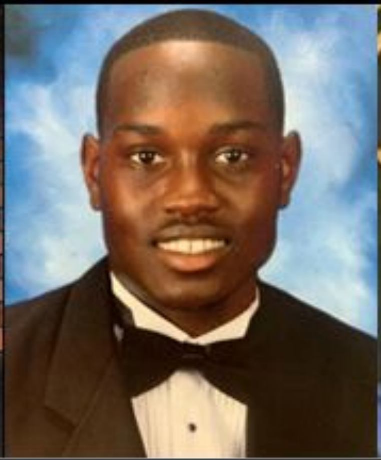
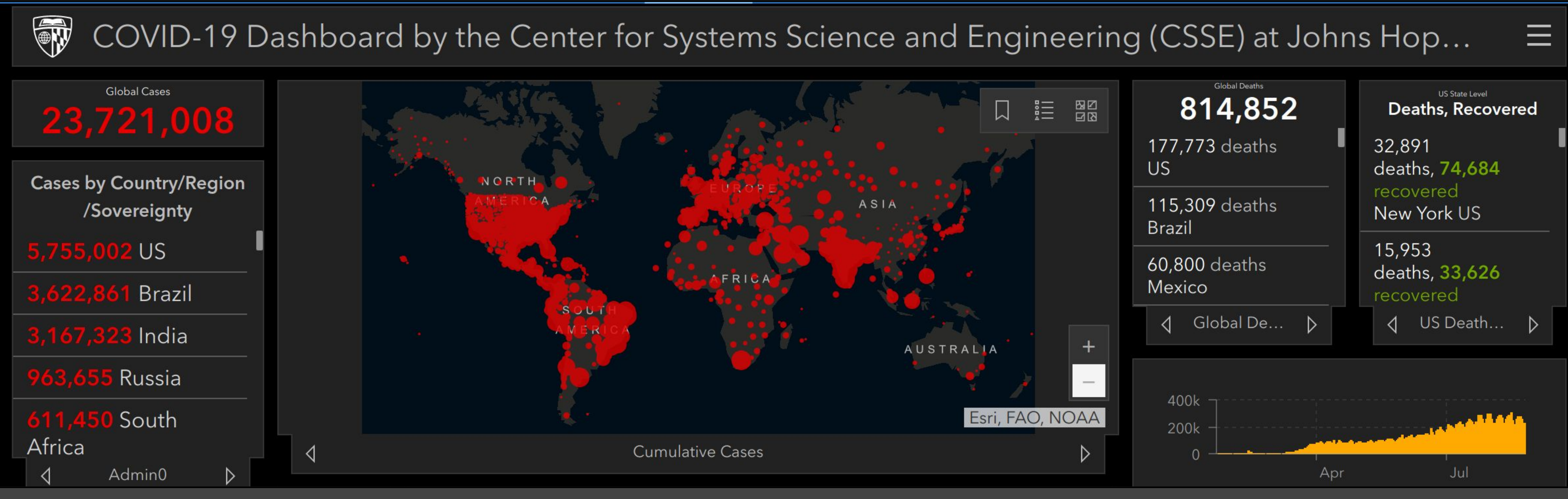
NLP, Ethics, & Social Change

Eve Fleisig & Rediet Abebe



NEW YORKERS FOR IMMIGRANT JUSTICE!
nyic
IMMIGRANT RIGHTS ARE HUMAN RIGHTS
AFRICAN SERVICES COMMITTEE
www.africanservices.org



COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hop…

Global Cases
23,721,008

Cases by Country/Region/Sovereignty

5,755,002 US

3,622,861 Brazil

3,167,323 India

963,655 Russia

611,450 South Africa

Admin0

Cumulative Cases

Global Deaths
814,852

177,773 deaths US

115,309 deaths Brazil

60,800 deaths Mexico

Global De…

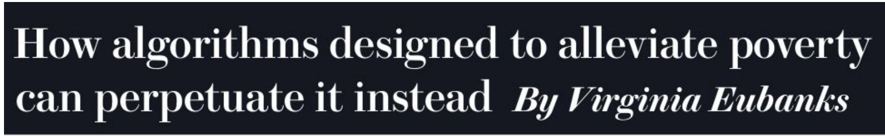
US State Level
Deaths, Recovered

32,891 deaths, 74,684 recovered New York US

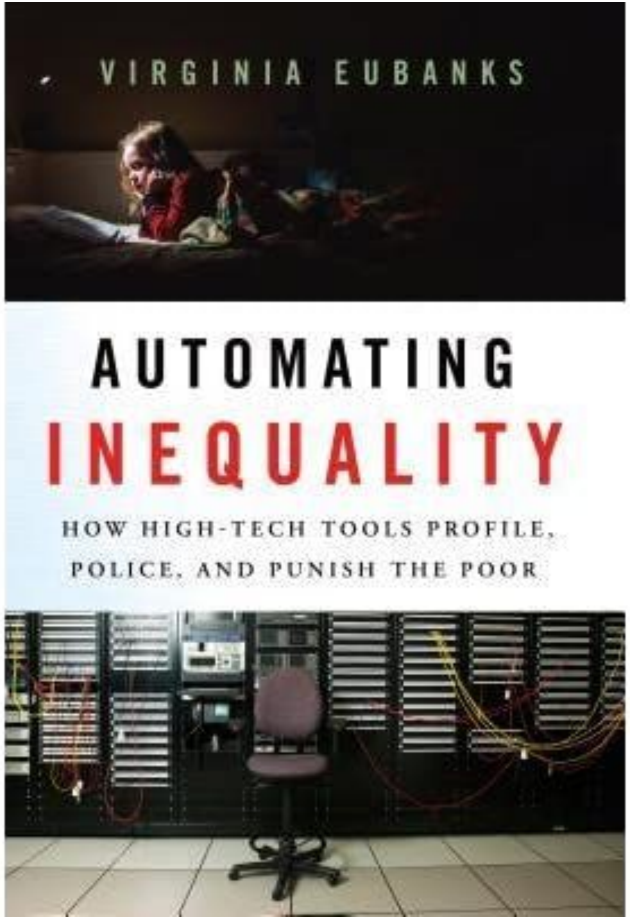
15,953 deaths, 33,626 recovered

US Death…

Anti-Black racial bias in healthcare algorithm.



How algorithms designed to alleviate poverty can perpetuate it instead *By Virginia Eubanks*



VIRGINIA EUBANKS

AUTOMATING INEQUALITY

HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR

Pitfalls of computing for social good.

Solutionism: tendency to assume computing will solve social problems.

Tinkering: take problematic sociopolitical systems as fixed and optimize around them.

Diversion: distract from the root of problems and other forms of addressing them.

Does computing have *any* role to play?

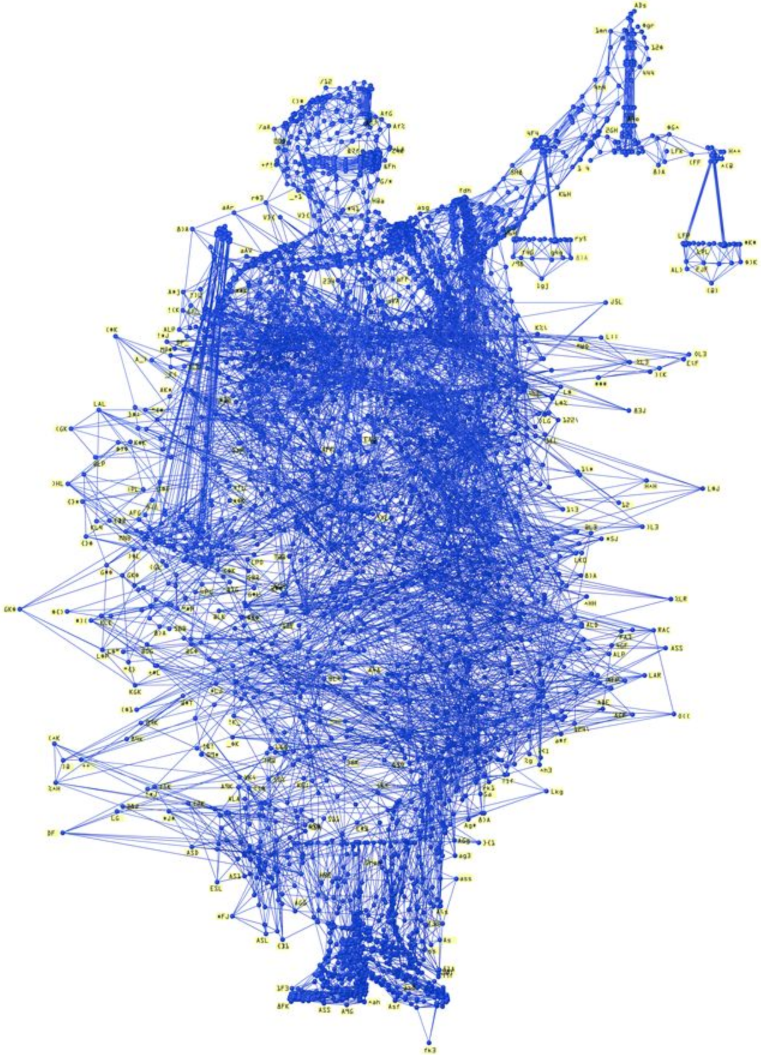
Does computing have *any* role to play?

