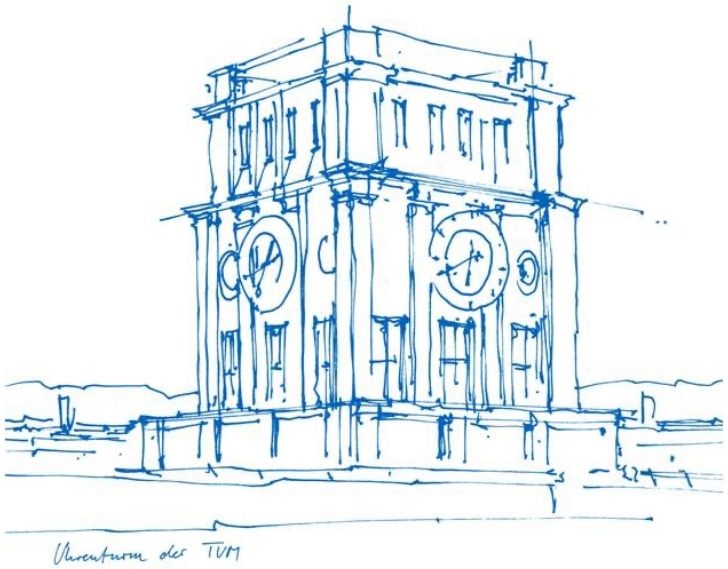
Computational social science for the study of marginalized groups



Stefanie Walter

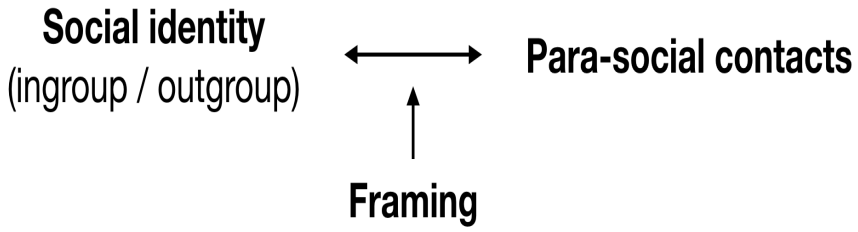
Technical University of Munich

Why study media representations of marginalized groups

Social identity theory (Tajfel and Turner 1979)

Para-social contacts hypothesis (Schiappa et al. 2005)

Framing (Entman 1993)



Who is represented and why it matters

Media as:

- Reflection of social realities (e.g., Brooks & Hébert, 2006)
- Brokers between citizens (cf., Ramasubramanian 2013: 54)
- Source of information about groups or issues related to minorities and diversity (Bleich et al. 2015)
- Space for the participation in public (Bleich et al. 2015)

Why paying attention to marginalized groups in academic research matters

"Being included (…) is an act or registration. Seeing yourself reflected in data collection exercises can positively shape how you perceive your own identity. (…) if you appear (…) it is hard to claim you do not exist." (Guyan, 2022: 6)

Why it matters how marginalized groups are represented

Media coverage affects:

- Perceptions of and attitudes towards groups (e.g., Bos et al. 2016)
- Stereotypes (Schemer, 2014)
- Behaviours (e.g., Koopmans, 1996) and vote choices (e.g., Burscher et al. 2015)

- Ingroup identities (e.g., Zerback & Karadas, 2022)
- Role models (Young et al. 2013)

Computational Social Science (CSS)

New opprotunities for the study or marginalized groups

- "Measurement revolution" (cf. Jungherr, 2018)

Data-driven objectivity:

- Increased objectivity and accuracy (Mayer-Schönberger, 2013)

What do we mean by bias?

Prejudice for or against a person, group, idea, or thing

Can be expressed directly or more subtly

In NLP these may refer to:

- Allocational harms
- **Representational harms**

(Blodgett et al. 2020)

What do we mean by bias?

Prejudice for or against a person, group, idea, or thing

Can be expressed directly or more subtly

In NLP these may refer to:

- Allocational harms
- **Representational harms**

(Blodgett et al. 2020)

Bias in Large Language Models

"The presence of systematic misrepresentations, attribution errors, or factual distortions that result in favoring certain groups or ideas, perpetuating stereotypes, or making incorrect assumptions based on learned patterns"

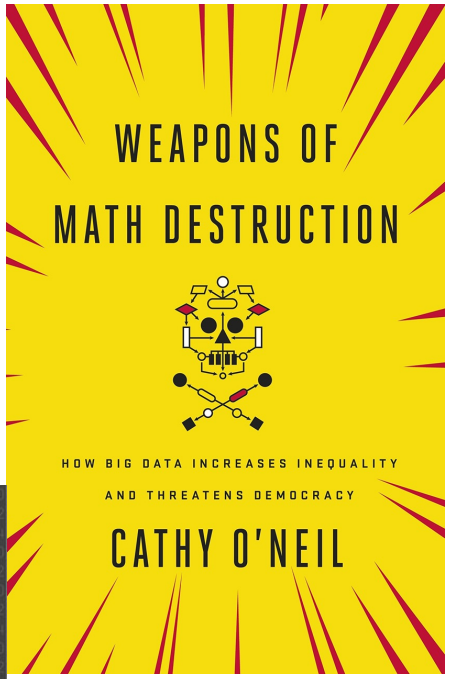
(Ferrara, 2023: 2)

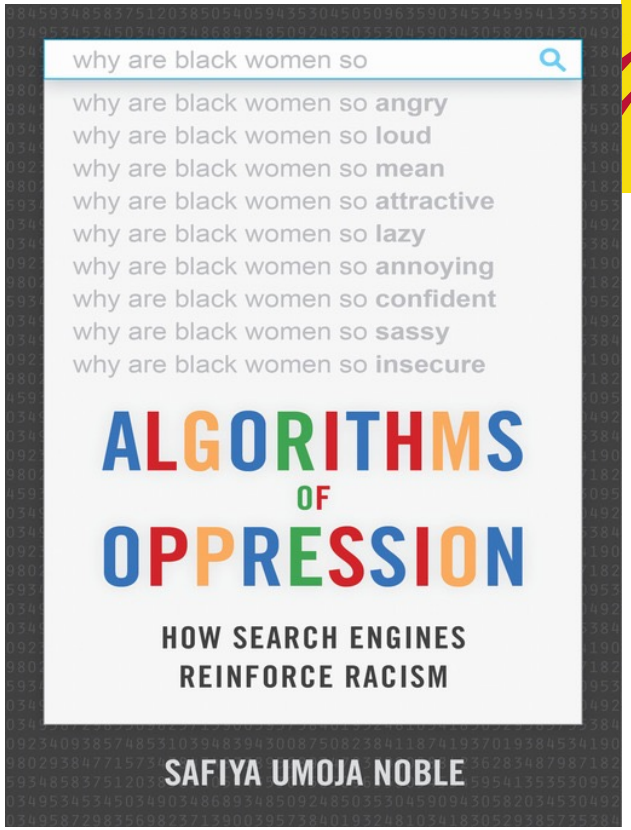
What are the challenges?

Biased representations

Danger of AI & big data in "masking and deepening social inequality" (Noble, 2018, p.1)

Algorithmically driven data failures specific to marginalized groups



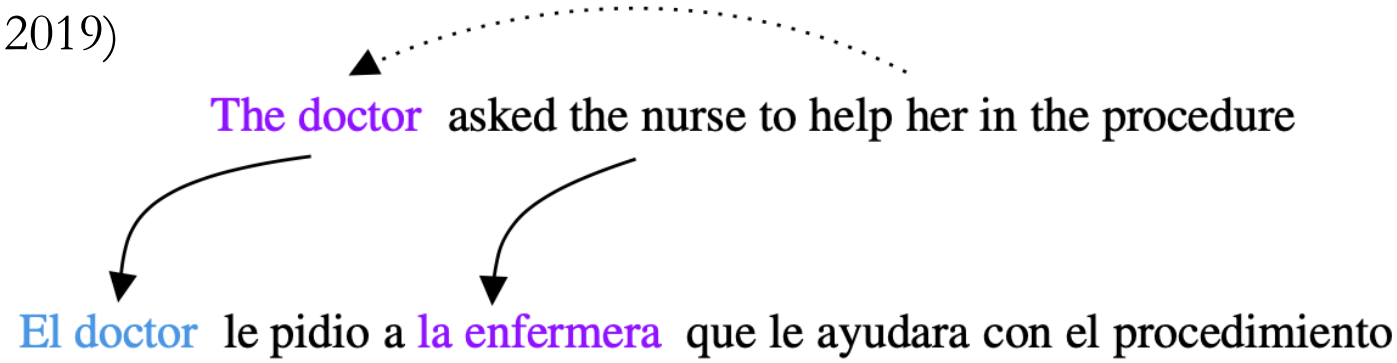


Machine translation

Demographic factors (e.g., gender, age) shaping language use

But: Likely to get lost in translation (Vanmassenhove et al. 2019)

