Embedding Bias

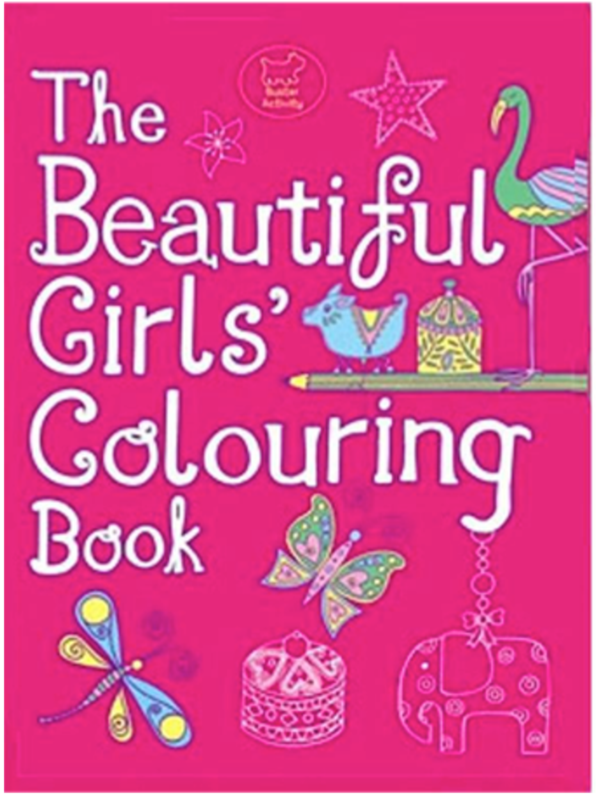
Man is to computer programmer as woman is to x

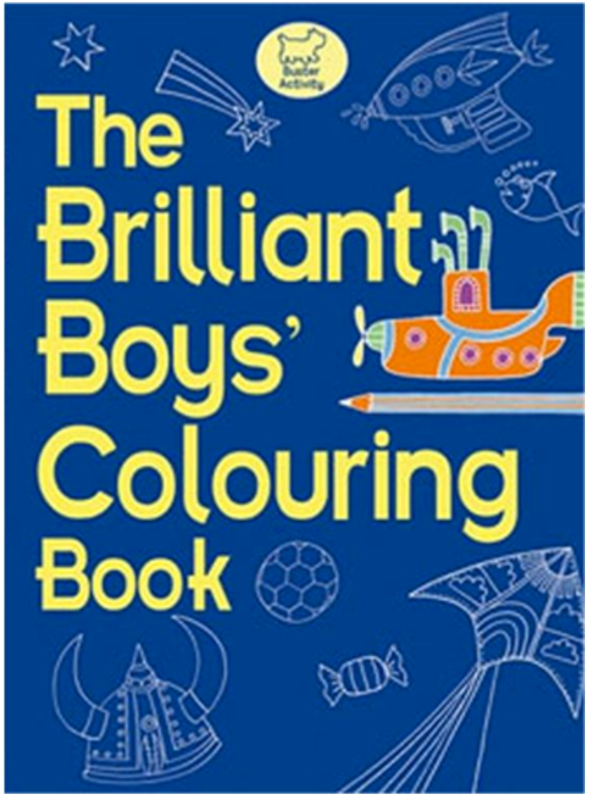
Ellie Lasater-Guttmann
Lecture 13

Which would you rather be?

Brilliant or Beautiful

What does bias look like?





How would you like to spend your day?

- Building Things
- Fixing things
- Meal prepping
- Music making
- Camping
- Fishing
- Swimming
- Boating
- Space exploring

or

- Pretend play
- Dolls and stuffed animals
- Dollhouses
- Playhouses
- Fun with pets
- Hair and nail salon
- Fun fitness
- Reading
- Jewelry making

Thanks to FatherMag.com

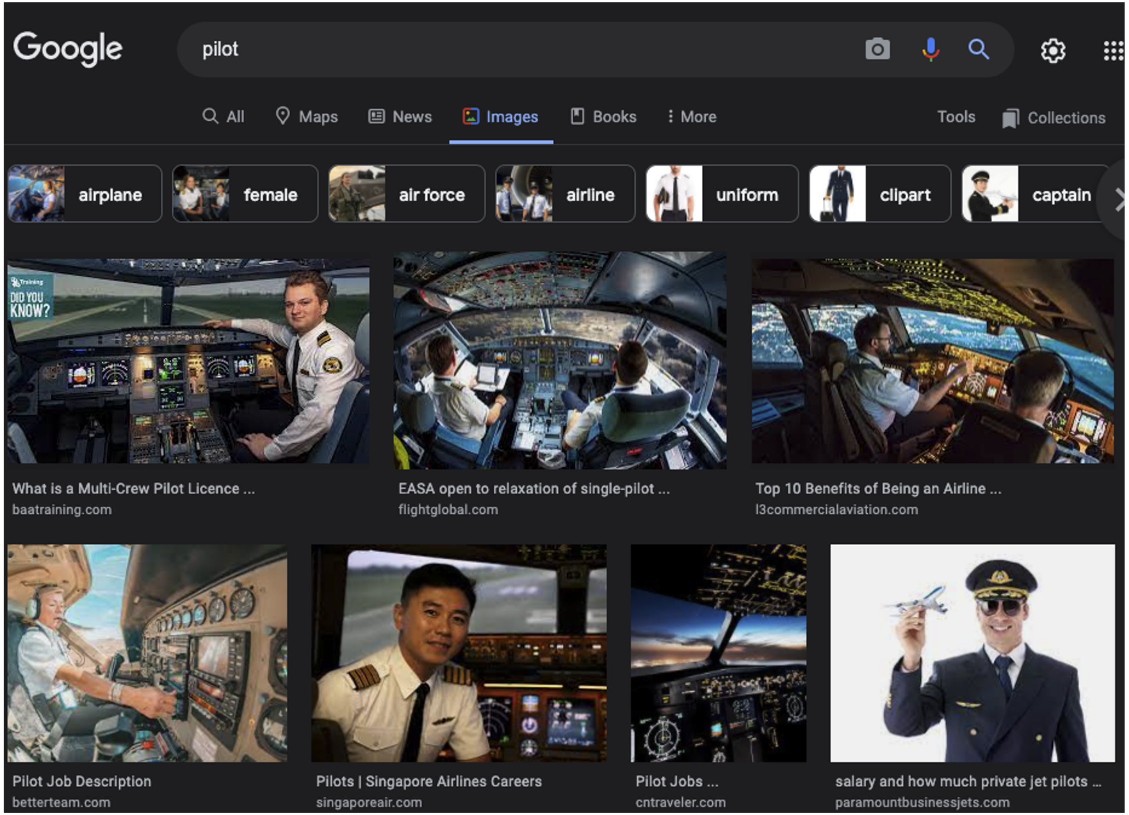
Bias - do you have more examples?

Bias - what are we talking about?