Roles for Computing in Social Change

Abebe, Barocas, Kleinberg, Levy,
Raghavan, & Robinson (FAT* ’20)

Social change is the work of many hands.

How can computing support (not supplant) other routes to broad social change?

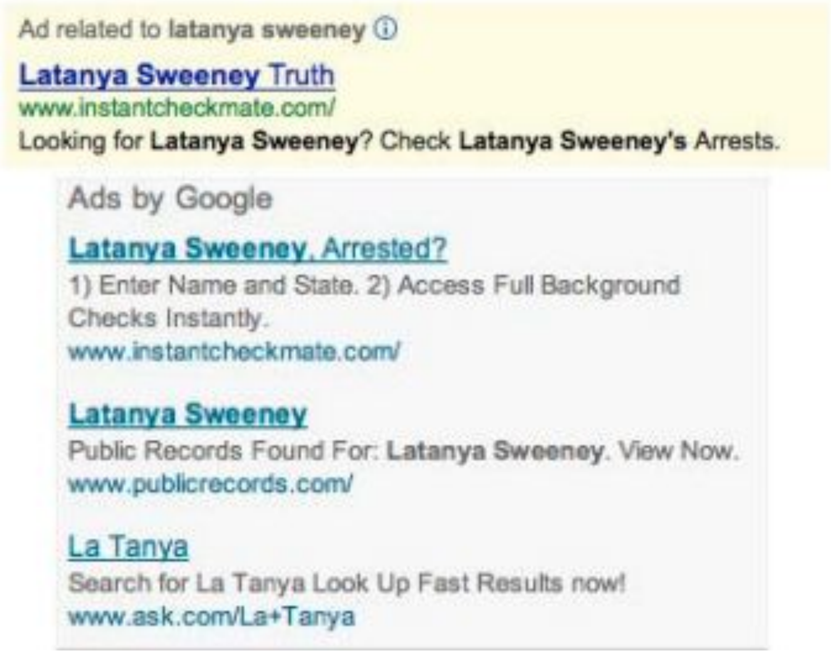


Computing as diagnostic.

Computing can help us measure social problems and diagnose how they manifest in technical systems.

Computing as diagnostic.

Sweeney (2013): racial bias in ad delivery



Ad related to **latanya sweeney** ⓘ

Latanya Sweeney Truth

www.instantcheckmate.com/

Looking for **Latanya Sweeney**? Check **Latanya Sweeney's** Arrests.

Ads by Google

Latanya Sweeney, Arrested?

1) Enter Name and State. 2) Access Full Background Checks Instantly.

www.instantcheckmate.com/

Latanya Sweeney

Public Records Found For: **Latanya Sweeney**. View Now.

www.publicrecords.com/

La Tanya

Search for La Tanya Look Up Fast Results now!

www.ask.com/La+Tanya

Computing as diagnostic.

