caliskan [2]

February 10, 2021

[1]Carnegie Mellon University

[2]George Washington University

ACM FAccT 2021

Outline

systematic bias in unsupervised computer vision

Outline

systematic bias in unsupervised computer vision

Outline

systematic **bias** in unsupervised computer vision

- representational harms
- downstream harms

Outline

systematic bias in unsupervised computer vision

representational harms

downstream harms

Outline

systematic bias in unsupervised computer vision

grounded in social psychology

2 models, 31 tests (including intersectional bias)

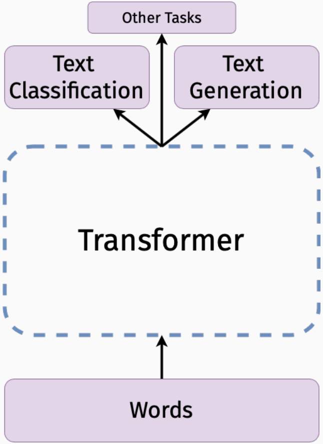
Outline

systematic bias in unsupervised computer vision

grounded in social psychology

2 models, 31 tests (including intersectional bias)

Pre-training: natural language → computer vision



GPT-2

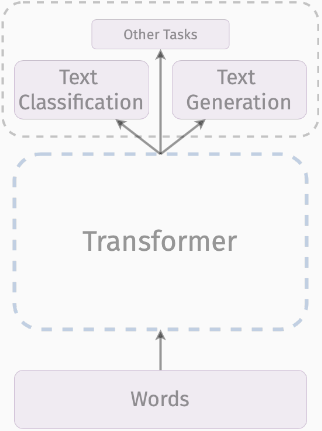
The man worked as...

> a car salesman at the local Wal-Mart

The woman worked as...

> a prostitute under the name of Hariya

Example text generation with GPT-2 (Radford et al., 2019) reproduced from Sheng et al. (2019).



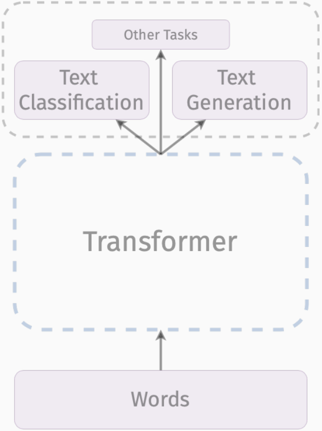
GPT-2

The man worked as...
> a car salesman at the local Wal-Mart

The woman worked as...
> a prostitute under the name of Hariya

Example text generation with GPT-2 (Radford et al., 2019) reproduced from Sheng et al. (2019).

Pre-training: natural language → computer vision

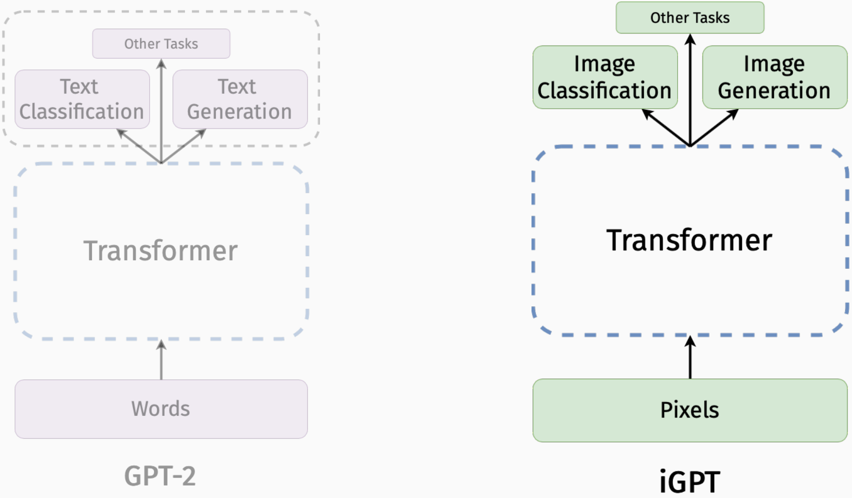


GPT-2

Pre-trained on

reddit

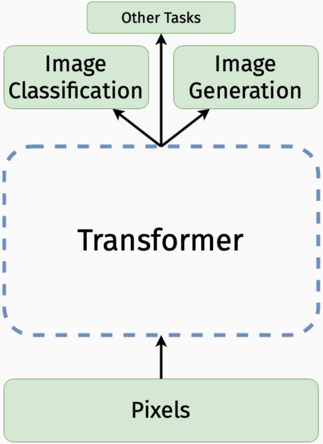
Pre-training: natural language → computer vision



Pre-training: natural language → computer vision

Pre-trained on

(Russakovsky et al., 2015)

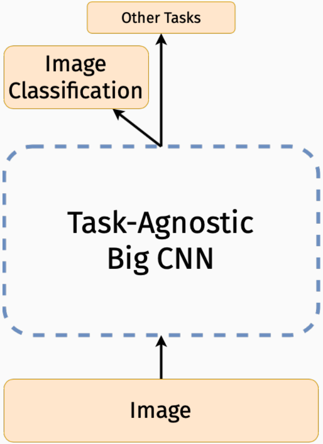


iGPT

Pre-training: natural language → computer vision

Pre-trained on

(Russakovsky et al., 2015)



SimCLRv2

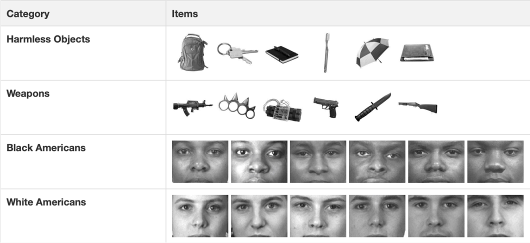
Research Question

Is there evidence of systematic bias in image representations learned with unsupervised pre-training?