Signs of bias in the world

- Different **treatment** depending on **identity**
- Different **ideas** about someone depending on **identity**
- Different **expectations** about someone depending on **identity**
- Different **representation** depending on **identity**

Gender Bias in AI



What is a Multi-Crew Pilot Licence ...
baatraining.com

EASA open to relaxation of single-pilot ...
flightglobal.com

Top 10 Benefits of Being an Airline ...
l3commercialaviation.com

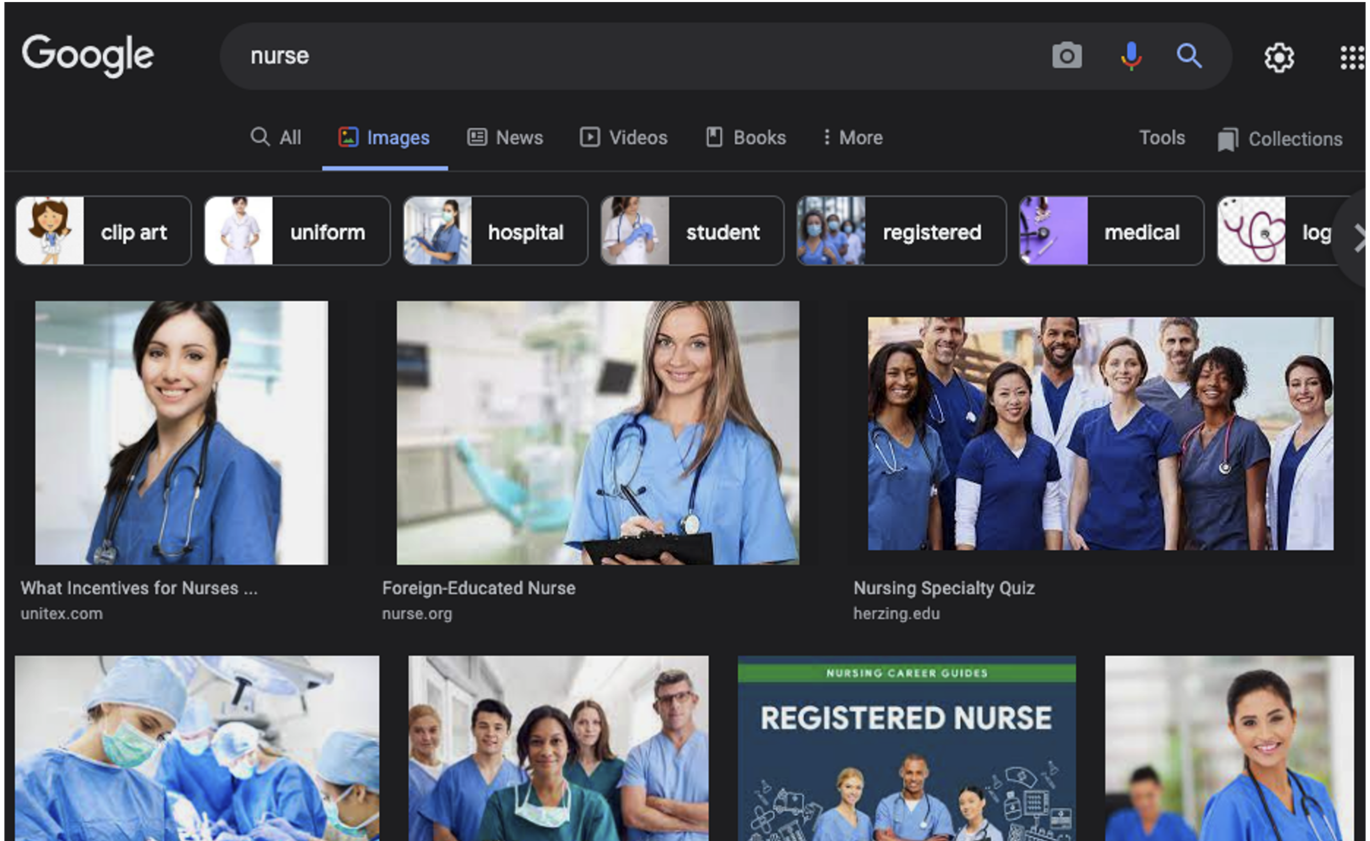
Pilot Job Description
betterteam.com

Pilots | Singapore Airlines Careers
singaporeair.com

Pilot Jobs ...
cntraveler.com

salary and how much private jet pilots ...
paramountbusinessjets.com

Gender Bias in AI



What Incentives for Nurses ...
unitex.com

Foreign-Educated Nurse
nurse.org

Nursing Specialty Quiz
herzing.edu

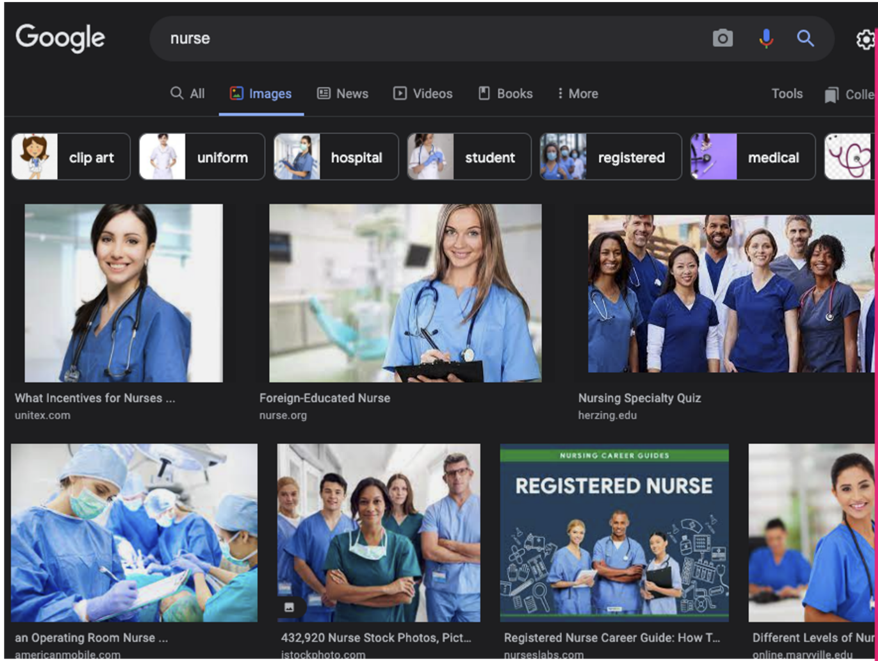
So, what’s the harm?

What is wrong with the examples we’ve seen here?

Representational Harms

A representational harm occurs when a system **reinforces** the **subordination** of some groups along **identity lines.**

What is an example of a representational harm?



What Incentives for Nurses ...
unitex.com

Foreign-Educated Nurse
nurse.org

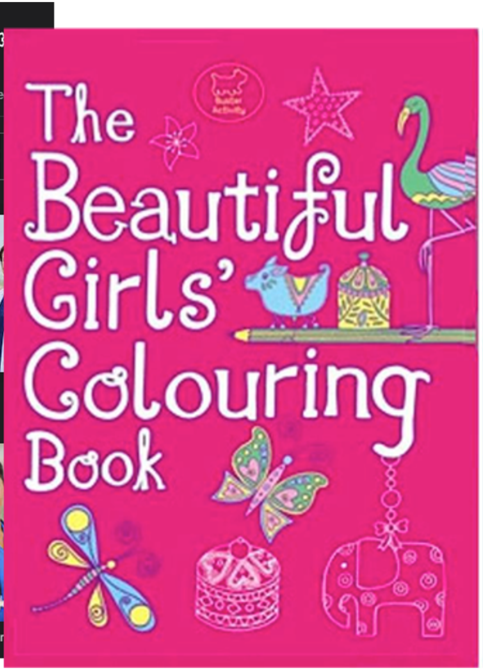
Nursing Specialty Quiz
herzing.edu

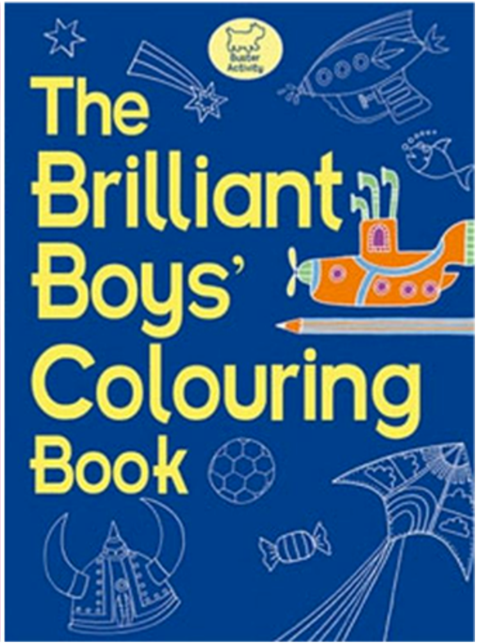
an Operating Room Nurse ...
americanmobile.com

