Bias in NLP systems

There is a classic riddle: *A man and his son get into a terrible car crash. The father dies, and the boy is badly injured. In the hospital, the surgeon looks at the patient and exclaims, "I can't operate on this boy, he's my son!"* **How can this be?**

(Rudinger et al. 2018)

Fairness

- Representational: are different groups represented differently, reinforcing societal stereotypes (e.g. women are homemakers, men are breadwinners)?
- Allocational: so certain groups lose out on professional or financial opportunities (e.g. women receive lower credit limits)?
- Relationship to interpretability: when a system we do not understand makes decisions with significant societal consequences, there’s a greater potential for unfair outcome

Fairness in machine learning

Apple Card is accused of gender bias. Here's how that can happen

By Evelina Nedlund, CNN Business

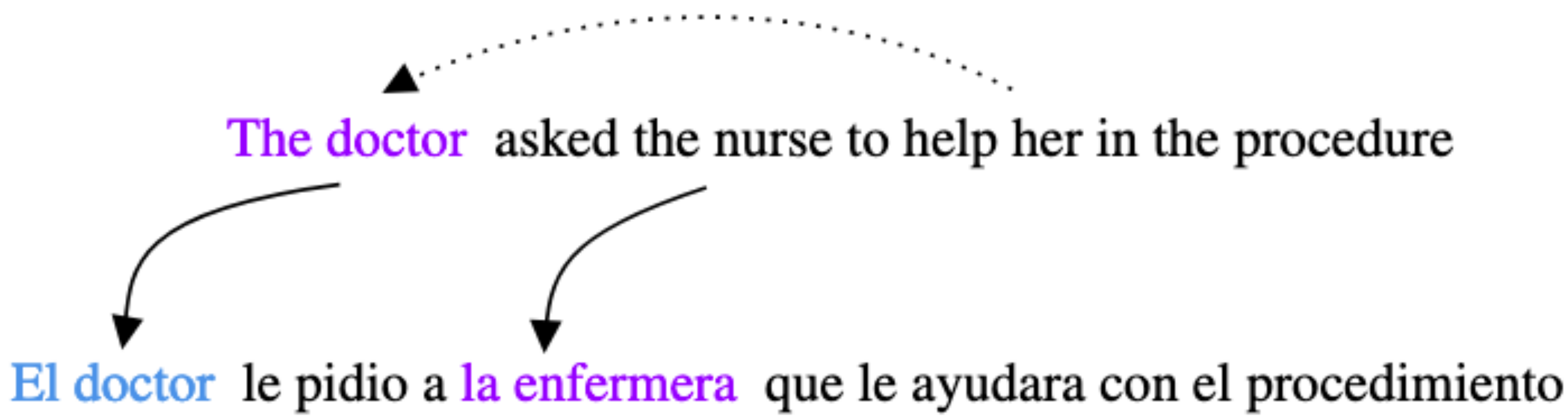
Updated 2:04 PM ET, Tue November 12, 2019

New York (CNN Business) — Some Apple Card customers say the credit card's issuer, Goldman Sachs, is giving women far lower credit limits, even if they share assets and accounts with their spouse. But it's impossible to know if the Apple Card -- or any other credit card -- discriminates against women, because creditworthiness algorithms are notoriously opaque.

"It's such a mystery we are seeing," said Sara Rathner, travel and credit cards expert at NerdWallet. "Because we don't know exactly what those algorithms are looking for, it can be hard to say if there might be some bias built into them."

https://www.cnn.com/2019/11/12/business/apple-card-gender-bias/index.html

Fairness in machine learning

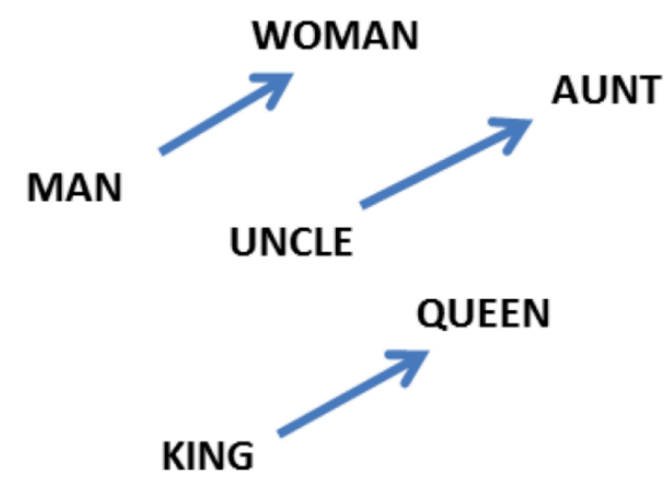


Gender bias leads to an incorrect translation!