Sweeney (2013): racial bias in ad delivery

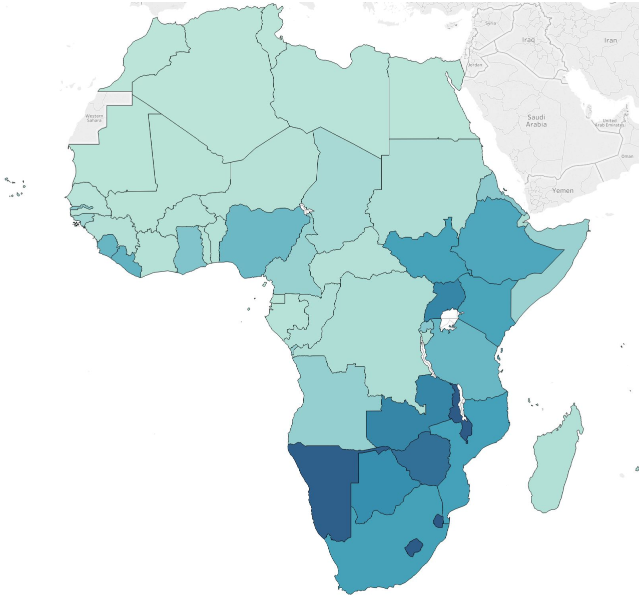
Buolamwini & Gebru (2018): gender and skin-tone bias in facial analysis

Bolukabsi et al. (2016): Cliskan et al. (2017): bias in word embeddings

Obermeyer et al. (2019), racial bias in healthcare algorithms

Koenecke et al. (2020): racial bias in speech recognition

Abebe et al. (2019): health information bias in search engines

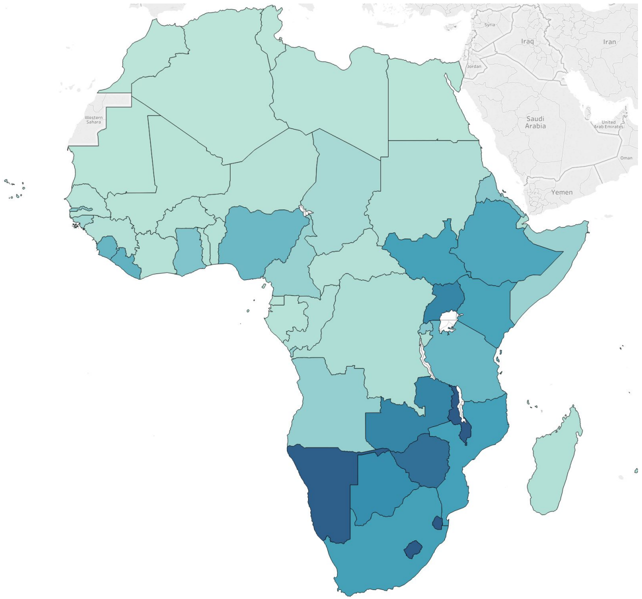


Using Search Queries to Understand Health Information Needs in Africa

Abebe, Hill, Vaughan, Small, & Schwartz
(ICWSM, ’19)

Drugs: drug, treatment, patients, abuse, therapy, drugs, resistance, antiretroviral, substance

Symptoms: pain, sign, lymph, swollen, nodes, sore, symptom, symptoms, throat, infection



Breastfeeding: positive, baby, mother, breastfeeding, breast, mothers, child, born, feeding, babies