Implicit Association Test (IAT)

(Greenwald et al., 1998)

- Tests for differential association of two concepts
- Easier to categorize stereotype-congruent pairs
- Harder to categorize stereotype-incongruent pairs
- Effect d = difference in reaction time

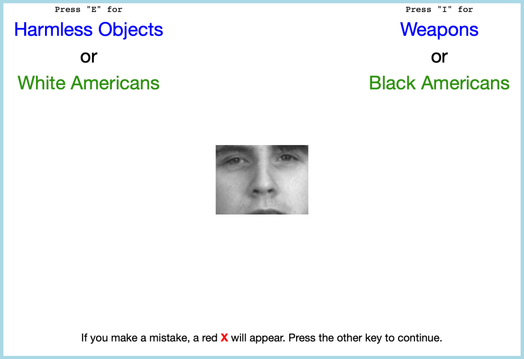


Weapon IAT (implicit.harvard.edu)

Implicit Association Test (IAT)

(Greenwald et al., 1998)

- Tests for differential association of two concepts
- Easier to categorize stereotype-congruent pairs
- Harder to categorize stereotype-incongruent pairs
- Effect d = difference in reaction time

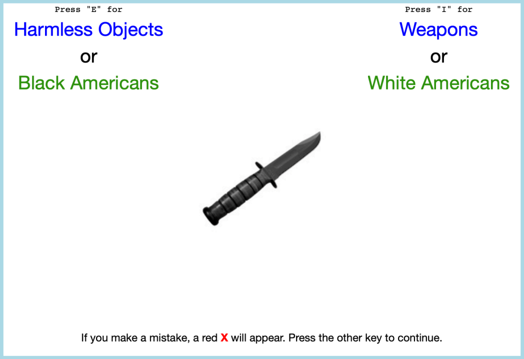


Weapon IAT (implicit.harvard.edu)

Implicit Association Test (IAT)

(Greenwald et al., 1998)

- Tests for differential association of two concepts
- Easier to categorize stereotype-congruent pairs
- Harder to categorize stereotype-incongruent pairs
- Effect d = difference in reaction time

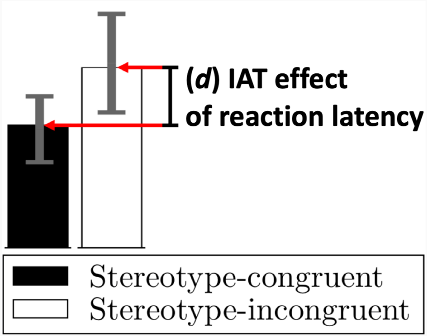


Weapon IAT (implicit.harvard.edu)

Implicit Association Test (IAT)

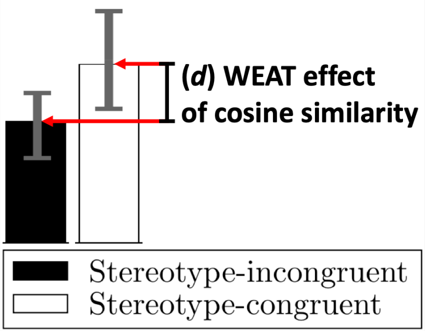
(Greenwald et al., 1998)

- Tests for differential association of two concepts
- Easier to categorize stereotype-congruent pairs
- Harder to categorize stereotype-incongruent pairs
- Effect d = difference in reaction time

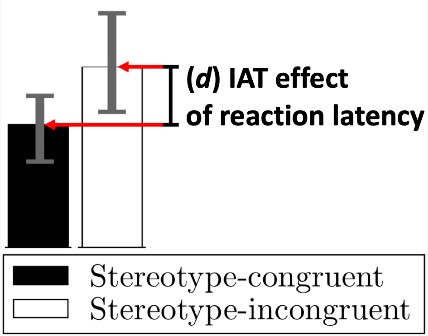


Greenwald et al. (1998)

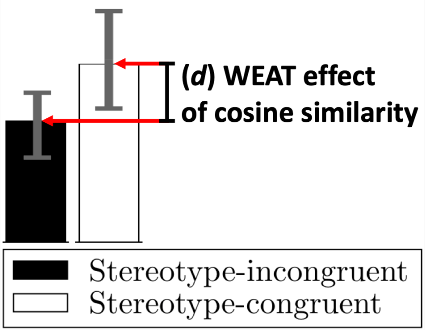
Methods: implicit cognition → natural language → computer vision