(Stanovsky et al, 2019)

Bias in word embeddings

$$\overrightarrow{\mathrm{man}} - \overrightarrow{\mathrm{woman}} \approx \overrightarrow{\mathrm{computer\ programmer}} - \overrightarrow{\mathrm{homemaker}}.$$



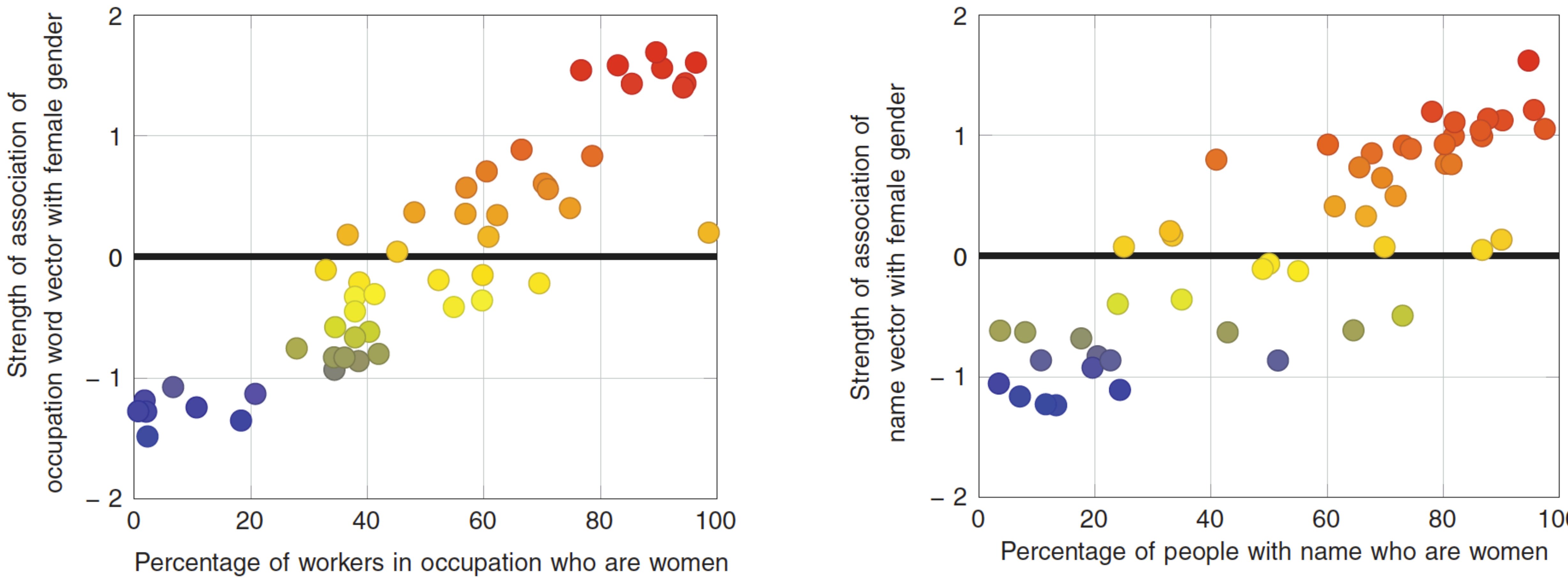
Extreme *she*	Extreme *he*
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

	Gender stereotype *she*-*he* analogies	
sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant
	Gender appropriate *she*-*he* analogies	
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

(Projection onto *he* - *she*)

(Bolukbasi et al 2016)

Biases in word embeddings induced from a corpus reflect social reality



- It is undesirable for the meaning of “doctor” to include “male”!
- Faithfulness to corpus statistics can lead to fairness issues

(Caliskan et al 2017)

The Implicit Association Test

Sequence	1	2	3	4	5
Task description	*Initial target-concept discrimination*	*Associated attribute discrimination*	*Initial combined task*	*Reversed target-concept discrimination*	*Reversed combined task*
Task instructions	● BLACK WHITE ●	● pleasant unpleasant ●	● BLACK ● pleasant WHITE ● unpleasant ●	BLACK ● ● WHITE	BLACK ● ● pleasant ● WHITE unpleasant ●
Sample stimuli	MEREDITH ○ ○ LATONYA ○ SHAVONN HEATHER ○ ○ TASHIKA KATIE ○ BETSY ○ ○ EBONY	○ lucky ○ honor poison ○ grief ○ ○ gift disaster ○ ○ happy hatred ○	○ JASMINE ○ pleasure PEGGY ○ evil ○ COLLEEN ○ ○ miracle ○ TEMEKA bomb ○	○ COURTNEY ○ STEPHANIE SHEREEN ○ ○ SUE-ELLEN TIA ○ SHARISE ○ ○ MEGAN NICHELLE ○	○ peace LATISHA ○ filth ○ ○ LAUREN ○ rainbow SHANISE ○ accident ○ ○ NANCY