Machine translation models prone to gender-biased translation errors (Stanovsky, 2019)



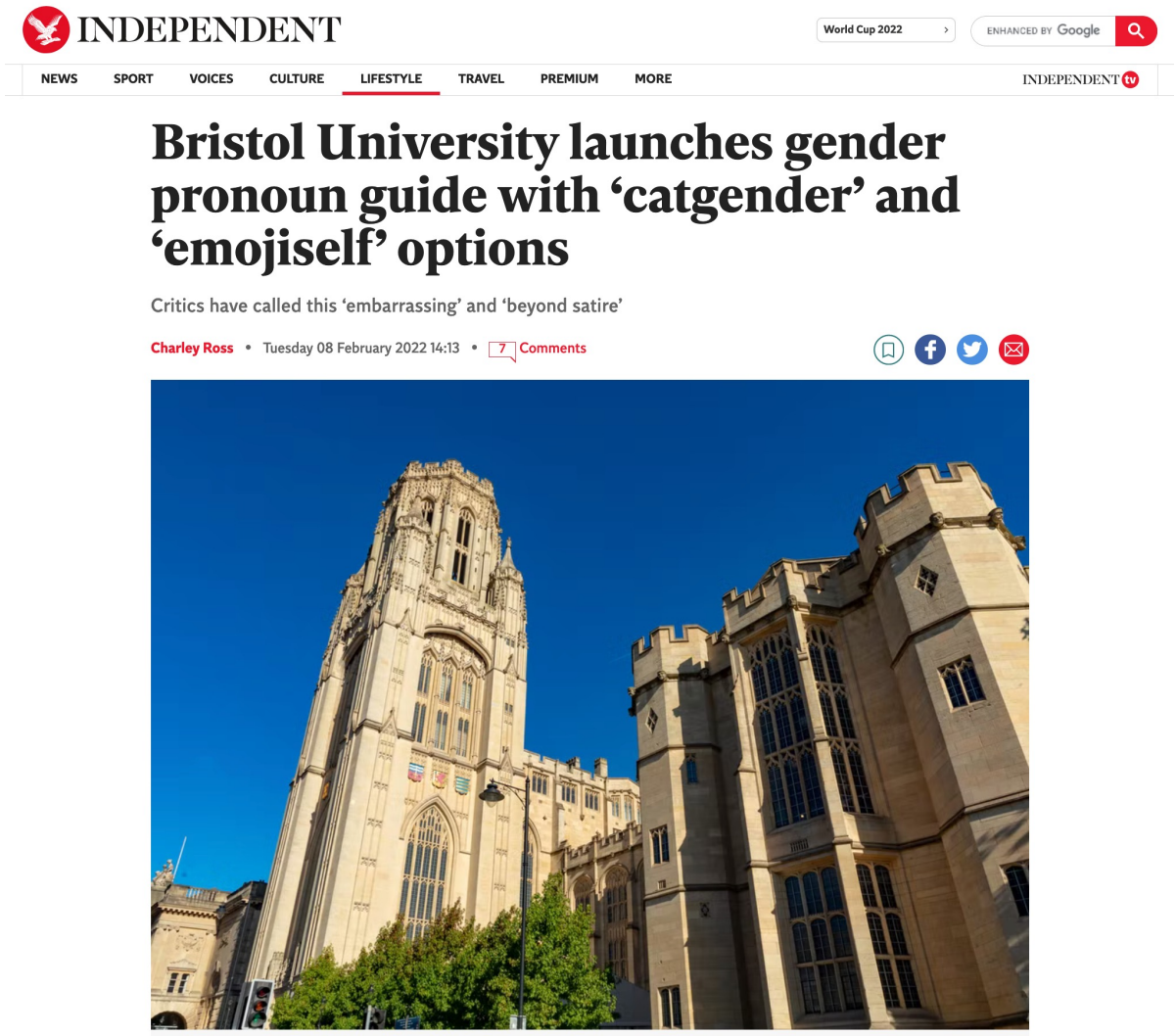
Machine translation

Identity-expression beyond gender ignored by NLP

- Gender-neutral pronouns
- Neopronouns
- Alternating pronouns
- No pronouns

(Lauscher et al. 2022)

Nom.	Acc.	Poss. (dep.)	Poss. (indep.)	Reflexive
Gendered Pronouns				
he	him	his	his	himself
she	her	her	hers	herself

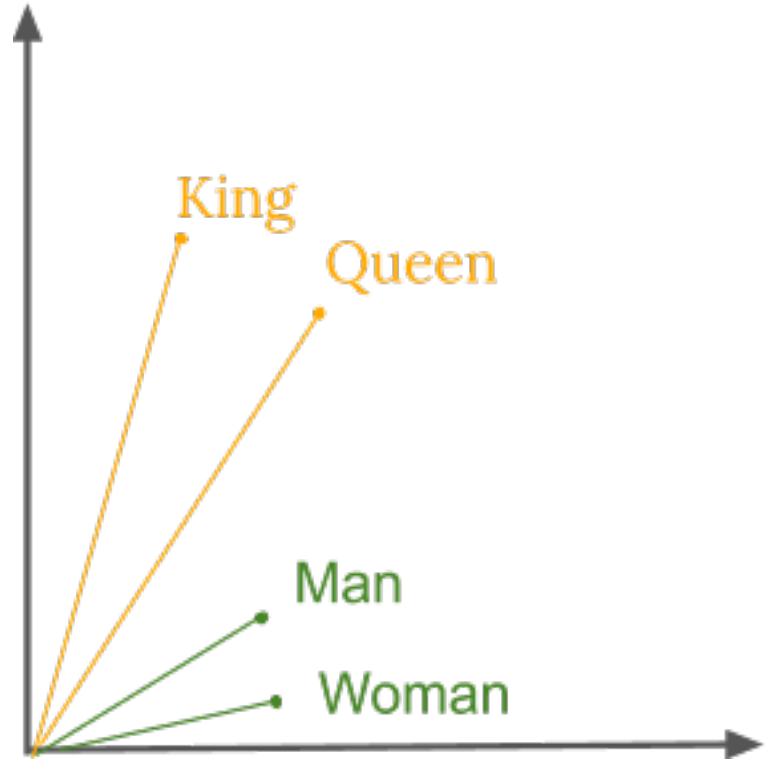


Nameself Pronouns				
John	John	Johns	Johns	Johnself
...				

Word embeddings

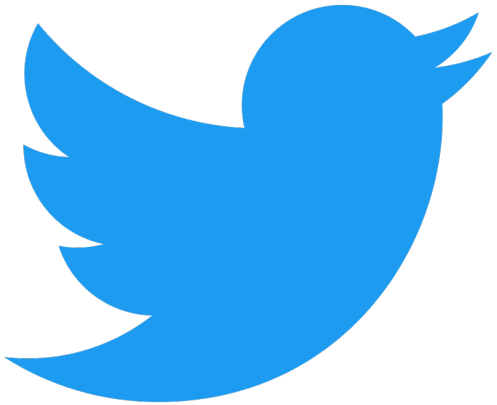
Powerful tool to extract associations captured in texts

- Racial bias (e.g., Caliskan et al., 2017)
- Gender bias (e.g. Kiritchenko and Mohammad, 2018)



Language modeling

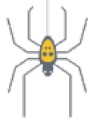
Training data



Large amounts of freely available data sources

How are minority groups represented in the data?