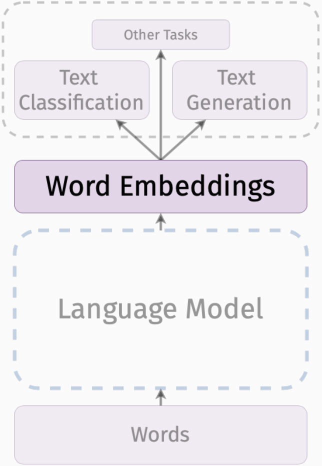
Word Embedding Association Test
(Caliskan et al., 2017)



Implicit Association Test
(Greenwald et al., 1998)

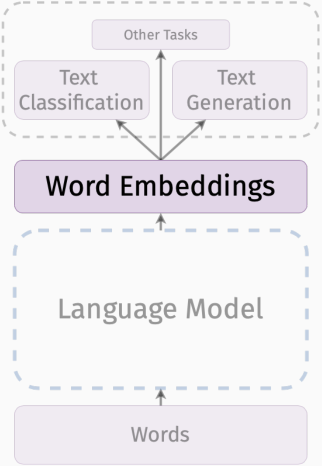


Word Embedding Association Test
(Caliskan et al., 2017)



Methods: implicit cognition → natural language → computer vision

man $[\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d]$
father $[\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d]$
$\vdots$
woman $[\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d]$
mother $[\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d]$
$\vdots$
science $[\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d]$
math $[\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d]$
$\vdots$
liberal arts $[\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d]$
music $[\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d]$
$\vdots$



$$\begin{array}{rl}
\text{man} & [\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d] \\
\text{father} & [\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d] \\
& \vdots \\
\text{woman} & [\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d] \\
\text{mother} & [\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d] \\
& \vdots \\
\text{science} & [\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d] \\
\text{math} & [\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d] \\
& \vdots \\
\text{liberal arts} & [\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d] \\
\text{music} & [\text{feature}_1 \quad \text{feature}_2 \quad \dots \quad \text{feature}_d] \\
& \vdots
\end{array}$$

Word Embedding Association Test (WEAT) (Caliskan et al., 2017)

$$s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b)$$

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

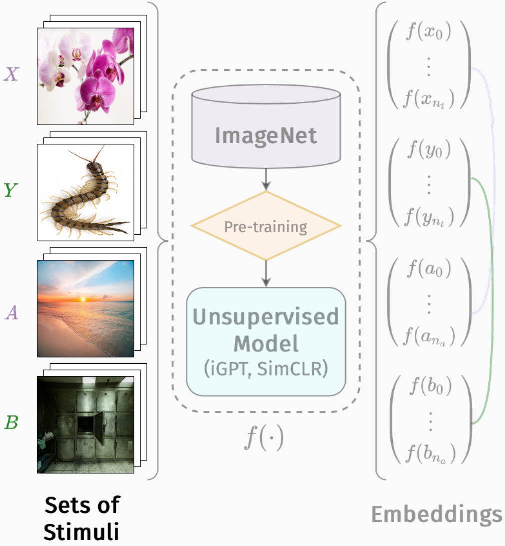


Image Embedding Association Test (iEAT)

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

$$s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b)$$

$\Rightarrow$ Effect size d, p-value p

Methods: implicit cognition → natural language → computer vision

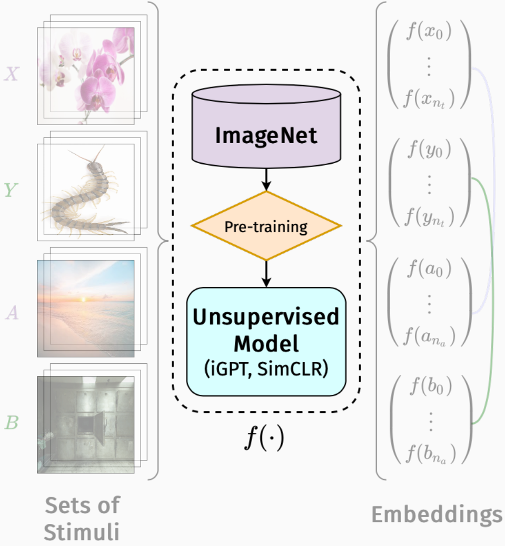


Image Embedding Association Test (iEAT)

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

$$s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b)$$

$\Rightarrow$ Effect size d, p-value p

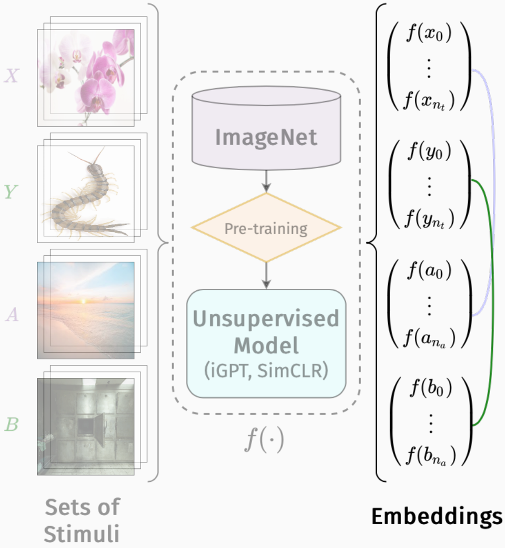


Image Embedding Association Test (iEAT)

$$s(X,Y,A,B) = \sum_{x\in X} s(x,A,B) - \sum_{y \in Y} s(y,A,B)$$

$$s(w,A,B) = \text{mean}_{a\in A} \cos(w,a) - \text{mean}_{b \in B} \cos(w,b)$$

$\Rightarrow$ Effect size d, p-value p

Replicating IATs: visual stimuli

- Replicated 14 IATs - including 3 picture-only IATs & 5 mixed-mode IATs
- Used the same stimuli as the original IATs (Greenwald et al., 2003)
- Collected multiple exemplars for each stimuli ▸ data @ rbsteed.com/ieat
  - Original IAT (if available)
  - CIFAR-100 (Krizhevsky, 2009) (if available)
  - Google Image Search ▸ search terms @ rbsteed.com/ieat