(Greenwald et al 1998)

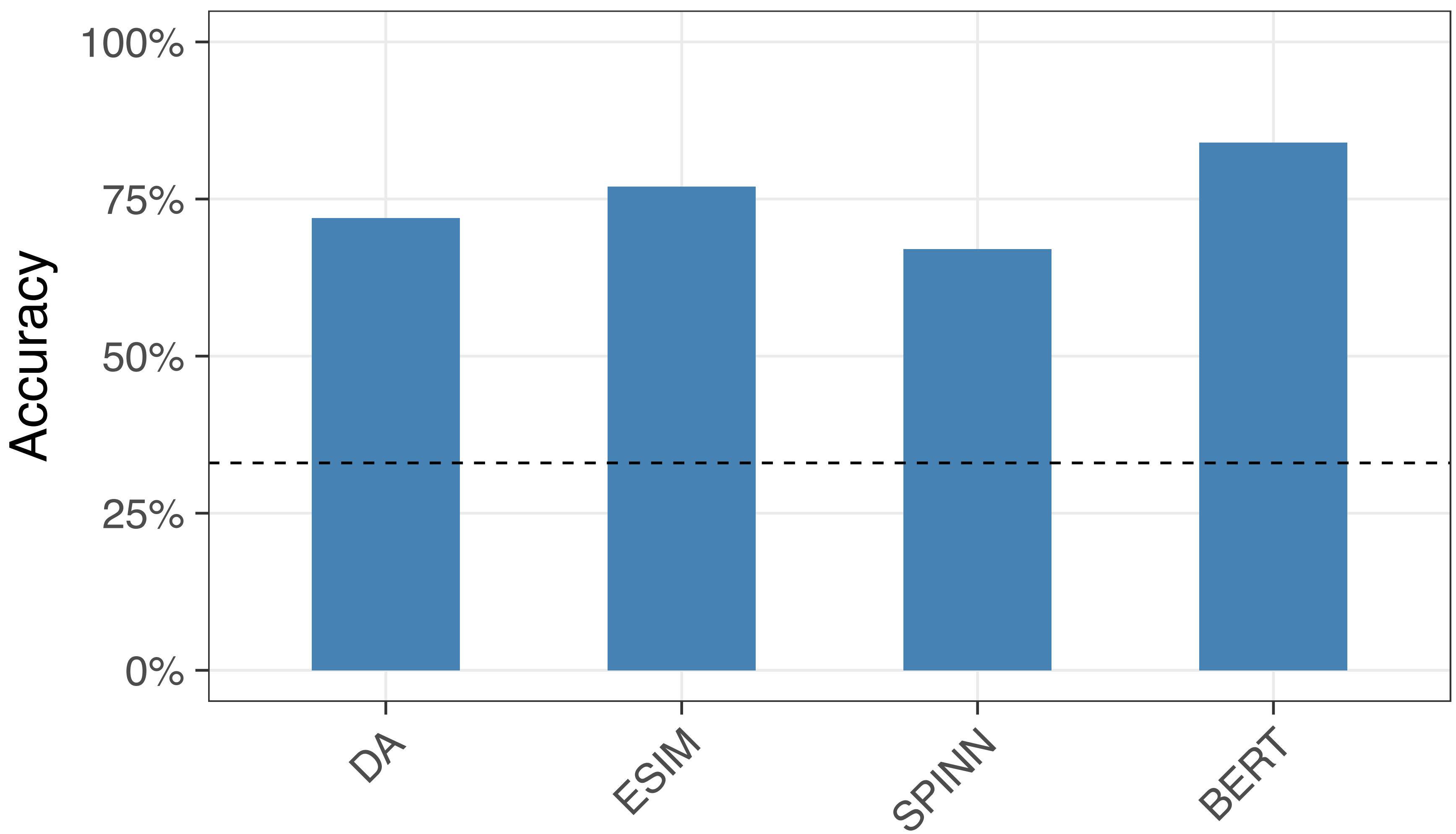
Word embeddings show similar biases to humans in an Implicit Association Test