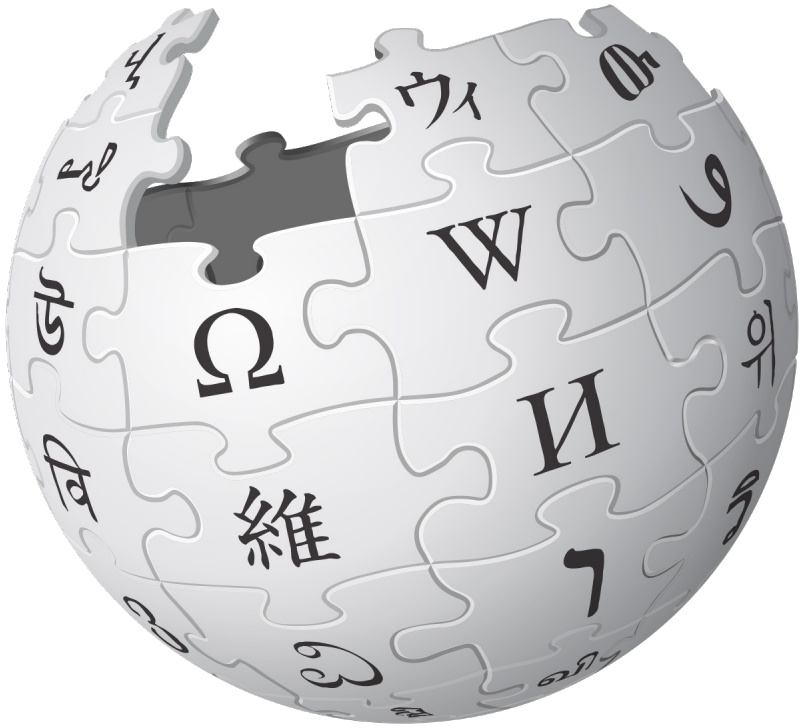
Common Crawl



Wikipedia

'Neutral point of view' policy:

"Representing fairly, proportionately, and, as far as possible, without editorial bias"



But:

Geographical bias (Beytía, 2020)

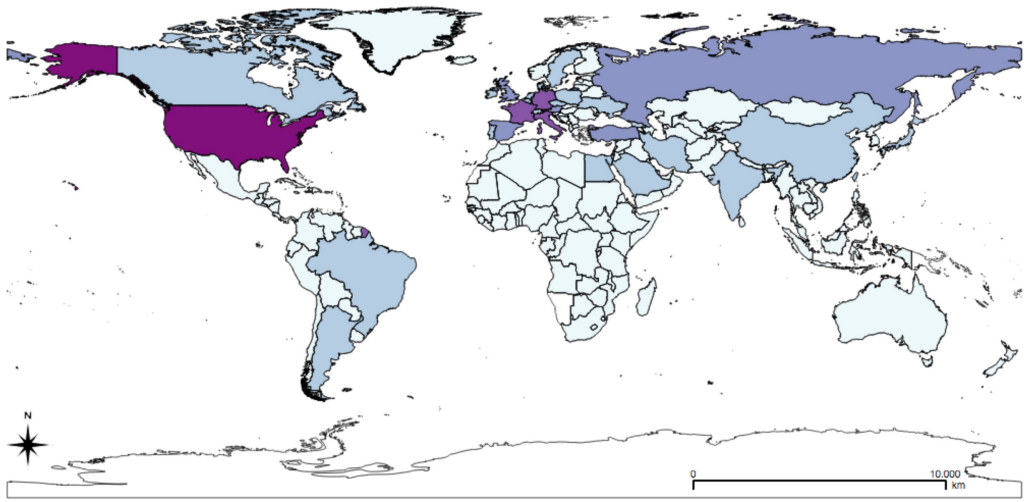
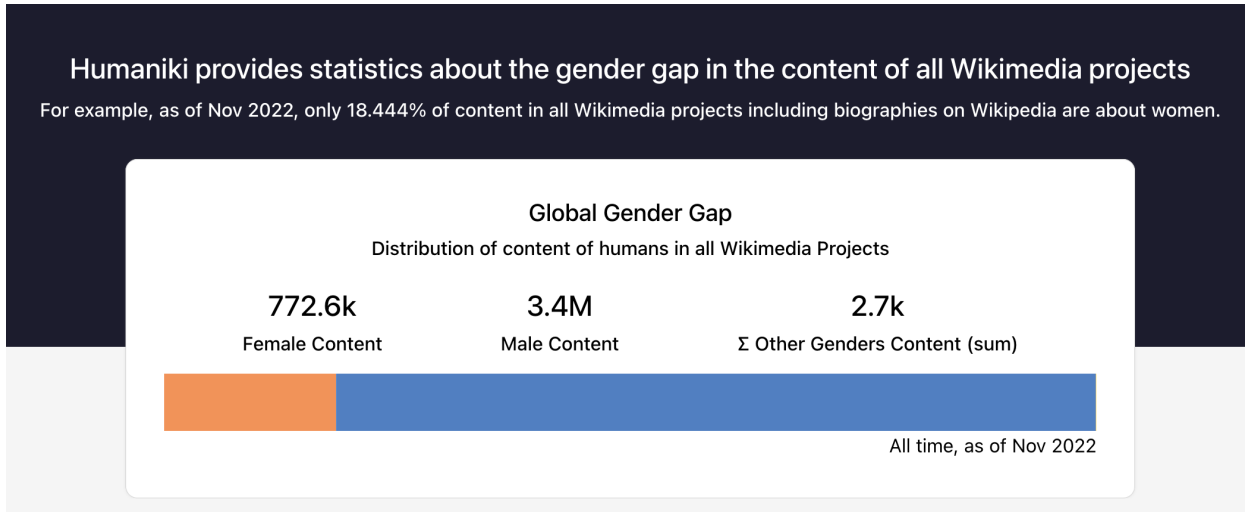


Figure 3: Spatial concentration of the biographical coverage in Wikipedia considering the internal positioning of the articles (BCI accumulation)

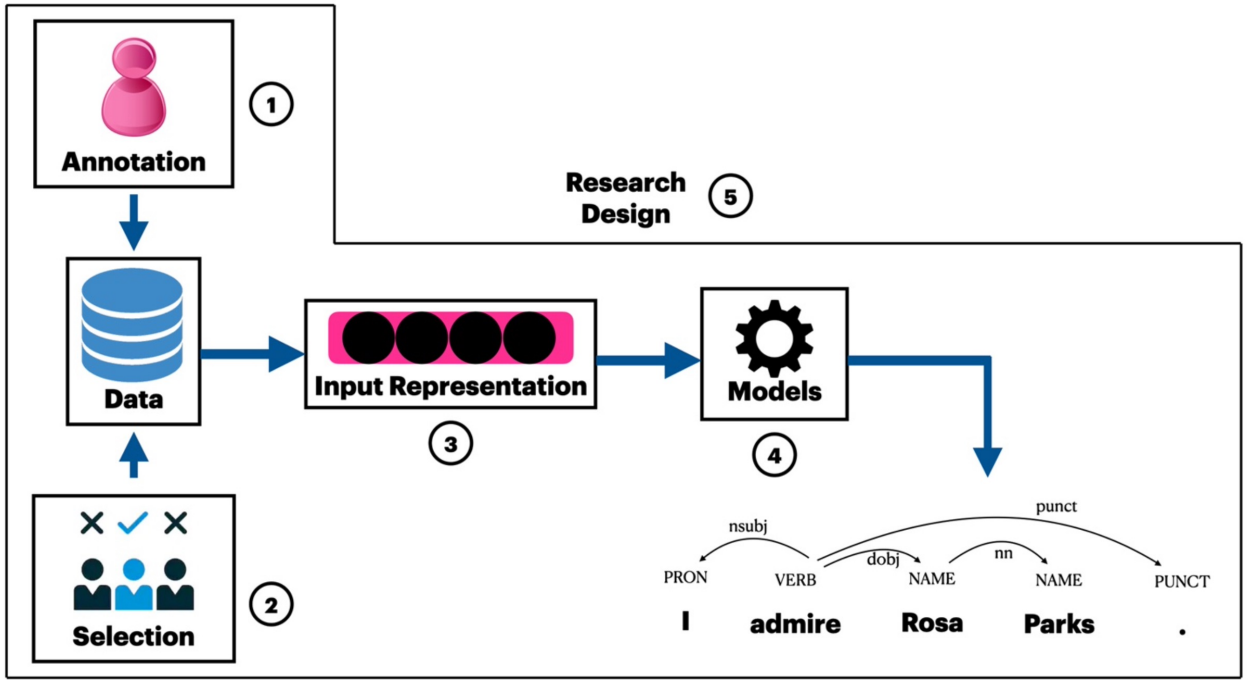
Gender bias in text and visual content (Beytía et al. 2022)



(https://whgi.wmflabs.org/gender-by-dob)

5 sources of bias in NLP

1. Selection bias (input data)
2. Label bias (annotations)
3. Bias from input representations (semantic bias)
4. Bias overamplification (models)
5. Research design



(Hovy & Prabhumoye, 2021)

ChatGPT

ST The doctor yelled at the nurse because she was late. Who was late?