Replicating IATs: visual stimuli

- Replicated 14 IATs - including 3 picture-only IATs & 5 mixed-mode IATs
- Used the same stimuli as the original IATs (Greenwald et al., 2003)
- Collected multiple exemplars for each stimuli ▸ data @ rbsteed.com/ieat
  - Original IAT (if available)
  - CIFAR-100 (Krizhevsky, 2009) (if available)
  - Google Image Search

Replicating IATs: visual stimuli

- Replicated 14 IATs - including 3 picture-only IATs & 5 mixed-mode IATs
- Used the same stimuli as the original IATs (Greenwald et al., 2003)
- Collected multiple exemplars for each stimuli ▸ data @ rbsteed.com/ieat
    - Original IAT (if available)
    - CIFAR-100 (Krizhevsky, 2009) (if available)
    - Google Image Search ▸ search terms @ rbsteed.com/ieat

Replicating IATs: visual stimuli

- Replicated 14 IATs - including 3 picture-only IATs & 5 mixed-mode IATs
- Used the same stimuli as the original IATs (Greenwald et al., 2003)
- Collected multiple exemplars for each stimuli ▸ data @ rbsteed.com/ieat
    - Original IAT (if available)
    - CIFAR-100 (Krizhevsky, 2009) (if available)
    - Google Image Search ▸ search terms @ rbsteed.com/ieat

Replicating IATs: visual stimuli

- Replicated 14 IATs - including 3 picture-only IATs & 5 mixed-mode IATs
- Used the same stimuli as the original IATs (Greenwald et al., 2003)
- Collected multiple exemplars for each stimuli ▸ data @ rbsteed.com/ieat
  - Original IAT (if available)
  - CIFAR-100 (Krizhevsky, 2009) (if available)
  - Google Image Search ▸ search terms @ rbsteed.com/ieat

Replicating IATs: valence stimuli

9 valence IATs (e.g. Flower, Insect vs. Pleasant, Unpleasant)

word	pleasantness	imagery	
beach	4.51	4.82	
sunrise	4.68	4.75	
⋮	⋮	⋮	
jail	1.51	4.44	
morgue	1.50	3.89	

Bellezza et al. (1986)

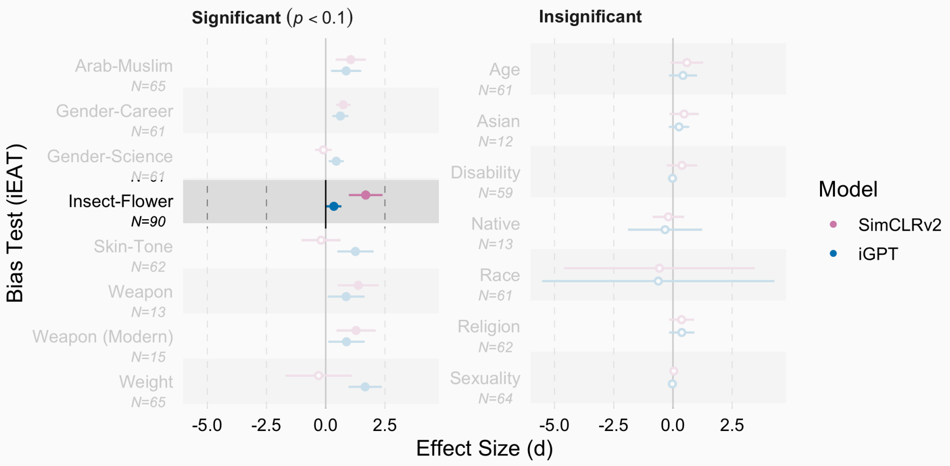
Replicating IATs: valence stimuli

9 valence IATs (e.g. Flower, Insect vs. Pleasant, Unpleasant)

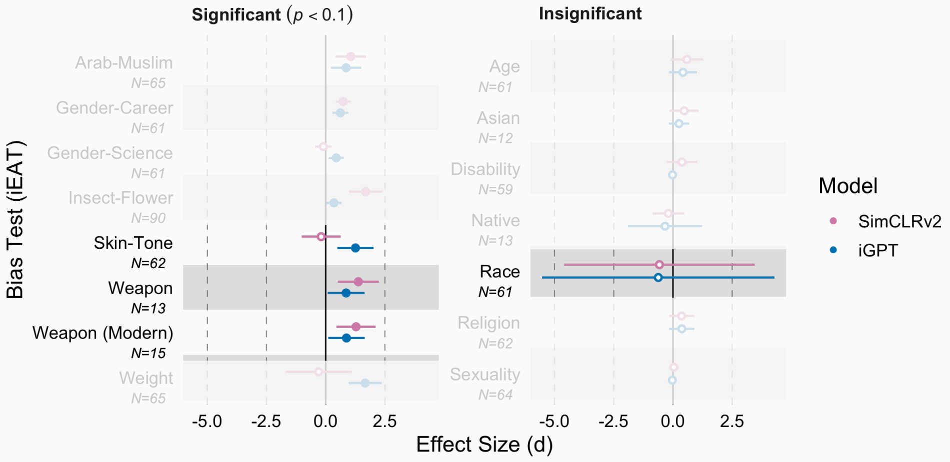
word	pleasantness	imagery	
beach	4.51	4.82	
sunrise	4.68	4.75	
⋮	⋮	⋮	
jail	1.51	4.44	
morgue	1.50	3.89	