432,920 Nurse Stock Photos, Pict...
istockphoto.com

Registered Nurse Career Guide: How T...
nurseslabs.com

Different Levels of Nur
online.maryville.edu





What is an example of a representational harm?

What about this type of bias?

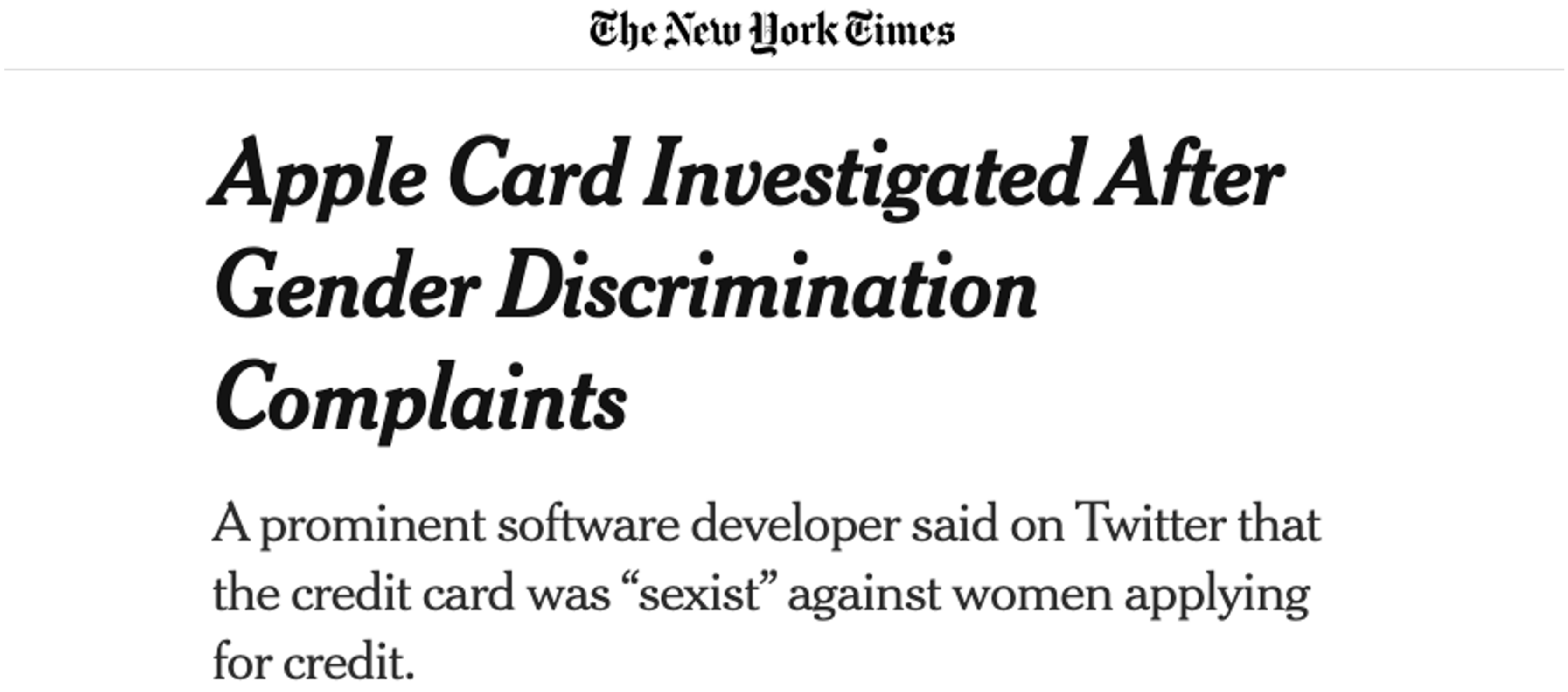
Air conditioning temperatures are set according to the resting metabolic rate of a **154-pound, 40 year-old man**. This overestimates women’s metabolic rates by 35%+

- As office temperatures get warmer, **women perform better on cognitive tasks while men perform worse**

https://www.nature.com/articles/nclimate2741

https://journals.plos.org/plosone/article/authors?id=10.1371/journal.pone.0216362

What about this type of bias?



The New York Times

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was “sexist” against women applying for credit.

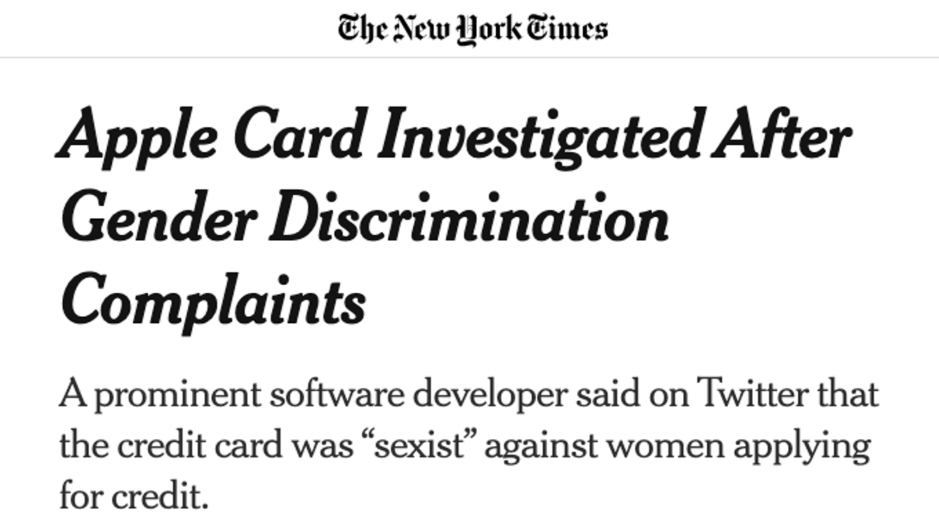
So, what’s the harm?

What is wrong with the examples we’ve seen here?

Allocative Harms

An allocative harm is when a system **allocates or withholds** certain **identity groups** an **opportunity** or a **resource**.

What is an example of an allocative harm?