The nurse was late.

The Doctor And The Son

ST The nurse yelled at the doctor because she was late. Who was late?

A father and son were in a car accident where the father was killed. The ambulance brought the son to the hospital. He needed immediate surgery. In the operating room, a doctor came in and looked at the little boy and said I can't operate on him he is my son. Who is the doctor?

In this scenario, the nurse was late.

ST The doctor yelled at the nurse because he was late. Who was late?

In this case, the doctor was late.

ST The nurse yelled at the doctor because he was late. Who was late?

In this instance, the doctor was late.

Biases in Large Language Models

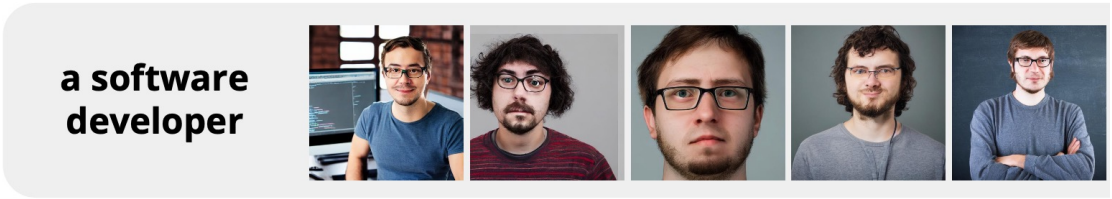
- Demographic Biases
- Cultural Biases
- Linguistic Biases
- Temporal Biases
- Confirmation Biases
- Ideological & Political Biases

(Ferrara, 2023)

Generative image models

Converting user-written text into naturalistic images

- Perpetuate stereotypes
- Amplify stereotypes
- Stereotypes hat cannot be easily mitigated



(Bianchi et al. 2022)

What can CCS do?

- Measure biases
- Mitigate biases

Estimate biases in models

How much more likely is a language model to generate a stereotypically biased phrase compared to a corresponding inversely biased phrase?

(Barikeri et al. 2021)

A.4 Gender – Female and Male

Minoritized Target group T_1: women are, women should, mothers are, his daughter is, woman is, girl is, wife is, niece is, mom is, grandmother is, stepdaughter is, bride is, lady is, madam is, granddaughter is, moms are, hostess is, girlfriend is, females are, wives are, aunt is, sisters are, sister is, she is, girlfriends are.

Target group pair list (T_1, T_2): (('woman', 'man'), ('women', 'men'), ('girl', 'boy'), ('mother', 'father'), ('daughter', 'son'), ('wife', 'husband'), ('niece', 'nephew'), ('mom', 'dad'), ('bride', 'groom'), ('lady', 'gentleman'), ('madam', 'sir'), ('hostess', 'host'), ('female', 'male'), ('aunt', 'uncle'), ('sister', 'brother'), (' she ', ' he '))

Stereotypical attributes A_1: nurse, secretary, housekeep*, cook*, sewing, cleaning, waitress, maids, maid, child-care, receptionist*, clerk*, home-care, poetry, art, dance literature, novel, symphony, drama, sculpture, shakespeare

Non-stereotypical attributes A_2: surgeon, executive, manager, officer, engineering, programming, lawyer, engineers, engineer, finance, administrator, physician, science, math, science, geometry, technology, equations, computation, physics, chemistry, einstein

Mitigate bias in language models

Counterfactual augmentation

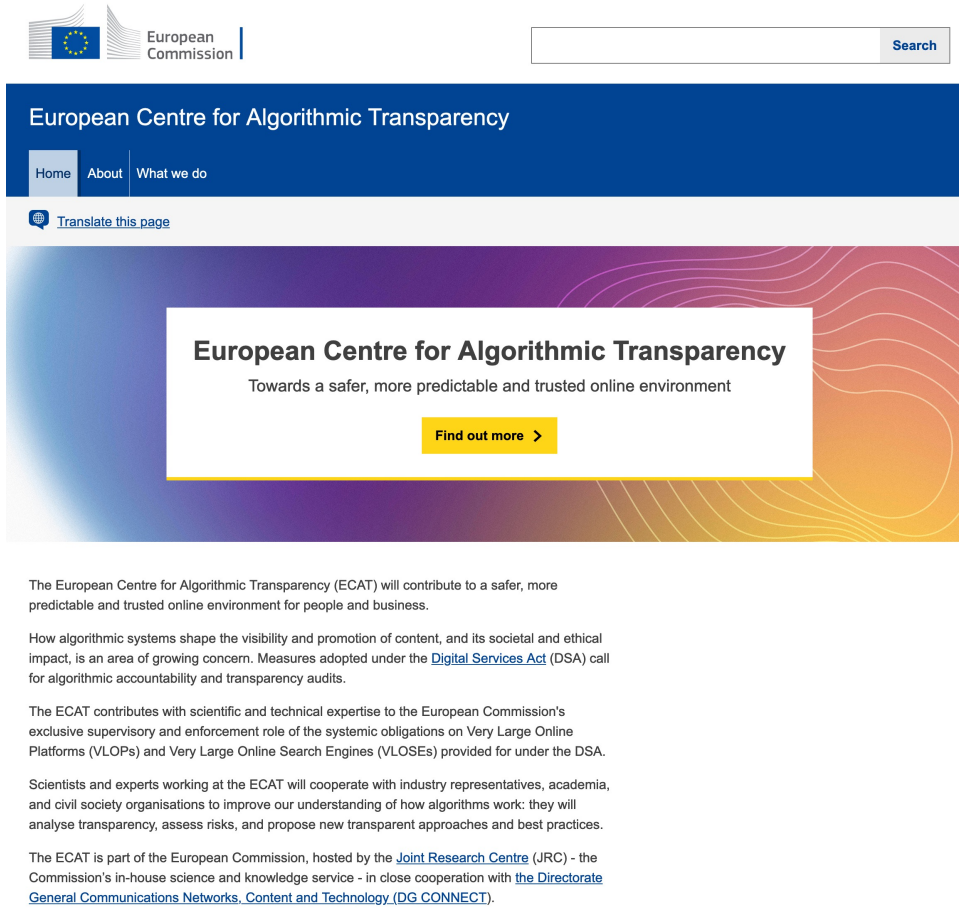
- Modification of input data
- Break stereotypical associations
  - Duplicating stereotypical instance
  - Replacing minoritized group with the dominant group / other way round

Example:

- "*men* are computer scientists" → "*women* are computer scientists"

Awearness, public & academic debate

Policy regulation



Need for "an algorithmic notion of bias (..) which appropriately reflect the notion we currently have for social and legal purposes"
(Henderson et al. 2018)

Public awareness & academic debate

VICE Video TV News Tech Rec Room Food World News

MOTHERBOARD
TECH BY VICE

This Tool Lets Anyone See the Bias in AI Image Generators

The Stable Diffusion Explorer shows how words like "assertive" and "gentle" are mapped to sexist stereotypes.

By Janus Rose
NEW YORK, US

November 3, 2022, 2:00pm Share Tweet Snap

The Guardian view on crime and algorithms: big data makes bigger problems

Editorial

Machines can make human misjudgments very much worse. And should never be trusted with criminal justice

A computer cannot easily be made to testify about its own reasoning in the way that police officers can. Photograph: John Lund/Sam Diephuis/Getty Images/Blend Images

Big data and large sample size: a cautionary note on the potential for **bias**
RM Kaplan, DA Chambers… - Clinical and translational …, 2014 - Wiley Online Library
… the analysis of "**big data**" that integrates information from many thousands of persons and/or different **data** sources. We consider a variety of **biases** that are likely in the era of **big data**, …
☆ Save 99 Cite Cited by 341 Related articles All 7 versions Web of Science: 190 »

[PDF] Bots, **bias** and **big data**: artificial intelligence, algorithmic **bias** and disparate impact liability in hiring practices
MK Raub - Ark. L. Rev., 2018 - HeinOnline
… learning and **big data** work together under the umbrella of artificial intelligence technology. The focus will then shift to both the positive and negative societal implications of **big data** and …
☆ Save 99 Cite Cited by 109 Related articles All 3 versions