Target words	Attribute words	Original finding				Our finding			
		Ref.	N	d	P	N_T	N_A	d	P
Flowers vs. insects	Pleasant vs. unpleasant	(*5*)	32	1.35	10^{-8}	25 × 2	25 × 2	1.50	10^{-7}
Instruments vs. weapons	Pleasant vs. unpleasant	(*5*)	32	1.66	10^{-10}	25 × 2	25 × 2	1.53	10^{-7}
European-American vs. African-American names	Pleasant vs. unpleasant	(*5*)	26	1.17	10^{-5}	32 × 2	25 × 2	1.41	10^{-8}
European-American vs. African-American names	Pleasant vs. unpleasant from (*5*)	(*7*)	Not applicable			16 × 2	25 × 2	1.50	10^{-4}
European-American vs. African-American names	Pleasant vs. unpleasant from (*9*)	(*7*)	Not applicable			16 × 2	8 × 2	1.28	10^{-3}
Male vs. female names	Career vs. family	(*9*)	39k	0.72	$<10^{-2}$	8 × 2	8 × 2	1.81	10^{-3}
Math vs. arts	Male vs. female terms	(*9*)	28k	0.82	$<10^{-2}$	8 × 2	8 × 2	1.06	.018
Science vs. arts	Male vs. female terms	(*10*)	91	1.47	10^{-24}	8 × 2	8 × 2	1.24	10^{-2}
Mental vs. physical disease	Temporary vs. permanent	(*23*)	135	1.01	10^{-3}	6 × 2	7 × 2	1.38	10^{-2}
Young vs. old people's names	Pleasant vs. unpleasant	(*9*)	43k	1.42	$<10^{-2}$	8 × 2	8 × 2	1.21	10^{-2}

(Caliskan et al 2017)

The challenge set approach

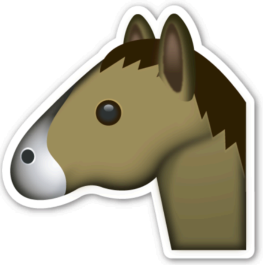
- Take a system trained on a standard dataset (e.g. MNLI)…