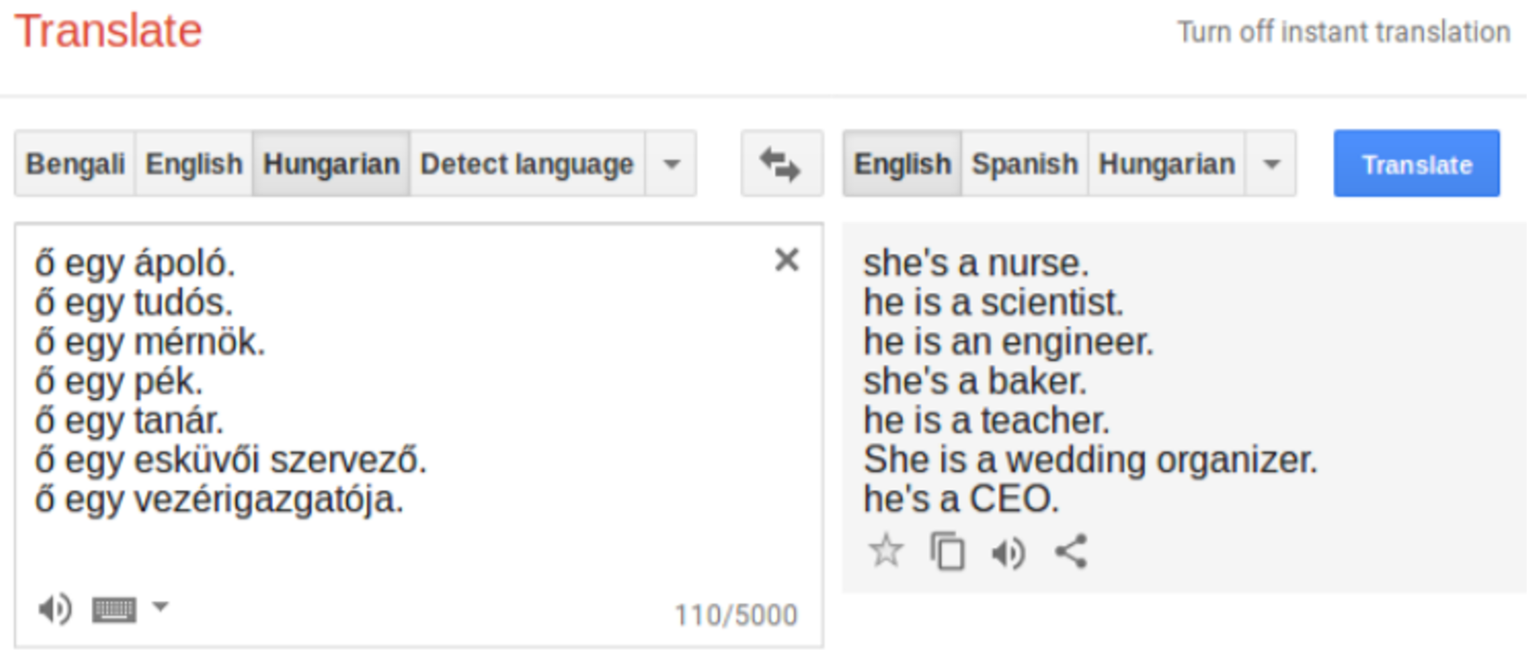
The New York Times

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was “sexist” against women applying for credit.

What is an example of an allocative harm?

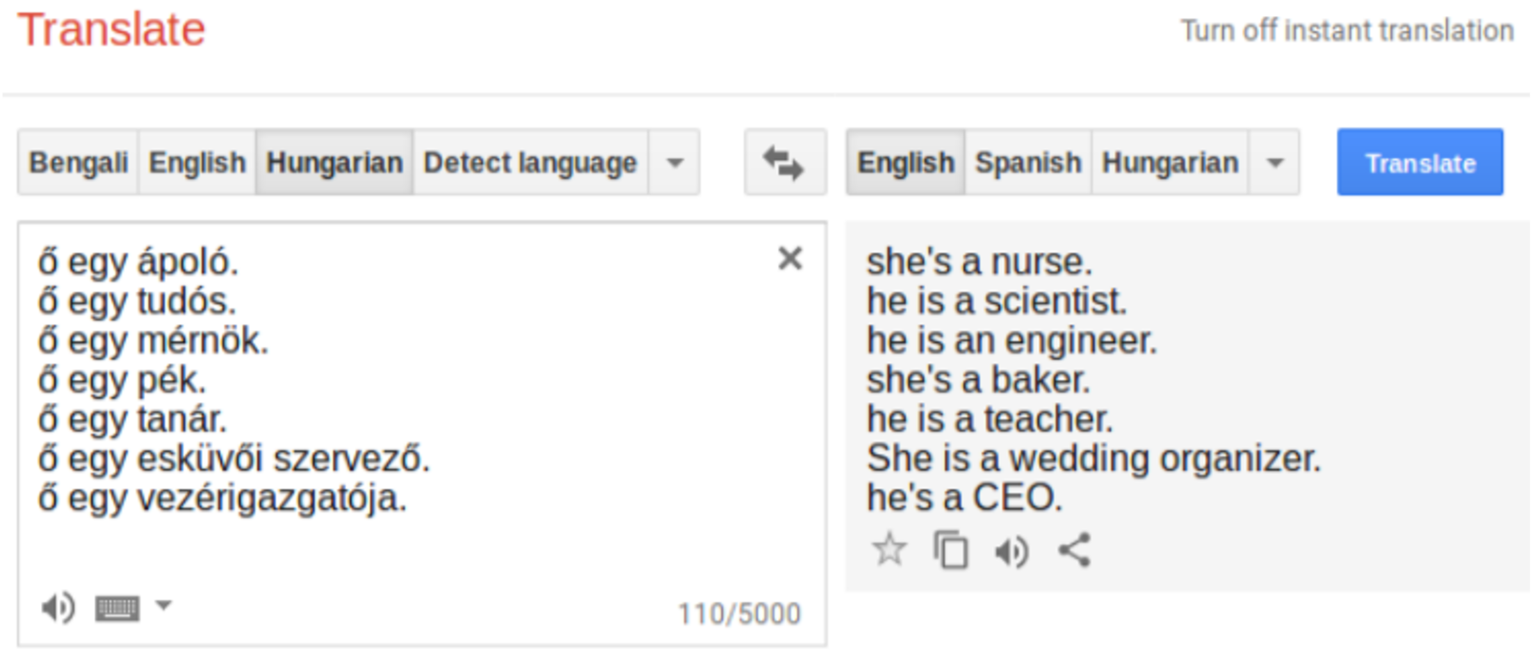
Gender Bias in NLP



ő egy ápoló.
ő egy tudós.
ő egy mérnök.
ő egy pék.
ő egy tanár.
ő egy esküvői szervező.
ő egy vezérigazgatója.

she's a nurse.
he is a scientist.
he is an engineer.
she's a baker.
he is a teacher.
She is a wedding organizer.
he's a CEO.