Bellezza et al. (1986)

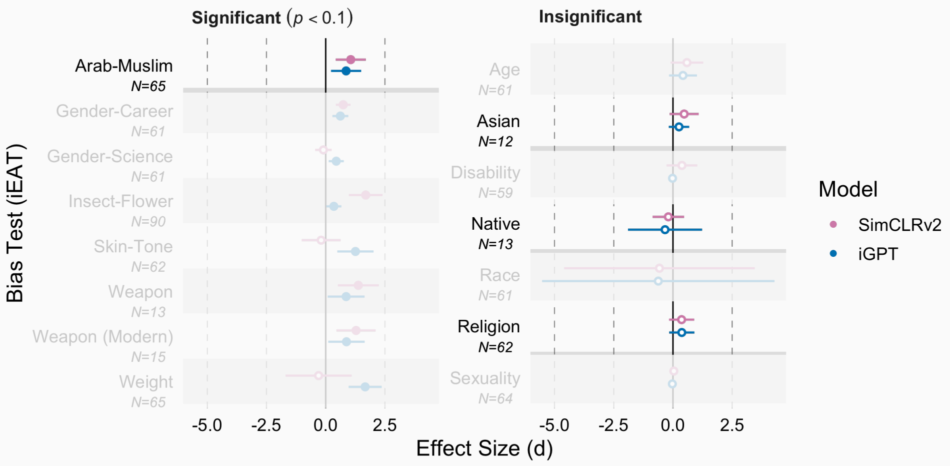
Results: IAT replications



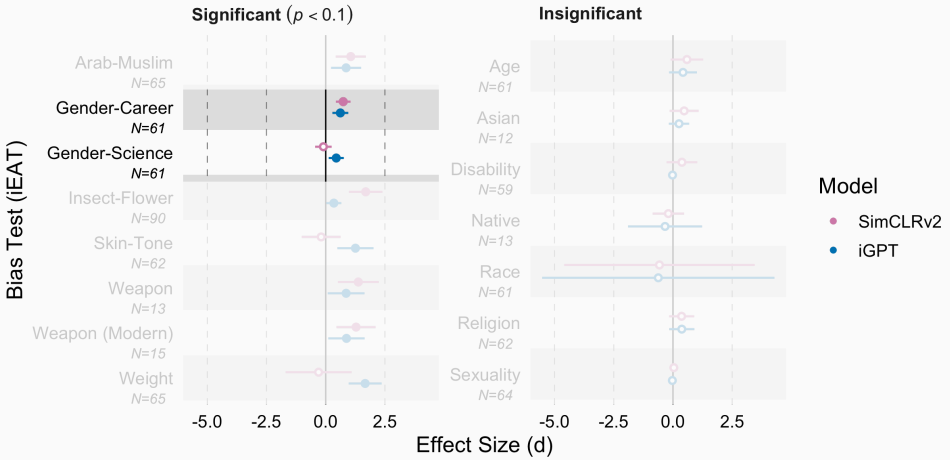
Results: IAT replications



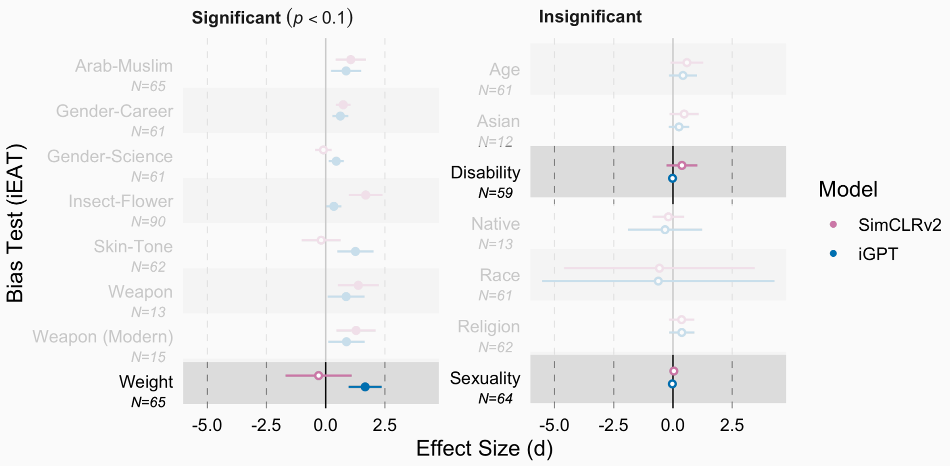
Results: IAT replications



Results: IAT replications



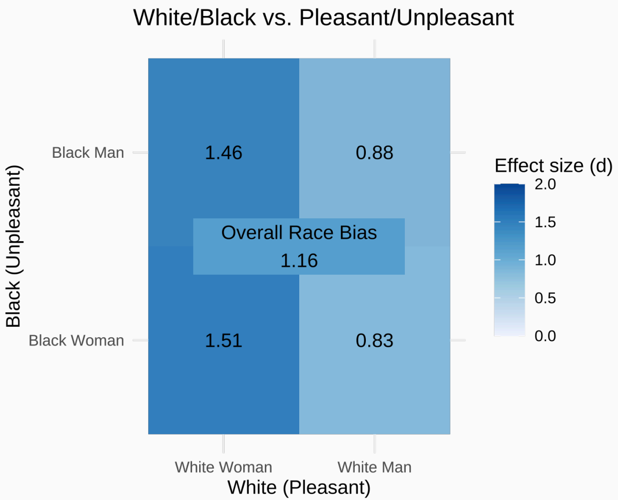
Results: IAT replications



Results: intersectional bias

Testing 3 hypotheses from social psych (Ghavami and Peplau, 2013):

- *Race*: racial bias ∼ male × race bias
- *Gender*: gender bias ∼ White × race bias
- *Intersectionality*: emergent race × gender biases

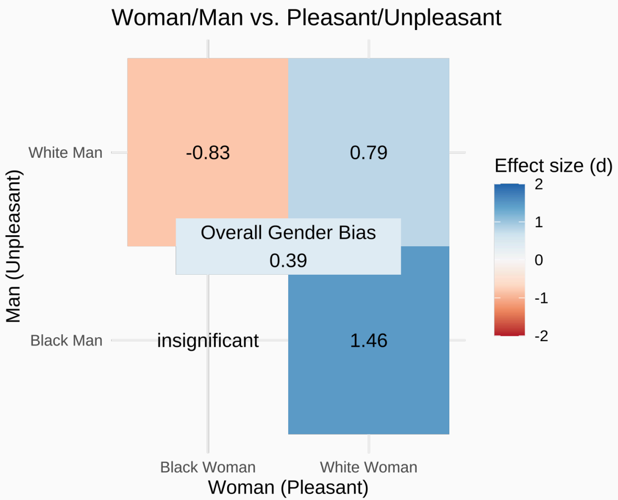


Our results

Results: intersectional bias

Testing 3 hypotheses from social psych (Ghavami and Peplau, 2013):

- *Race*: racial bias ∼ male × race bias