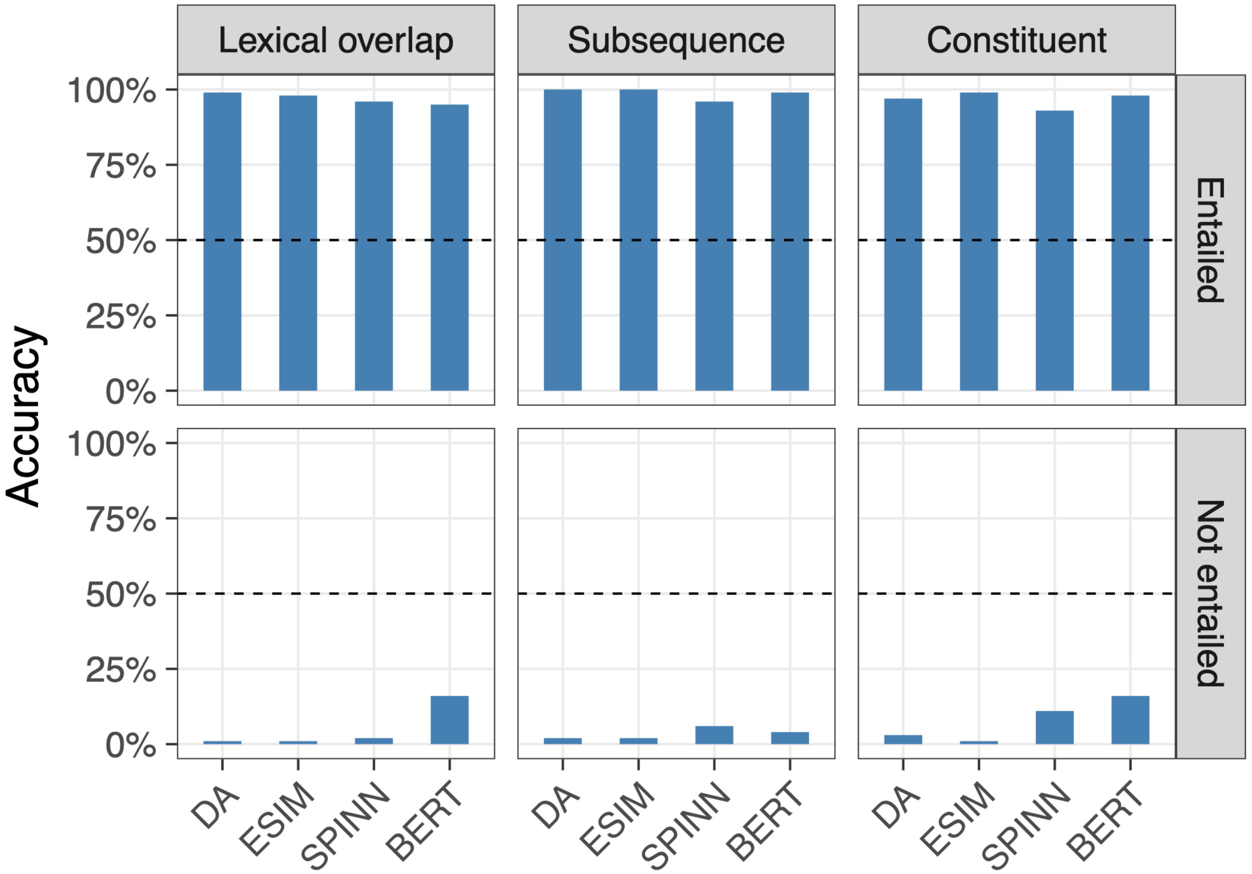


The challenge set approach

- And test it on a dataset constructed to test for a particular deficiency or bias

Results on HANS





Heuristic makes correct prediction

Heuristic makes incorrect prediction

Coreference resolution and the Winograd Schema Challenge

- In many cases, coreference resolution requires going beyond syntactic constraints and using “world knowledge”

- *The trophy didn’t fit into the suitcase because* ***it*** *was too large.*

- *The trophy didn’t fit into the suitcase because* ***it*** *was too small.*

- *Joan made sure to thank Susan for all the help* ***she*** *had given.*