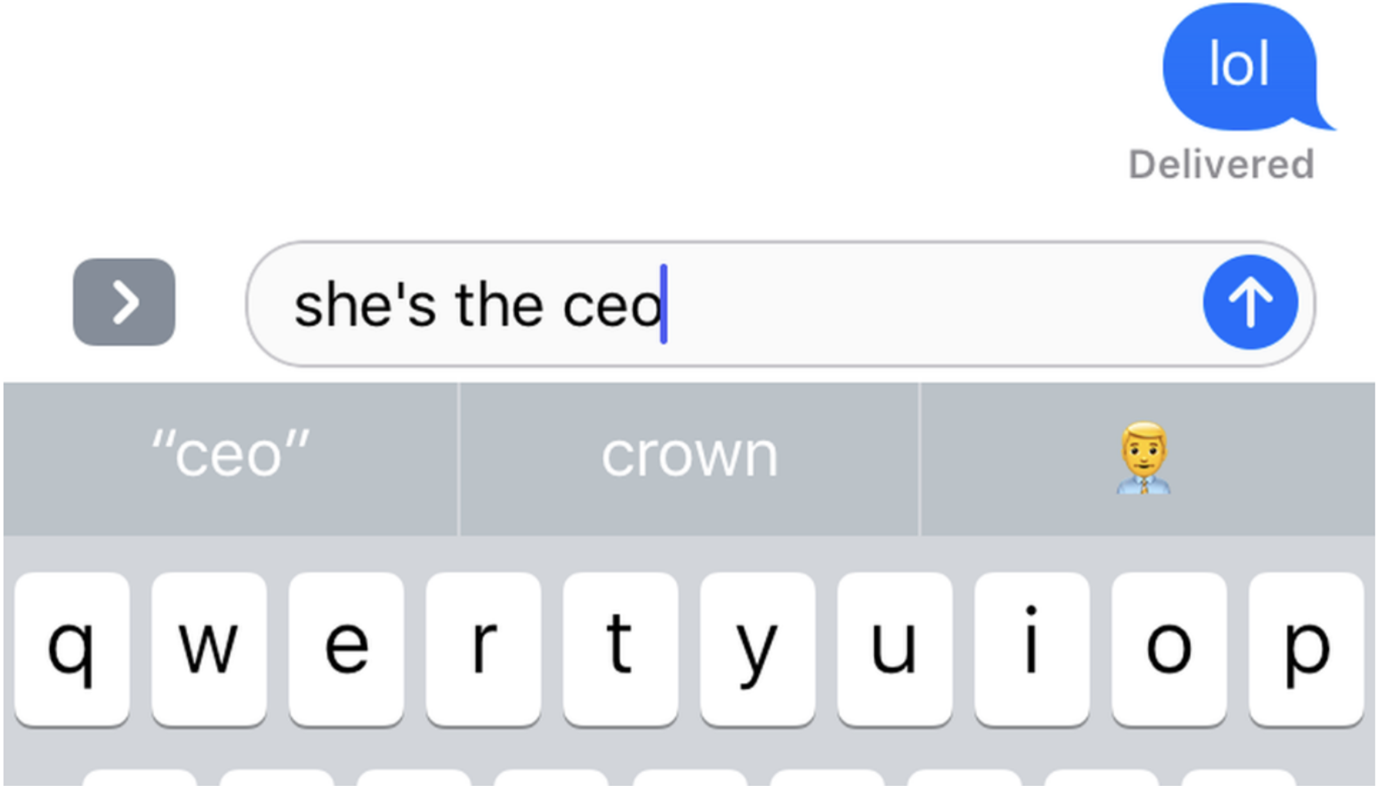
What’s the harm with this example?



ő egy ápoló.
ő egy tudós.
ő egy mérnök.
ő egy pék.
ő egy tanár.
ő egy esküvői szervező.
ő egy vezérigazgatója.

she's a nurse.
he is a scientist.
he is an engineer.
she's a baker.
he is a teacher.
She is a wedding organizer.
he's a CEO.

Gender Bias in AI

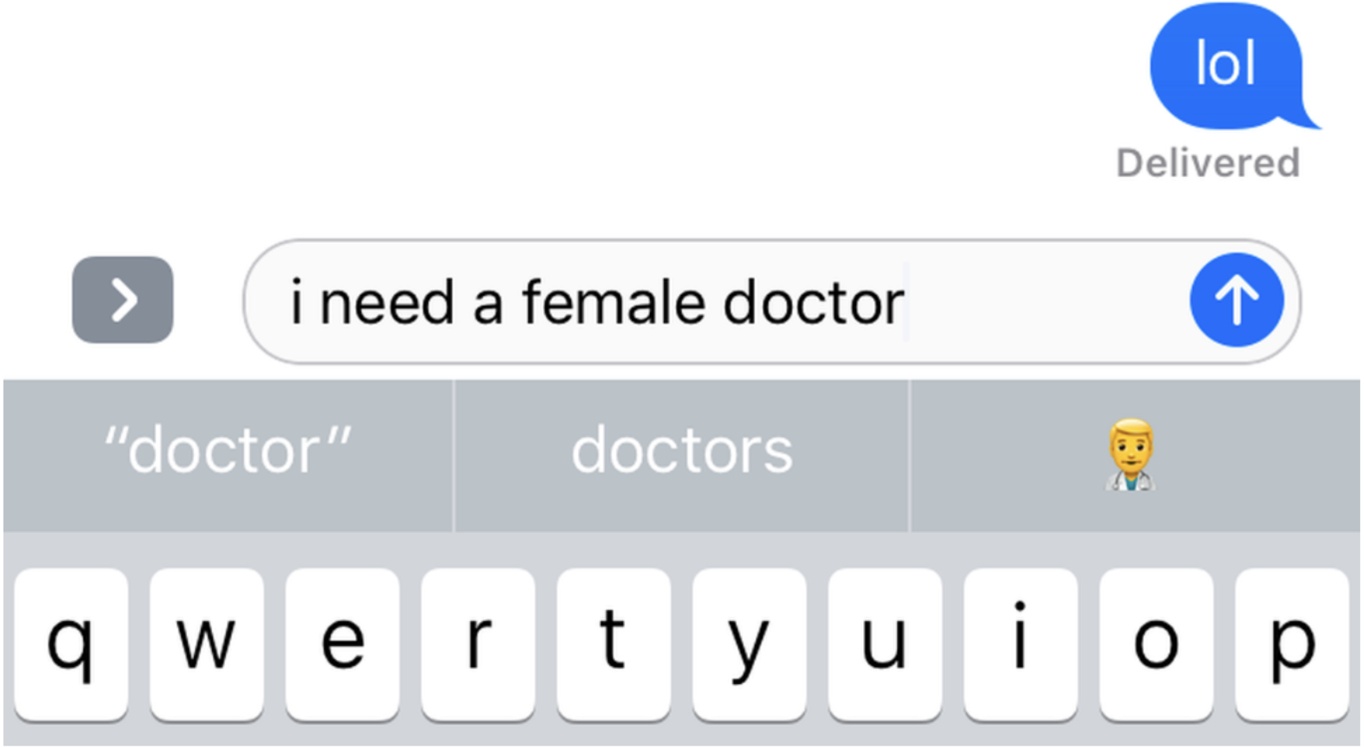


lol

Delivered

she's the ceo

What’s the harm?