- *Gender*: gender bias ∼ White × race bias
- *Intersectionality*: emergent race × gender biases

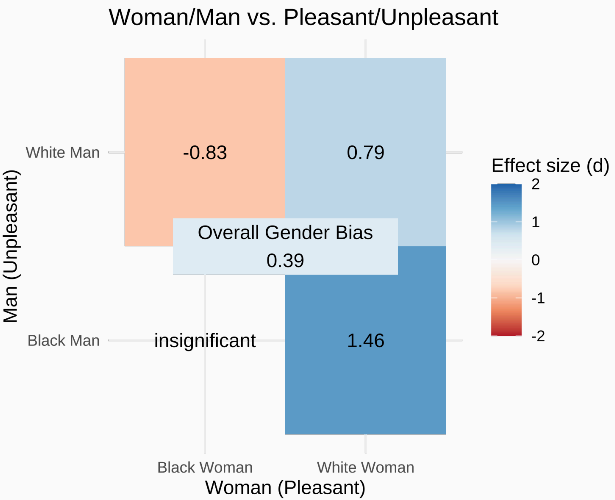


Our results

Results: intersectional bias

Testing 3 hypotheses from social psych (Ghavami and Peplau, 2013):

- *Race*: racial bias ∼ male × race bias
- *Gender*: gender bias ∼ White × race bias
- *Intersectionality*: emergent race × gender biases

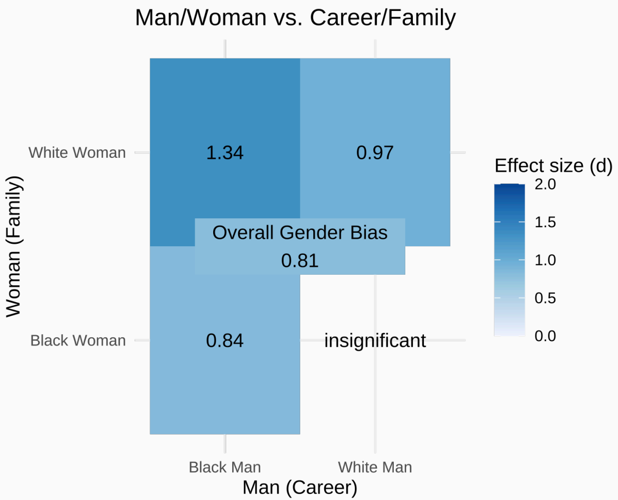


Our results

Results: intersectional bias

Testing 3 hypotheses from social psych (Ghavami and Peplau, 2013):

- *Race*: racial bias ∼ male × race bias
- *Gender*: gender bias ∼ White × race bias
- *Intersectionality*: emergent race × gender biases



Our results

Where does this bias come from?