(Levesque et al 2011)

The Winogender challenge

(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

(1b) **The paramedic** performed CPR on someone even though she/he/they knew it was too late.

(2b) The paramedic performed CPR on **someone** even though she/he/they was/were already dead.

(Correct answers in bold)

1. **OCCUPATION** , a person referred to by their occupation and a definite article, e.g., "the paramedic."

2. **PARTICIPANT** , a secondary (human) participant, e.g., "the passenger."

3. **PRONOUN** , a pronoun that is coreferent with either OCCUPATION or PARTICIPANT.

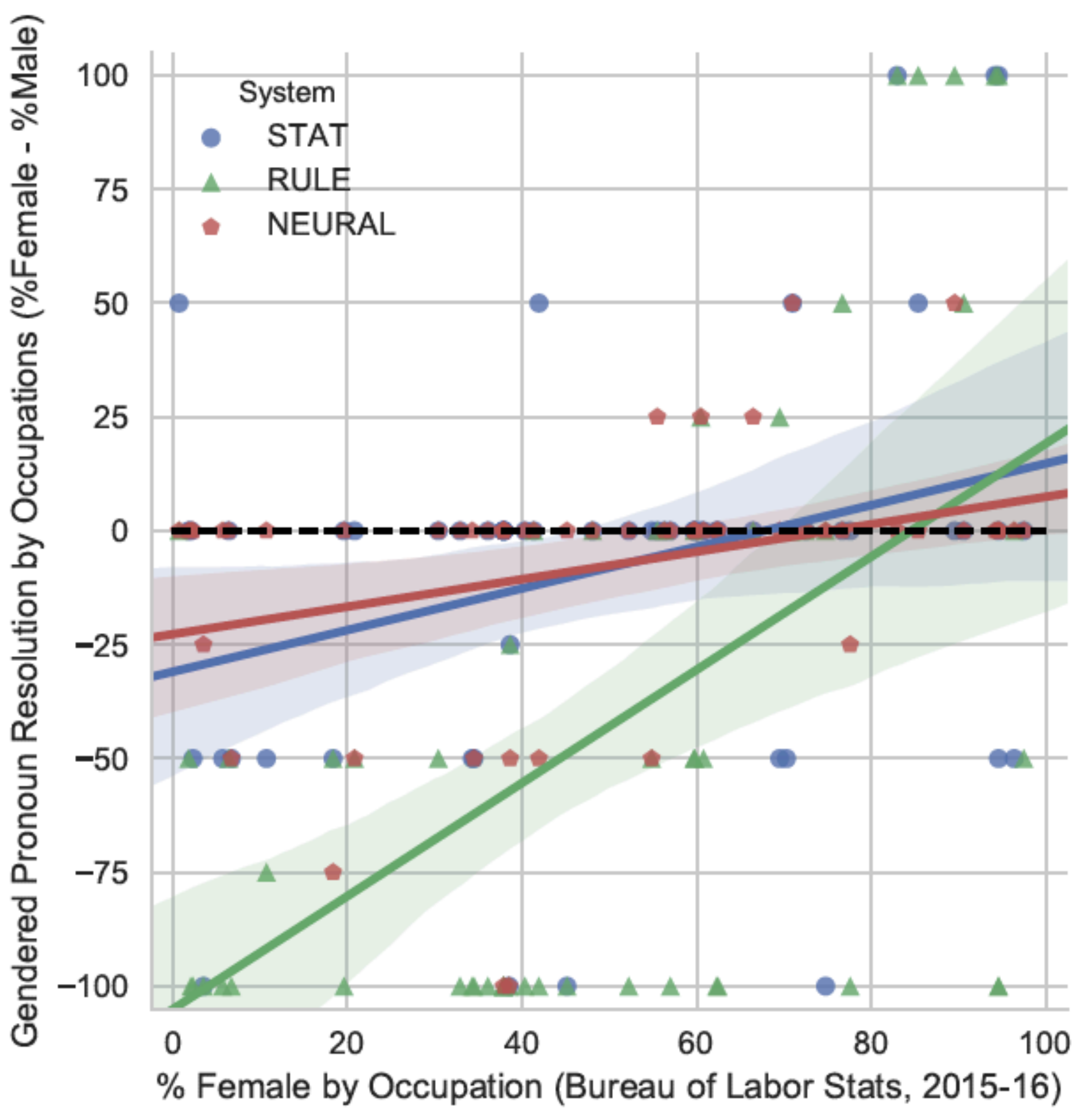
(Rudinger et al. 2018)