lol

Delivered

i need a female doctor

Gender Bias in Word Embeddings

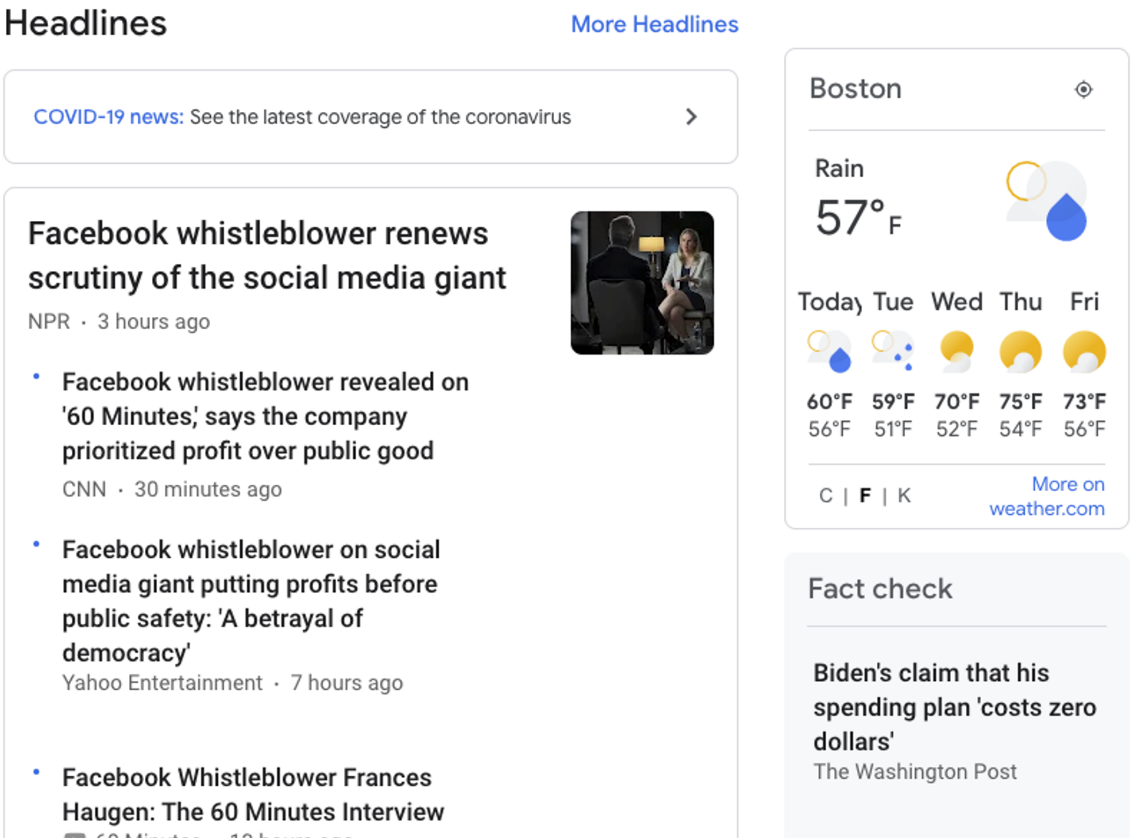
Can anyone describe the embedding model discussed in the reading?

Word2Vec: word-embedding model, learning associations between words in the corpus

Corpus in question: **Google News**

w2vNEWS: Word2Vec trained on Google News

Google News



Headlines

More Headlines

COVID-19 news: See the latest coverage of the coronavirus

Facebook whistleblower renews scrutiny of the social media giant

NPR · 3 hours ago

- Facebook whistleblower revealed on '60 Minutes,' says the company prioritized profit over public good
  CNN · 30 minutes ago
- Facebook whistleblower on social media giant putting profits before public safety: 'A betrayal of democracy'
  Yahoo Entertainment · 7 hours ago
- Facebook Whistleblower Frances Haugen: The 60 Minutes Interview
  60 Minutes · 10 hours ago

Boston

Rain

57°F

Today	Tue	Wed	Thu	Fri
60°F	59°F	70°F	75°F	73°F
56°F	51°F	52°F	54°F	56°F

C | **F** | K

More on weather.com

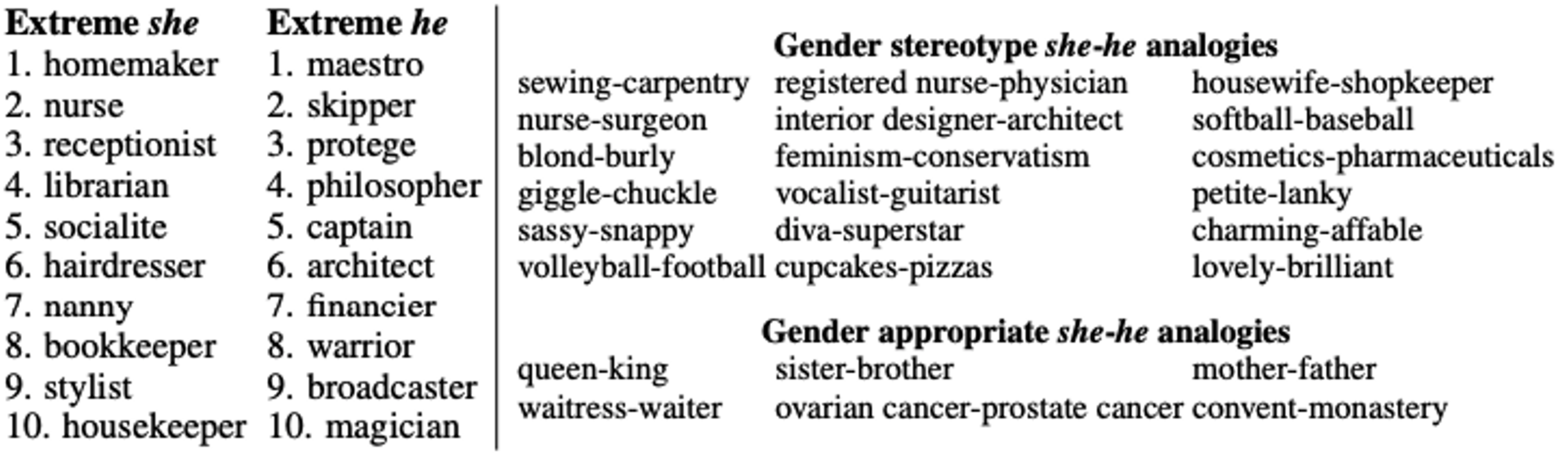
Fact check

Biden's claim that his spending plan 'costs zero dollars'

The Washington Post

What happened in w2vNEWS?

What happened in w2vNEWS?



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

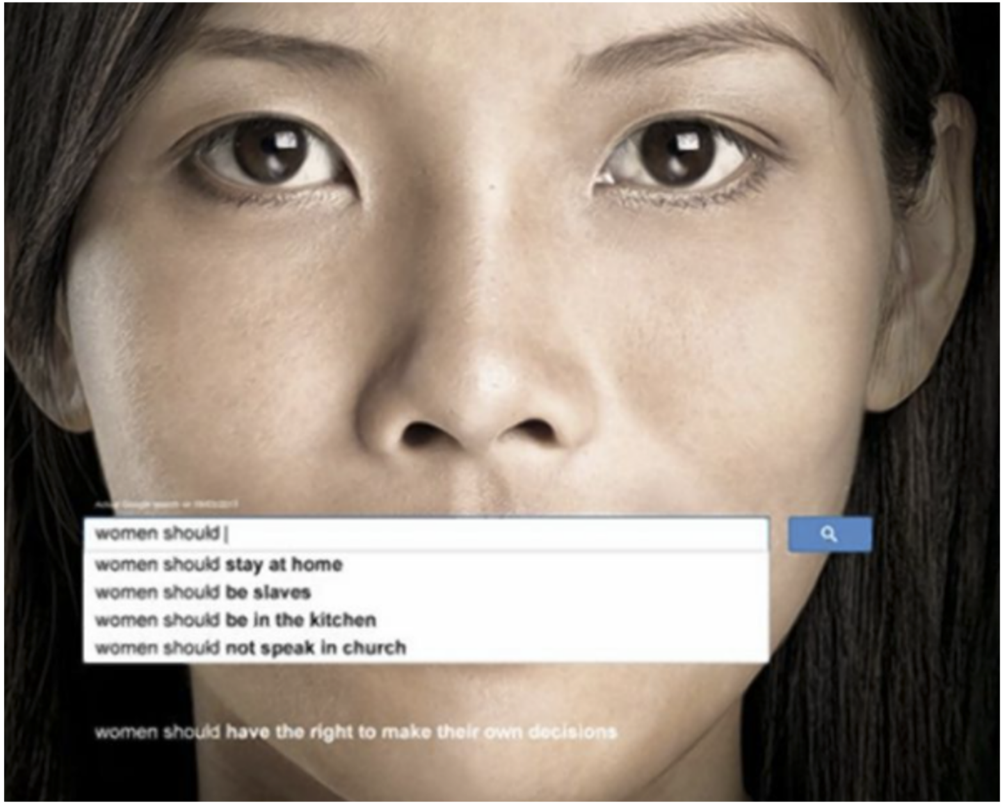
Gender appropriate *she*-*he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