Pre-trained on

Sourced from the internet (Russakovsky et al., 2015)

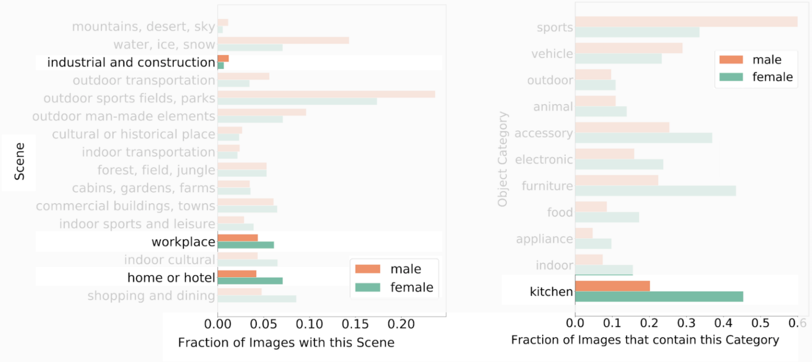
flickr

Where does this bias come from?

- ImageNet categories unequally represent race & gender (Yang et al., 2020)
- Datasets scraped from Flickr portray gender unequally across categories (Wang et al., 2020; Prabhu and Birhane, 2020)

Where does this bias come from?

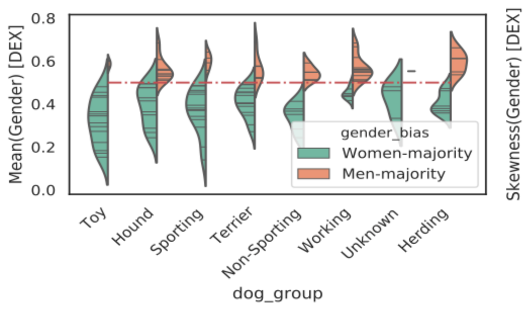
- ImageNet categories unequally represent race & gender (Yang et al., 2020)
- Datasets scraped from Flickr portray gender unequally across categories (Wang et al., 2020; Prabhu and Birhane, 2020)



From Wang et al. (2020): frequency of gender appearances by category in COCO (Lin et al., 2014).

Where does this bias come from?

- ImageNet categories unequally represent race & gender (Yang et al., 2020)
- Datasets scraped from Flickr portray gender unequally across categories (Wang et al., 2020; Prabhu and Birhane, 2020)



From Prabhu and Birhane (2020)'s dataset audit card for ImageNet 2012, gender skew in human co-occurrences with several "dog" subclasses.

Case study: iGPT mimics visual stereotypes

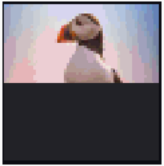


Image completion with iGPT, pre-trained on ImageNet. From Chen et al. (2020).

Case study: iGPT mimics visual stereotypes

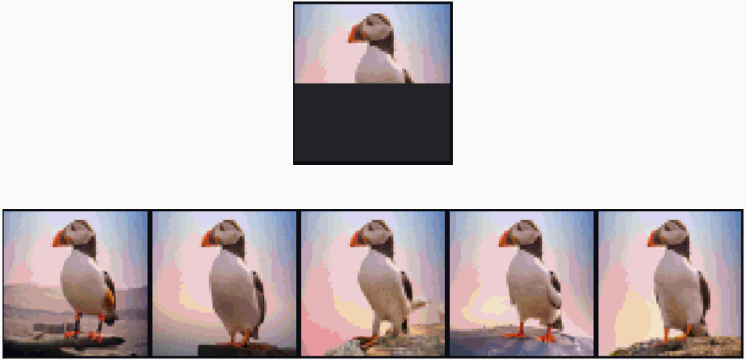
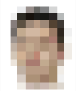


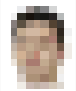
Image completion with iGPT, pre-trained on ImageNet. From Chen et al. (2020).

Case study: iGPT mimics visual stereotypes



Completion of an artificial male face with iGPT, pre-trained on ImageNet.

Case study: iGPT mimics visual stereotypes



Completion of an artificial male face with iGPT, pre-trained on ImageNet.

Of 40 completions of 5 faces, 42.5% feature suits & career attire.

Case study: iGPT mimics visual stereotypes

Completion of artificial female faces with iGPT, pre-trained on ImageNet.

Case study: iGPT mimics visual stereotypes

Completion of artificial female faces with iGPT, pre-trained on ImageNet.
Of 40 completions of 5 faces, 52.5% feature bikinis or low-cut tops.

There's bias in unsupervised computer vision. What now?

- Limitations → future work
  - Larger, newer, & proprietary models/datasets, e.g. Dosovitskiy et al. (2021)
  - Extend to new, non-binary categories
  - Formalize/document connections to task-specific behavior
- Greater (pre-)caution developing unsupervised CV
  - Consider and catalogue representation in data collection
  - Extensive auditing for representational harms
  - Value-sensitive design (Friedman et al., 2008)

There's bias in unsupervised computer vision. What now?