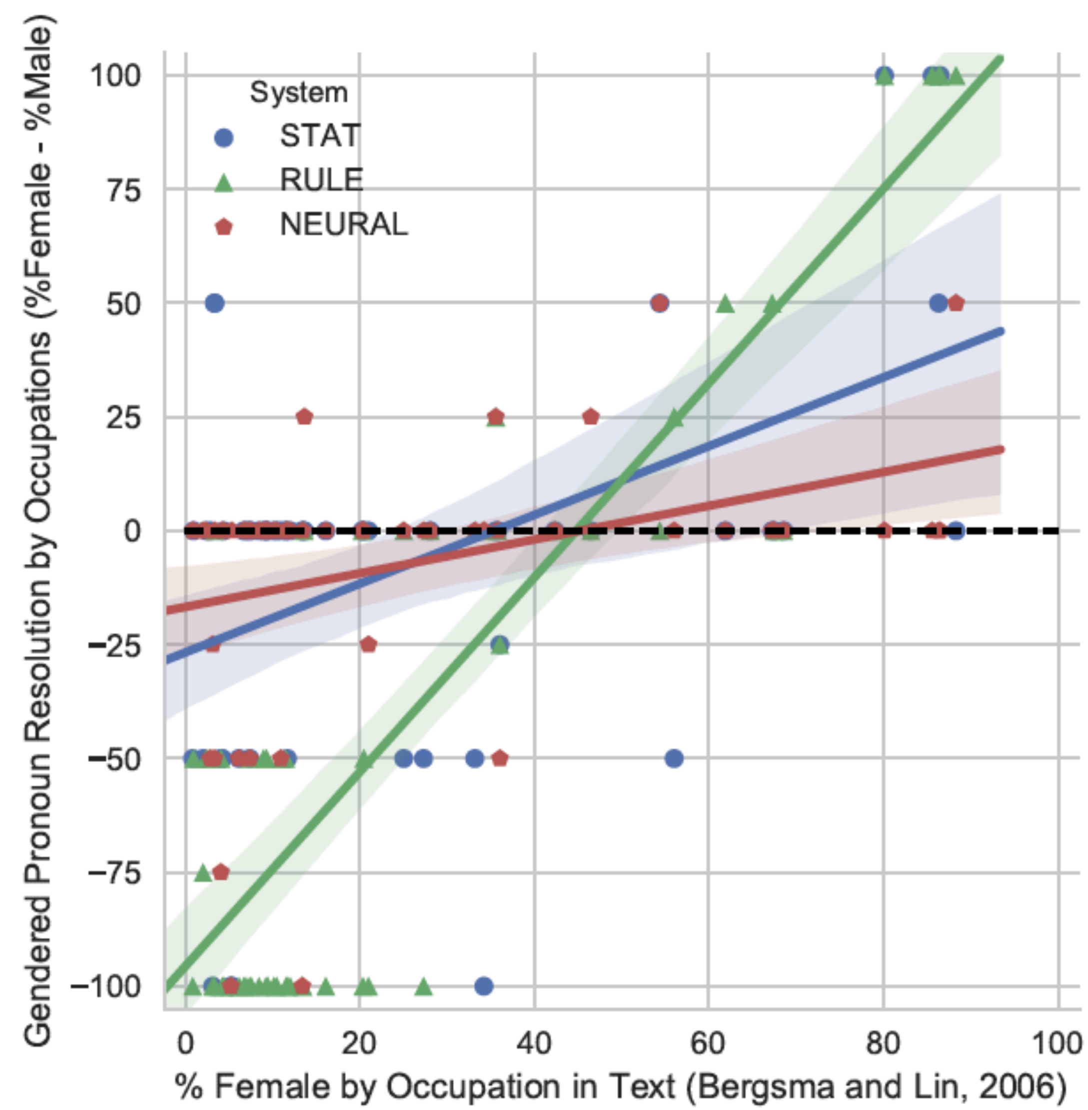
The Winogender challenge

- Tested three systems: RULE, STAT and NEURAL
- Male pronouns are more likely to be resolved as OCCUPATION than female or neutral (for all occupations); e.g. for NEURAL, 87% male vs 80% female and 36% neutral
- Neutral pronouns are often resolved as neither OCCUPATION nor PARTICIPANT
- 68% of male-female minimal pair test sentences are resolved differently by the RULE system; 28% for STAT; and 13% for NEURAL

(Rudinger et al. 2018)

The Winogender challenge





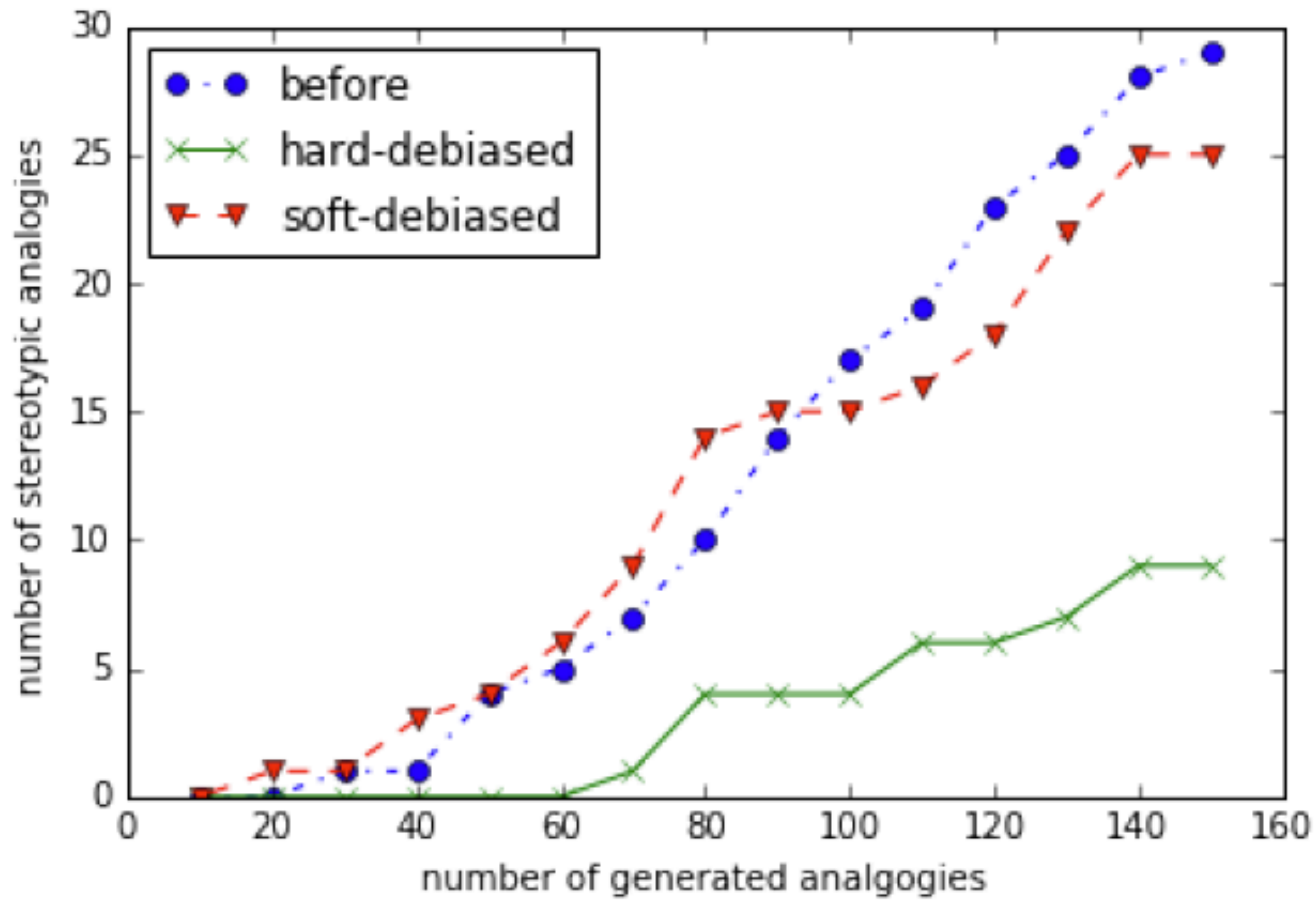
(Rudinger et al. 2018)