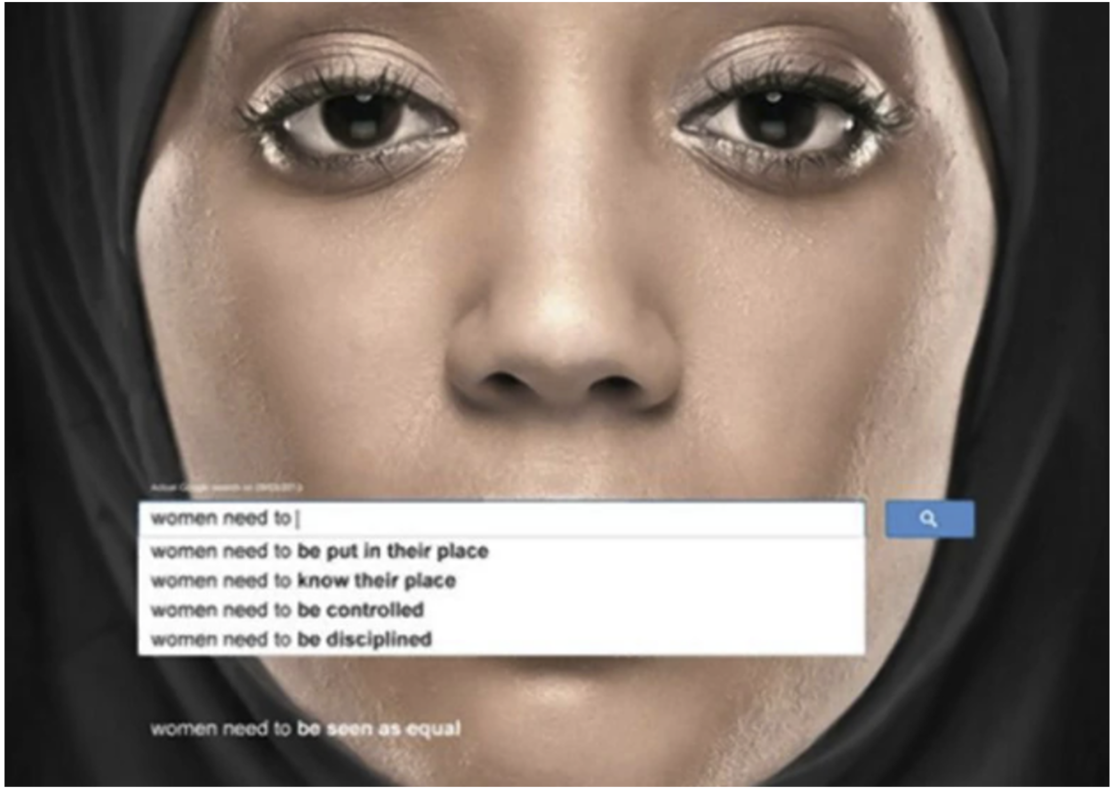
Man is to computer programmer as woman is to homemaker.

What are word embeddings used for?

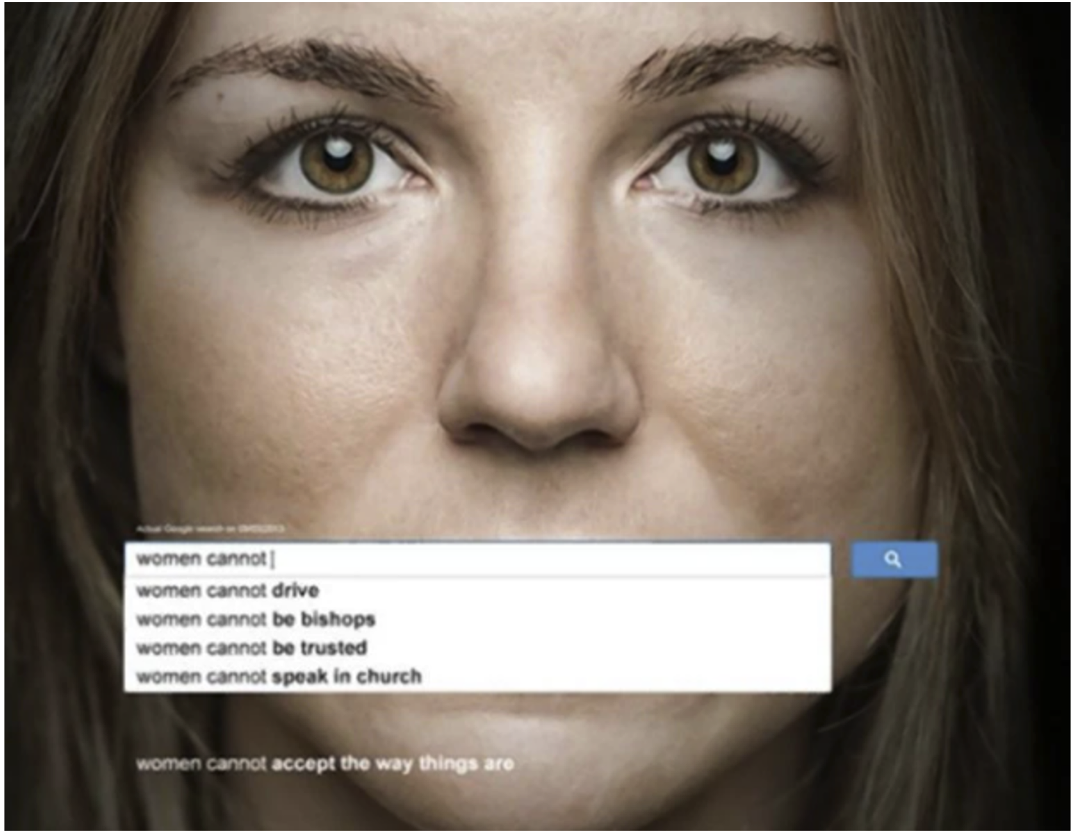
Gender Bias in Word Embeddings


Actual Google search on 09/03/2013
women should |
women should stay at home
women should be slaves
women should be in the kitchen
women should not speak in church
women should have the right to make their own decisions