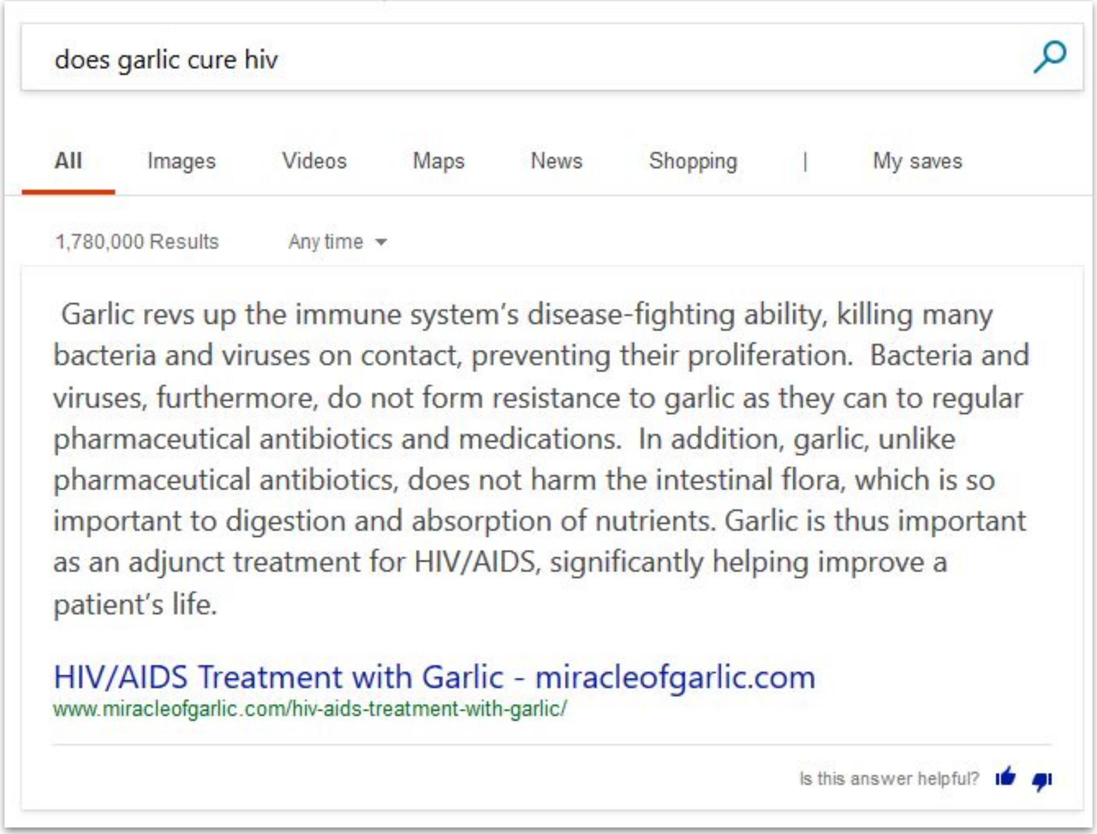
Stigma: stigma, issues, discrimination, related, ethical, legal, prevention, safety, pdf, workplace

healthy lifestyle: food, positive, people, person, diet, healthy, living, eat, good patients, medication, nutrition, foods, lifestyle

Natural Cures: cure, oil, black, healing, heal, healed, seed, herbs, natural, cures

Is there variance in the **quality of content** shown to users online?

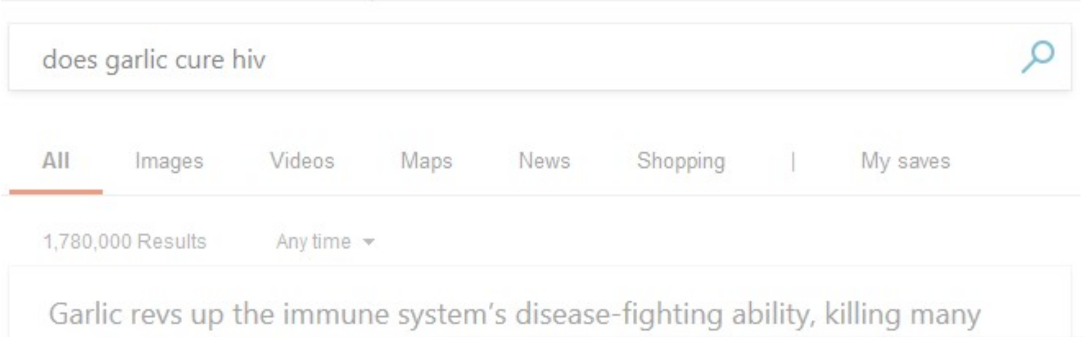
Natural cures



Garlic revs up the immune system's disease-fighting ability, killing many bacteria and viruses on contact, preventing their proliferation. Bacteria and viruses, furthermore, do not form resistance to garlic as they can to regular pharmaceutical antibiotics and medications. In addition, garlic, unlike pharmaceutical antibiotics, does not harm the intestinal flora, which is so important to digestion and absorption of nutrients. Garlic is thus important as an adjunct treatment for HIV/AIDS, significantly helping improve a patient's life.