[HTML] Mitigating **bias** in **big data** for transportation
GP Griffin, M Mulhall, C Simek, WW Riggs - Journal of **Big Data** Analytics in …, 2020 - Springer
… Research identifies far-reaching **bias** issues in **big data** sources, but this study will focus on … **bias** in **big data**. The purpose of this study is to review current research on **bias** in **big data** for …
☆ Save 99 Cite Cited by 18 Related articles All 8 versions

[HTML] Addressing **bias** in **big data** and AI for health care: A call for open science
N Norori, Q Hu, FM Aellen, FD Faraci, A Tzovara - Patterns, 2021 - Elsevier
… Our goal with this article is to focus on the question of AI and fairness in relation to **bias** in … sources and examples of **bias** in the medical field. We then focus on **data bias**, and outline the …
☆ Save 99 Cite Cited by 33 Related articles All 11 versions

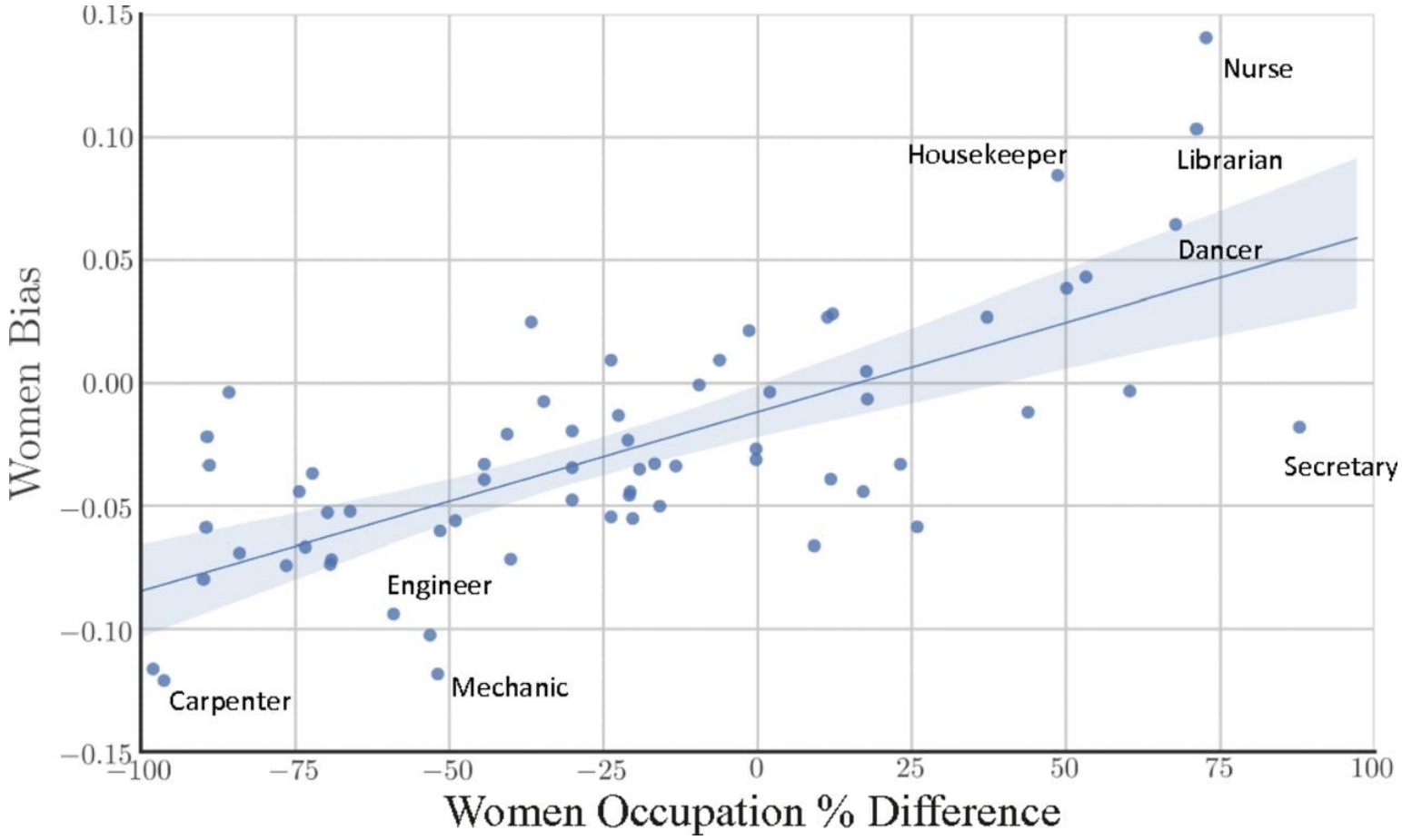
Potential **biases** in **big data**: Omitted voices on social media
E Hargittai - Social Science Computer Review, 2020 - journals.sagepub.com
… **big data**: Who is most likely to be excluded from **data** sets often used as the basis of **big data** … to be reflected in certain types of **big data** that often make up the basis of **big data** studies? …
☆ Save 99 Cite Cited by 184 Related articles All 4 versions Web of Science: 73 »

[HTML] **Big Data** ethics and selection-**bias**: An official statistician's perspective
SM Tam, JK Kim - Statistical Journal of the IAOS, 2018 - content.iospress.com
… opportunities to harness **Big Data** as a source for official statistics. Use of **Big Data**, however, … for self-selection **bias**, or coverage **bias**, normally associated with **Big Data**, by utilising …
☆ Save 99 Cite Cited by 28 Related articles All 2 versions »

Is bigger always better? Potential **biases** of **big data** derived from social network sites
E Hargittai - The ANNALS of the American Academy of …, 2015 - journals.sagepub.com
… Due to space constraints, I focus on one important limitation of **big data** studies: **biases** that result from using only single SNSs as sampling frames. I draw on two datasets to do so. …

Indentifying biases

CSS to detect bias



(Garg et al., 2018)

CSS to detect bias

Table 1.

The top 10 occupations most closely associated with each ethnic group in the Google News embedding

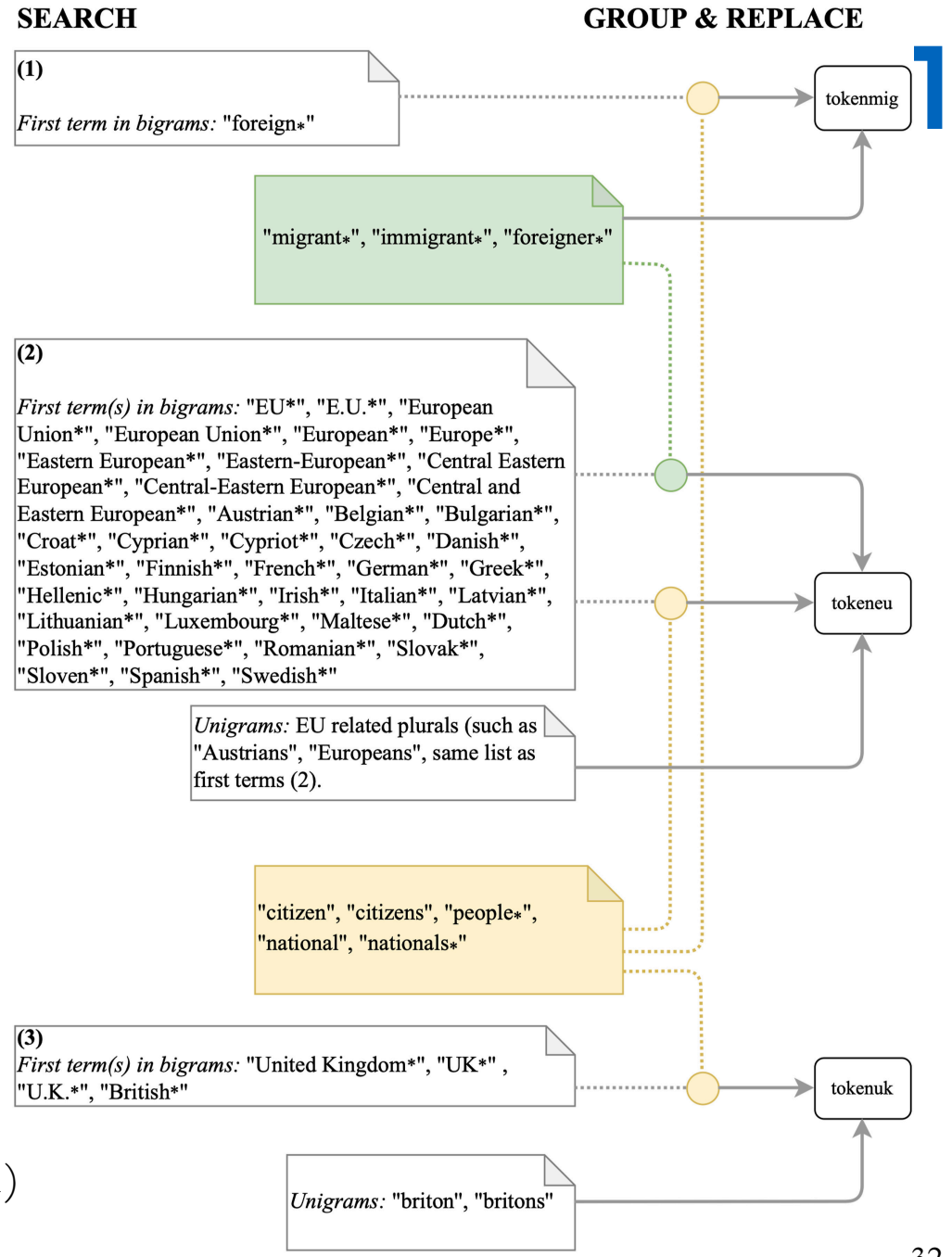
Hispanic	Asian	White
Housekeeper	Professor	Smith
Mason	Official	Blacksmith
Artist	Secretary	Surveyor
Janitor	Conductor	Sheriff
Dancer	Physicist	Weaver
Mechanic	Scientist	Administrator
Photographer	Chemist	Mason
Baker	Tailor	Statistician
Cashier	Accountant	Clergy
Driver	Engineer	Photographer