Gender Bias in Word Embeddings

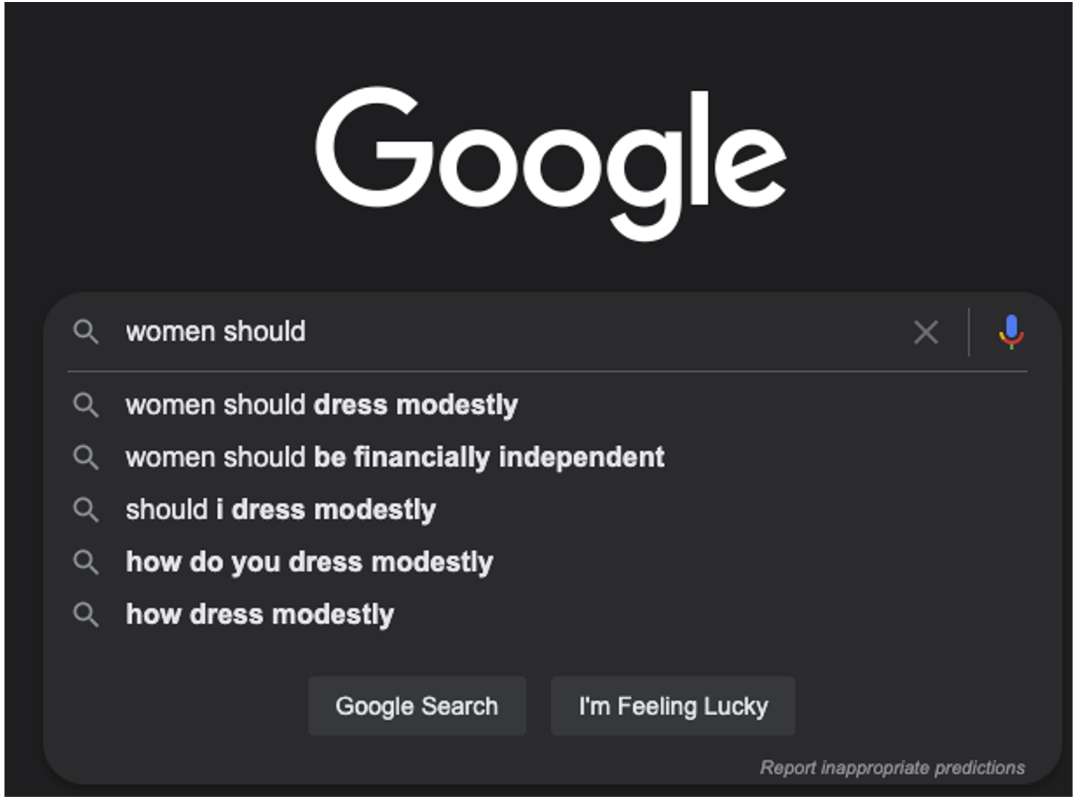

Actual Google search on
women need to |
women need to be put in their place
women need to know their place
women need to be controlled
women need to be disciplined
women need to be seen as equal

Gender Bias in Word Embeddings



Actual Google search on [illegible]
women cannot |
women cannot drive
women cannot be bishops
women cannot be trusted
women cannot speak in church
women cannot accept the way things are

Gender Bias in Word Embeddings



women should

women should **dress modestly**

women should **be financially independent**

should **i dress modestly**

how do you dress modestly

how dress modestly

So… what’s the harm?

Break yourselves into groups of 3 to brainstorm what types of harm are caused by word embedding models being biased in this way.

You’ll chat for 5 minutes and then we’ll debrief for 5 minutes.

BIG TAKEAWAY

NLP plays a role in representing people and allocating resources to them.

Should we do anything about our biased word embeddings?

Should we do anything about our biased word embeddings?

A possible answer: **No, it reflects the bias in the corpus.**

Should we do anything about our biased word embeddings?

Another possible answer: **Yes, because applications exacerbate the problems in the corpus**

BIG TAKEAWAYS

If we do nothing, our models will perpetuate representational and allocative harms.

How to debias…?

Organize yourself into groups of 3 to answer the following questions, using our reading(s) as a springboard. You'll be reporting back!:

1. What would a **unbiased** word embedding model look like?
2. Could an unbiased word embedding model cause **representational or allocative harms**?
3. What **steps would you take** to design an unbiased model?

What would debiased embeddings look like?

One definition of a debiased model: One cannot determine the gender association of a word by looking at its projection on any gendered pair.

Eg. “Nurse” is **equidistant from** “man” and “woman”

But remember… “mother” should not be equidistant from “man” and “woman”

What would debiased embeddings look like?

But… even when “nurse” is equidistant to “man” and “woman”...

“Nurse” is close to “receptionist,” “caregiver,” and “teacher”

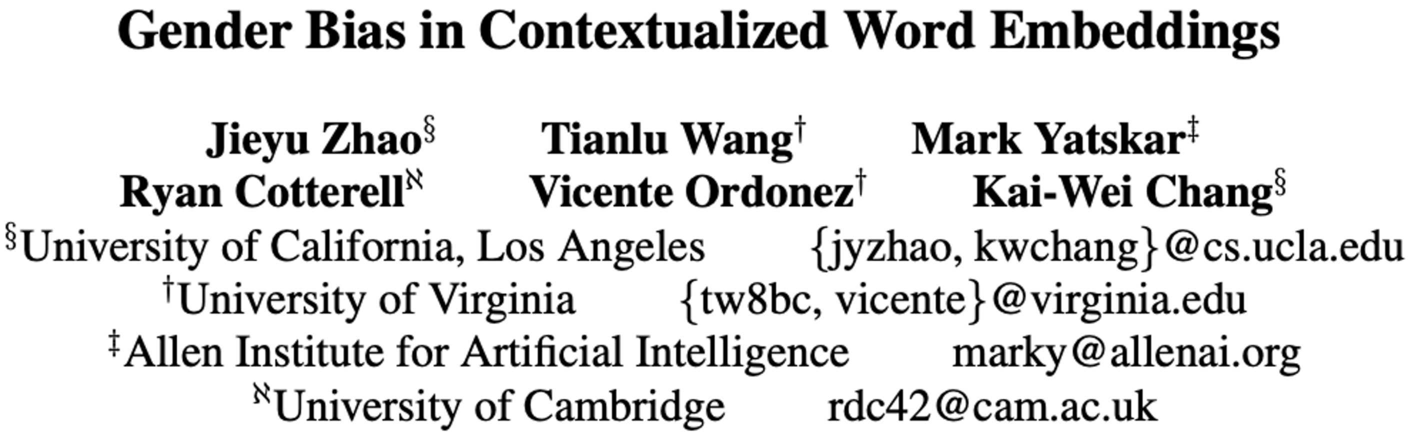
Debiasing - Strategy 1 - Post-processing

- **Reduce the bias for all words that are not inherently gendered (eg. “mother”)**
- They do that by zeroing the gender projection of each word on a predefined gender direction. (eg. similarity to the woman-man direction)
- This is an attempt at debiasing at the **post-processing stage**

Debiasing - Strategy 2 - Pre-processing

- Encourage gender to be represented in the final coordinate of each vector, so it can be excluded explicitly
- This is an attempt at debiasing at the **training stage**

Contextualized Embeddings



Gender Bias in Contextualized Word Embeddings

Jieyu Zhao[§] **Tianlu Wang**[†] **Mark Yatskar**[‡]
Ryan Cotterell[ℵ] **Vicente Ordonez**[†] **Kai-Wei Chang**[§]
[§]University of California, Los Angeles {jyzhao, kwchang}@cs.ucla.edu
[†]University of Virginia {tw8bc, vicente}@virginia.edu
[‡]Allen Institute for Artificial Intelligence marky@allenai.org
[ℵ]University of Cambridge rdc42@cam.ac.uk

BIG TAKEAWAYS

NLP plays a role in representing people and allocating resources to them. If we do nothing, our models will perpetuate representational and allocative harms.

But debiasing is hard!

Please take this survey!

https://tinyurl.com/AC295F21