Gender-inclusivity

- Researchers and computer systems often implicitly adopt views of gender that are inconsistent with biological reality and the existence of non-binary and transgender individuals:
    - Binary (man/woman)
    - Immutable (assigned at birth and can’t be changed)
    - Physiological (physical appearance determines with gender)
- Automatic gender detection, or use of pronouns based on first names, can lead to misgendering and other forms of harm

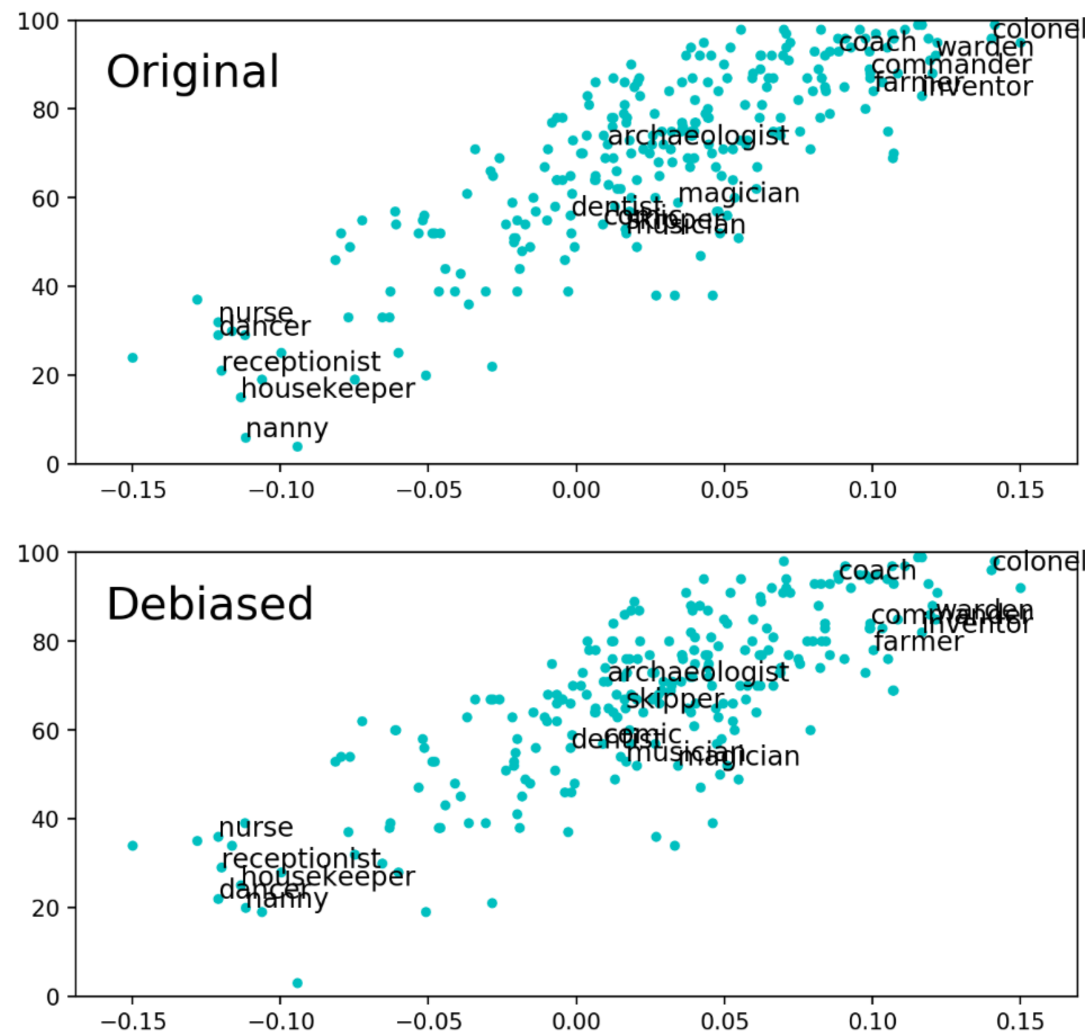
(Keyes, 2018; Cao & Daumé 2000)

Debiasing word embeddings



(Bolukbasi et al 2016)

Gender directions in vector space (e.g. *he* - *she*) are only one aspect of gender bias in word embeddings!



(b) The plots for GN-GLOVE embedding, before (top) and after (bottom) debiasing.

(Gonen & Goldberg, 2018)