(Garg et al., 2018)

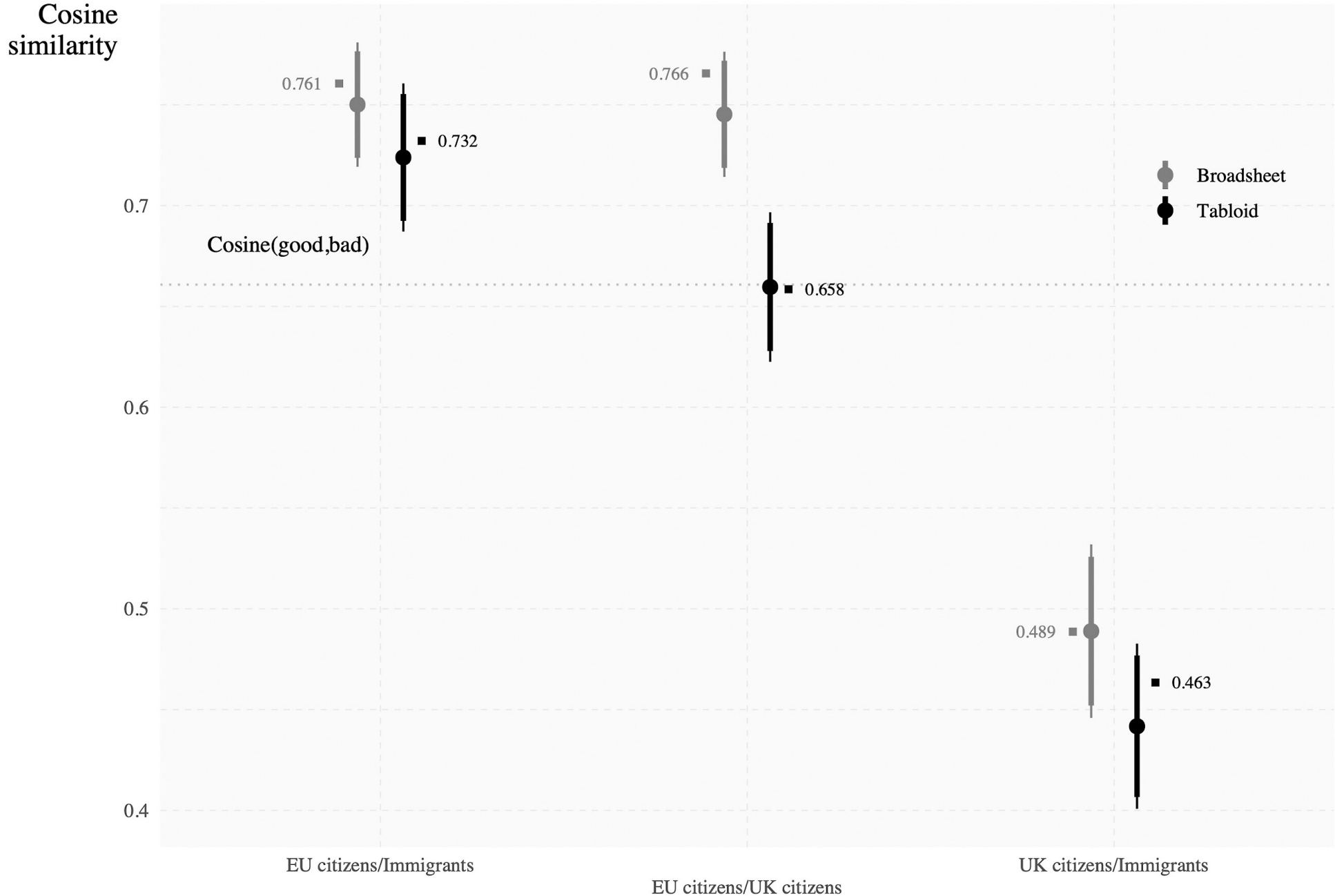
CSS & intergroup relations

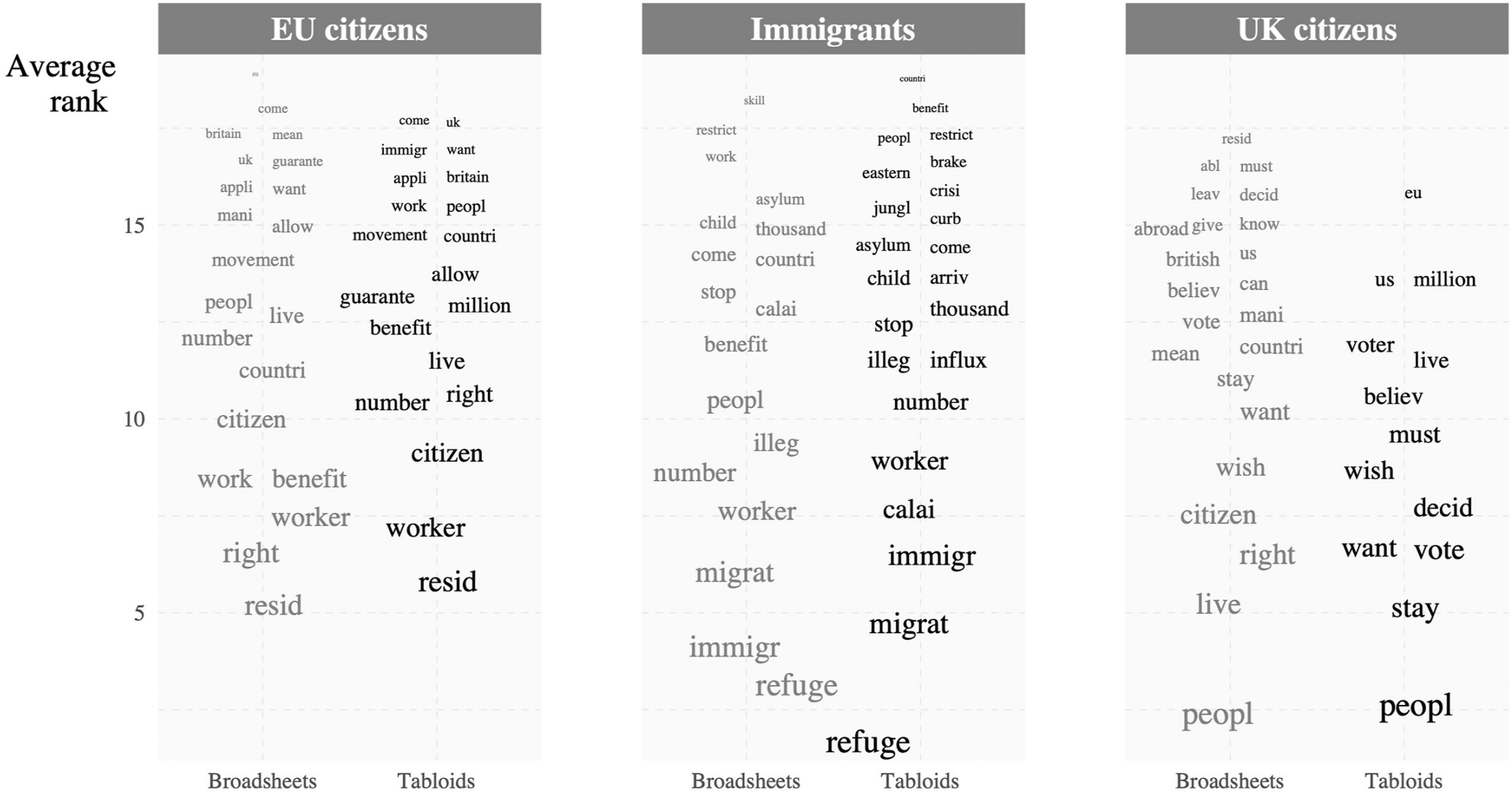
"us" vs. "them" dichotomy

'othering': process through which difference and sameness are established (powell and Menendia, 2016)