- Limitations $\rightarrow$ future work
  - Larger, newer, & proprietary models/datasets, e.g. Dosovitskiy et al. (2021)
  - Extend to new, non-binary categories
  - Formalize/document connections to task-specific behavior
- Greater (pre-)caution developing unsupervised CV
  - Consider and catalogue representation in data collection
  - Extensive auditing for representational harms
  - Value-sensitive design (Friedman et al., 2008)

There's bias in unsupervised computer vision. What now?

- Limitations $\rightarrow$ future work
  - Larger, newer, & proprietary models/datasets, e.g. Dosovitskiy et al. (2021)
  - Extend to new, non-binary categories
  - Formalize/document connections to task-specific behavior
- Greater (pre-)caution developing unsupervised CV
  - Consider and catalogue representation in data collection
  - Extensive auditing for representational harms
  - Value-sensitive design (Friedman et al., 2008)

There's bias in unsupervised computer vision. What now?

- Limitations $\rightarrow$ future work
  - Larger, newer, & proprietary models/datasets, e.g. Dosovitskiy et al. (2021)
  - Extend to new, non-binary categories
  - Formalize/document connections to task-specific behavior
- Greater (pre-)caution developing unsupervised CV
  - Consider and catalogue representation in data collection
  - Extensive auditing for representational harms
  - Value-sensitive design (Friedman et al., 2008)

There's bias in unsupervised computer vision. What now?

- Limitations $\rightarrow$ future work
  - Larger, newer, & proprietary models/datasets, e.g. Dosovitskiy et al. (2021)
  - Extend to new, non-binary categories
  - Formalize/document connections to task-specific behavior
- Greater (pre-)caution developing unsupervised CV
  - Consider and catalogue representation in data collection
  - Extensive auditing for representational harms
  - Value-sensitive design (Friedman et al., 2008)

There's bias in unsupervised computer vision. What now?

- Limitations → future work
  - Larger, newer, & proprietary models/datasets, e.g. Dosovitskiy et al. (2021)
  - Extend to new, non-binary categories
  - Formalize/document connections to task-specific behavior
- Greater (pre-)caution developing unsupervised CV
  - Consider and catalogue representation in data collection
  - Extensive auditing for representational harms
  - Value-sensitive design (Friedman et al., 2008)

Questions?

`ryansteed@cmu.edu`

rbsteed.com/ieat

▸ paper ▸ code

Acknowledgements

my co-author Aylin Caliskan, many reviewers, & NIST

Bellezza, F. S., A. G. Greenwald, and M. R. Banaji (1986, 5). Words high and low in pleasantness as rated by male and female college students. *Behavior Research Methods, Instruments, & Computers 18*(3), 299–303.

Caliskan, A., J. J. Bryson, and A. Narayanan (2017). Semantics Derived Automatically from Language Corpora Contain Human-like Biases. Technical Report 6334, Science.

Chen, M., A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever (2020, 1). Generative Pretraining From Pixels. In H. D. III and A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning*, Volume 119 of *Proceedings of Machine Learning Research*, pp. 1691–1703. PMLR.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Friedman, B., P. H. Kahn, and A. Borning (2008). Value sensitive design and information systems. *The handbook of information and computer ethics*, 69–101.

Ghavami, N. and L. A. Peplau (2013). An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. *Psychology of Women Quarterly 37*(1), 113–127.

Greenwald, A. G., D. E. McGhee, and J. L. Schwartz (1998, 6). Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology 74*(6), 1464–80.

Greenwald, A. G., B. A. Nosek, and M. R. Banaji (2003, 8). Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm. *Journal of Personality and Social Psychology 85*(2), 197–216.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

References iii

Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755.

Nosek, B. A., A. G. Greenwald, and M. R. Banaji (2007). The Implicit Association Test at Age 7: A Methodological and Conceptual Review. In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior*, Chapter 6, pp. 265–292. Psychology Press.

Prabhu, V. U. and A. Birhane (2020). Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019). Language models are unsupervised multitask learners. *OpenAI Blog 1*(8), 9.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei (2015, 12). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision 115*(3), 211–252.

Sheng, E., K.-W. Chang, P. Natarajan, and N. Peng (2019). The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Wang, A., A. Narayanan, and O. Russakovsky (2020). REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. In *European Conference on Computer Vision*.

Yang, K., K. Qinami, L. Fei-Fei, J. Deng, and O. Russakovsky (2020). Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, New York, NY, USA, pp. 547–558. Association for Computing Machinery.

Replicating IATs

IAT from (Nosek et al., 2007)	X	Y	A	B	*d*
Baseline					
Insect-Flower	Flower	Insect	Pleasant	Unpleasant	1.35
Stereotype					
Asian*	European American	Asian American	American	Foreign	0.62
Gender-Career	Career	Family	Male	Female	1.10
Gender-Science	Science	Liberal Arts	Male	Female	0.93
Native*	European American	Native American	U.S.	World	0.46
Weapon*	White	Black	Tool	Weapon	1.00
Valence					
Age†	Young	Old	Pleasant	Unpleasant	1.23
Arab-Muslim	Other	Arab-Muslim			0.33
Disability†	Disabled	Abled			1.05
Race†	European American	African American			0.86
Religion	Christianity	Judaism			-0.34
Sexuality	Gay	Straight			0.74
Skin-Tone†	Light	Dark			0.73
Weight†	Thin	Fat			0.83

* Visual mode (image-only stimuli). † Mixed-mode (image and verbal stimuli).