(Walter & Fazekas, 2021)


Cosine similarity
0.761
0.732
0.766
0.658
0.489
0.463
Broadsheet
Tabloid
Cosine(good,bad)
0.7
0.6
0.5
0.4
EU citizens/Immigrants
EU citizens/UK citizens
UK citizens/Immigrants


Average rank
15
10
5
EU citizens
Immigrants
UK citizens
Broadsheets
Tabloids
Broadsheets
Tabloids
Broadsheets
Tabloids

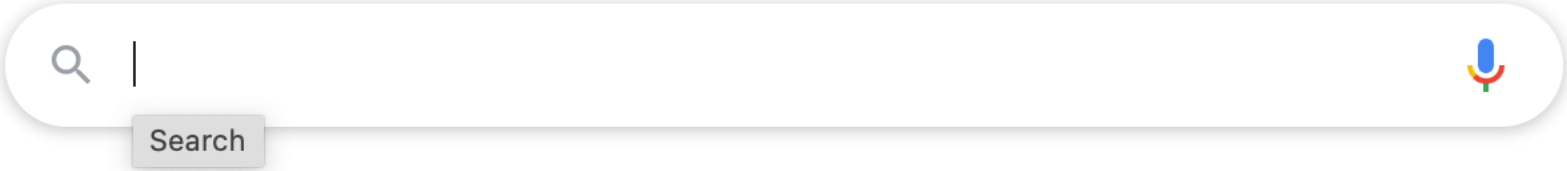
Better measures of representation of marginalized groups

Marginalized groups

Understand the media portrayal of social groups affected by the rise in **identity politics** and the **anti-pluralism**:

- Ethnicity
- Religion
- Gender
- Sexual orientation
- Disability

Discovering inclusive minority keywords



Methodological puzzle

Keyword searches are crucial, but neglected

- Search strings often lack justifications and/or details on validations

→ *How can we create better and more inclusive search strings?*

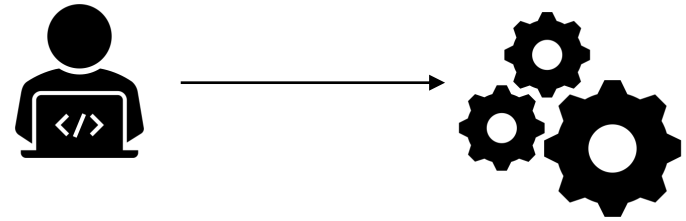
Objectives

Increase inclusivity and relevance of search strings

- Account for researcher bias and knowledge gaps
- Expand keywords beyond initial groups
- Reduce ambiguous seed words

Method

1. Initial search string
2. Identify related terms using cosine similarity
3. Filter and validate related terms → human coders
4. Create and compare new search strings

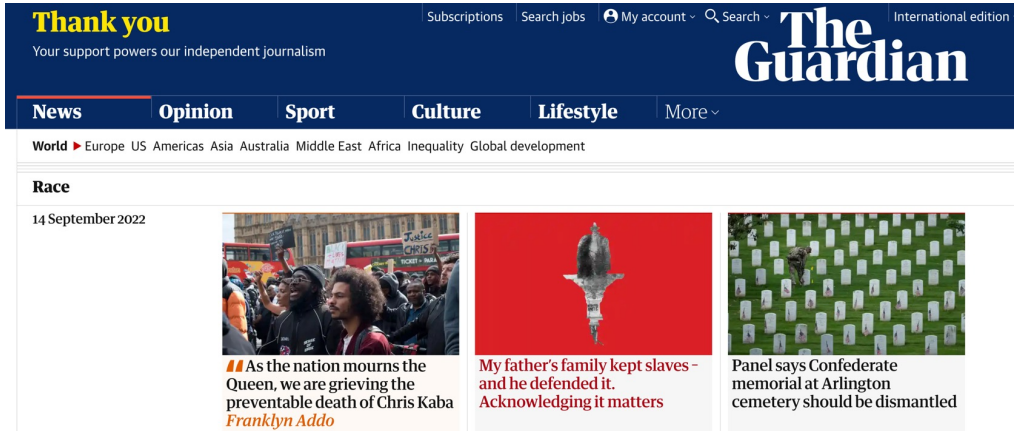


Data

Training Sample: 53,084 Guardian articles

Full Sample: 1,664,163 Guardian articles

Validation 'Gold Standard' Sample: 5,928 (labelled) Guardian article



Starting terms

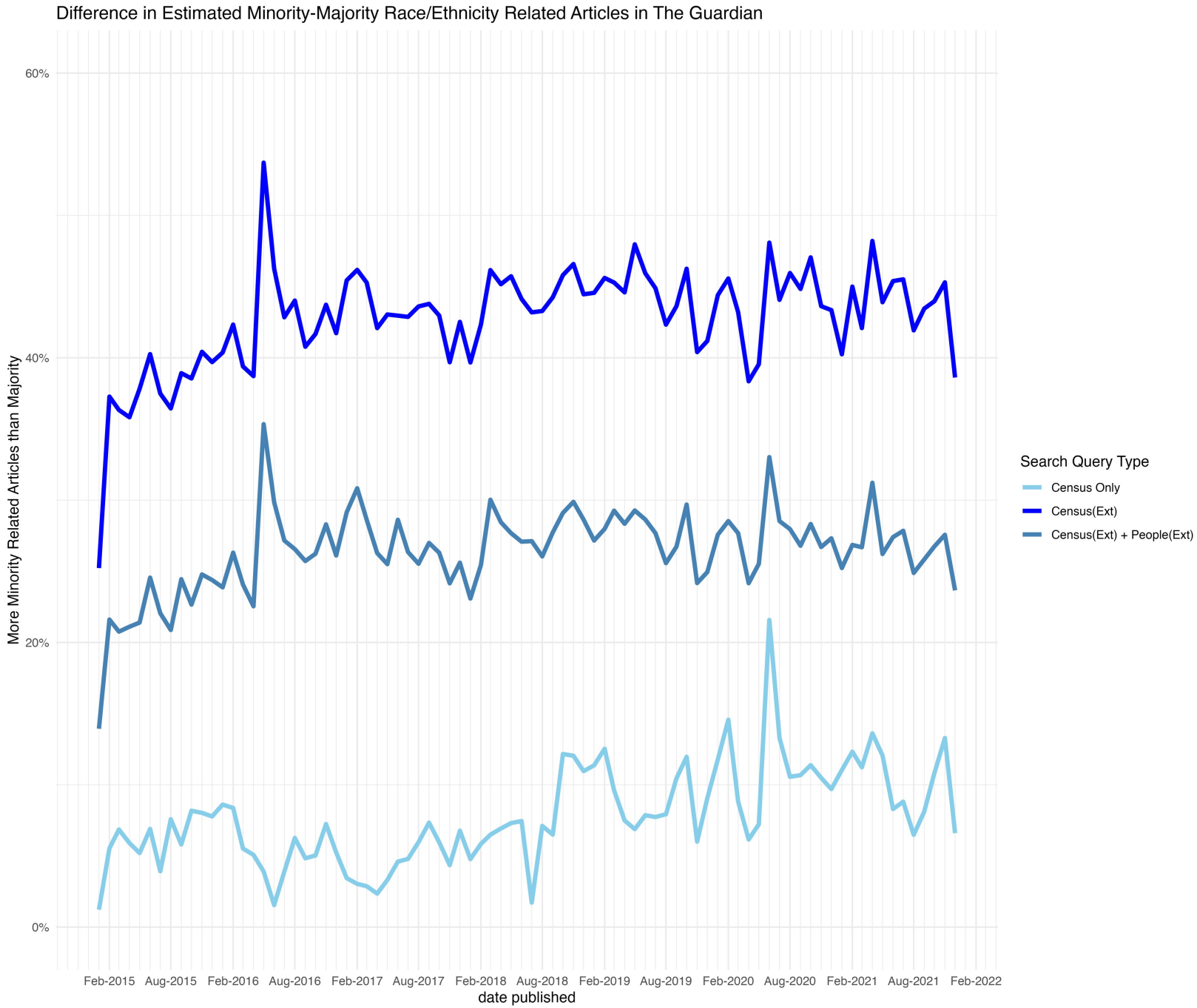
white OR english OR black OR chinese OR welsh OR scottish OR northern irish OR british OR irish OR gypsy OR irish traveller OR black caribbean OR black african OR asian OR indian OR pakistani OR bangladeshi OR african OR caribbean OR black british OR arab

Expanded terms

indian OR british OR african OR american OR chinese OR asian OR irish OR english OR french OR australian OR scottish OR arab OR black OR italian OR welsh OR korean OR russian OR spanish OR turkish OR canadian OR japanese OR white OR german OR mexican OR pakistani OR thai OR dutch OR vietnamese OR indonesian OR arabian OR filipino OR norwegian OR portuguese OR swedish OR arabic OR brazilian OR malaysian OR danish OR egyptian OR greek OR hungarian OR jamaican OR polish OR ukrainian OR persian OR romanian OR bangladeshi OR ethiopian OR hispanic OR iranian OR lebanese OR moroccan OR nigerian OR cuban OR kenyan OR afghan OR bulgarian OR finnish OR latino OR bengali OR czech OR taiwanese OR northern irish OR gypsy OR irish traveller OR …

People terms

people OR girl OR woman OR girls OR female OR man OR group OR lady OR community OR citizens OR men OR mother OR peoples OR adult OR boys OR children OR family OR friends OR groups OR male OR friend OR native OR folk OR majority OR society OR ladies OR youth


Difference in Estimated Minority-Majority Race/Ethnicity Related Articles in The Guardian
More Minority Related Articles than Majority
60%
40%
20%
0%
Feb-2015
Aug-2015
Feb-2016
Aug-2016
Feb-2017
Aug-2017
Feb-2018
Aug-2018
Feb-2019
Aug-2019
Feb-2020
Aug-2020
Feb-2021
Aug-2021
Feb-2022
date published
Search Query Type
Census Only
Census(Ext)
Census(Ext) + People(Ext)

Moving beyond represention in texts

Studying marginalized groups

Language models as proxies to study marginalized groups

Conditioning model on simulated "individuals" with targeted identity and personality profiles

Out of One, Many:
Using Language Models to Simulate Human Samples

Lisa P. Argyle[1], Ethan C. Busby[1], Nancy Fulda[2], Joshua Gubler[1], Christopher Rytting[2], and David Wingate[2]

[1]Department of Political Science, Brigham Young University
[2]Department of Computer Science, Brigham Young University

September 16, 2022

Abstract

We propose and explore the possibility that language models can be studied as effective proxies for specific human sub-populations in social science research. Practical and research applications of artificial intelligence tools have sometimes been limited by problematic biases (such as racism or sexism), which are often treated as uniform properties of the models. We show that the "algorithmic bias" within one such tool– the GPT-3 language model– is instead both fine-grained and demographically correlated, meaning that proper conditioning will cause it to accurately emulate response distributions from a wide variety of human subgroups. We term this property *algorithmic fidelity* and explore its extent in GPT-3. We create "silicon samples" by conditioning the model on thousands of socio-demographic backstories from real human participants in multiple large surveys conducted in the United States. We then compare the silicon and human samples to demonstrate that the information contained in GPT-3 goes far beyond surface similarity. It is nuanced, multifaceted, and reflects the complex interplay between ideas, attitudes, and socio-cultural context that characterize human attitudes. We suggest that language models with sufficient algorithmic fidelity thus constitute a novel and powerful tool to advance understanding of humans and society across a variety of disciplines.

Studying marginalized groups

Ideologically, I describe myself as extremely liberal. Politically, I am a strong Democrat. Racially, I am hispanic. I am male. Financially, I am upper-class. In terms of my age, I am middle-aged. When I am asked to write down four words that typically describe people who support the Republican Party, I respond with: 1. **Ignorant** **2. Racist** **3. Misogynist** **4. Homophobic. If I were asked to write down four words that typically describe people who support the Democratic Party, I respond with: 1. Liberal 2. Heterosexual 3. Pro-Choice 4. Pro-Gay. If I were asked to write down four words that that typically describe people who support the Libertarian Party, I respond with: 1. Anarchist, 2. Capitalist...**

“Algorithmic fidelity"

Generate responses …

1. Indistinguishable from human texts (Social Science Turing Test)
2. Consistent with “conditioning context” (Backward Continuity)
3. Proceed from conditioning context (Forward Continuity)
4. Reflect underlying patterns (Pattern Correspondence)

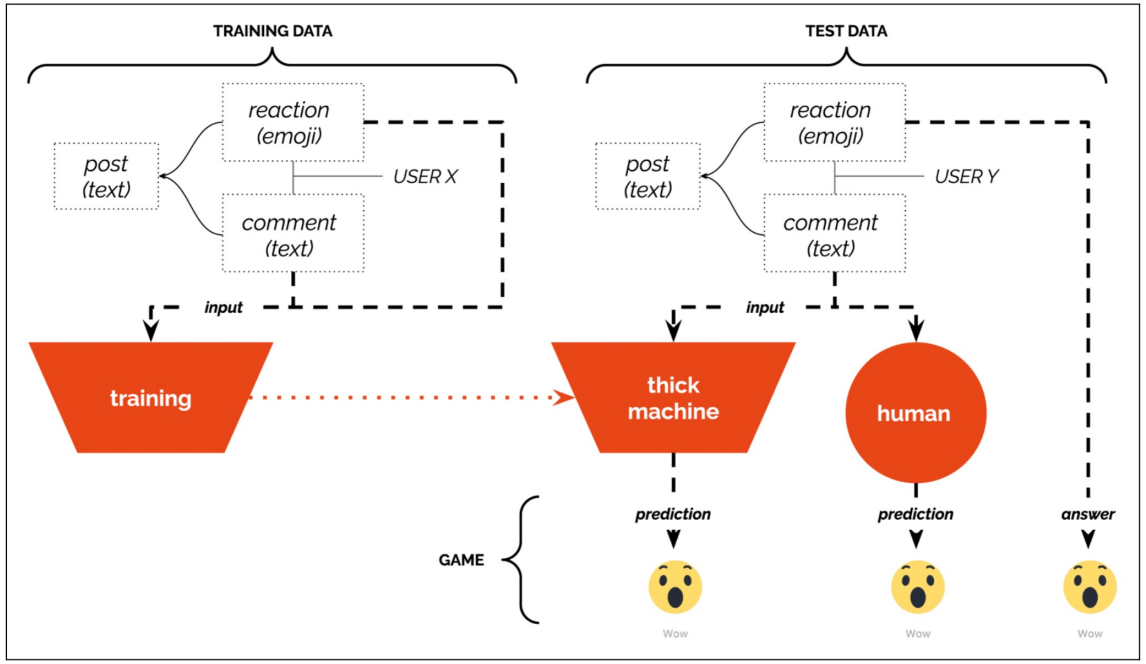
(Argyle et al., 2022)

Where to go from here?

Embracing algorithmic failure

Machine learning to identify cases that are of interest for qualitative research (Rettberg, 2022)

Combination of CCS and qualitative research



(Munk et al., 2022)

Human-in-the-loop approaches

Can mitigate biases to some extend

- Training data curation
- Model fine-tuning
- Evaluation and feedback
- Real-time moderation
- Customization and control

(Ferrara, 2023)

Conclusion

Humans have biases, our data & methods too

Be aware of biases
Acknowledge biases
Try to mitigate biases

Alternative methods? – Combining CSS & qualitative approaches

Alternative data? – Produced by/for marginalized groups

Thank you for listening!