HIV/AIDS Treatment with Garlic - miracleofgarlic.com
www.miracleofgarlic.com/hiv-aids-treatment-with-garlic/

Natural cures



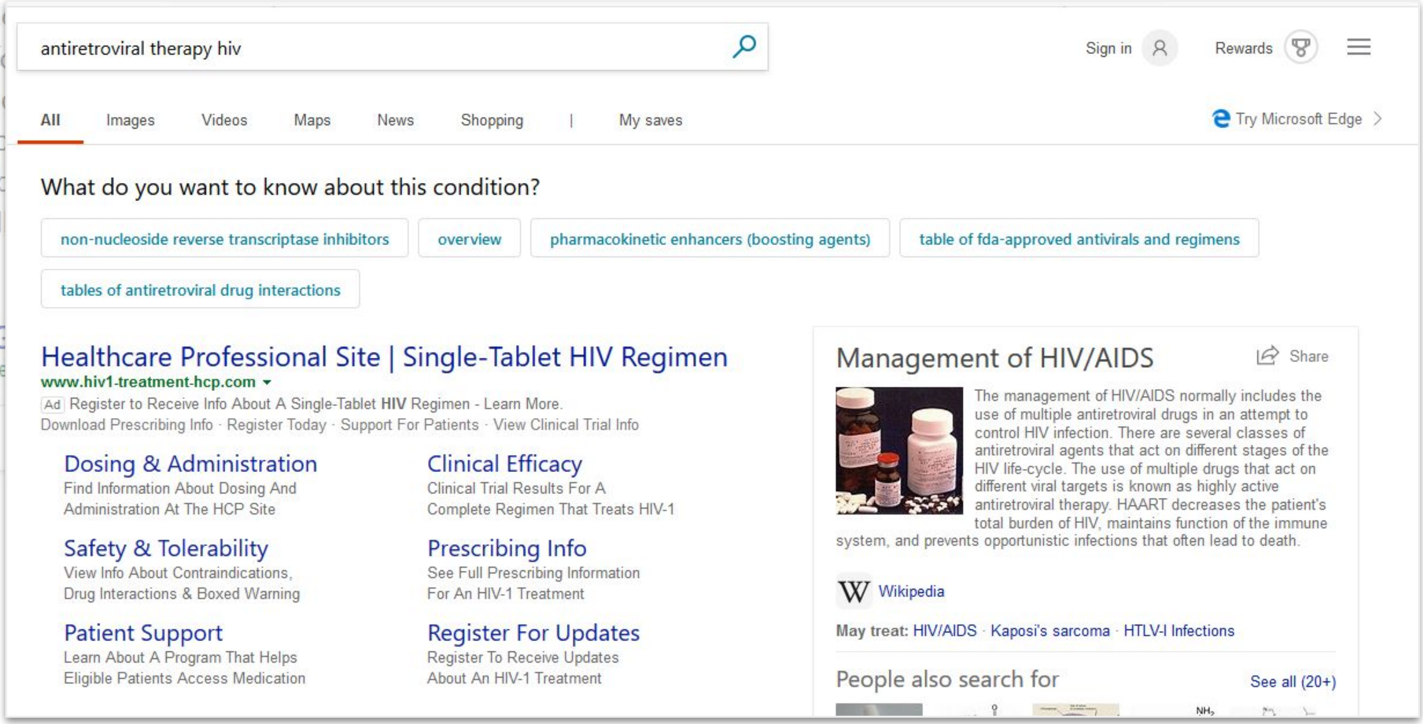
does garlic cure hiv

1,780,000 Results Any time

Garlic revs up the immune system's disease-fighting ability, killing many bacteria and viruses on conta
viruses, furthermore, do not f
pharmaceutical antibiotics an
pharmaceutical antibiotics, d
important to digestion and a
as an adjunct treatment for H
patient's life.

HIV/AIDS Treatment with G
www.miracleofgarlic.com/hiv-aids-treatme

HIV/AIDS drugs



antiretroviral therapy hiv

What do you want to know about this condition?

non-nucleoside reverse transcriptase inhibitors · overview · pharmacokinetic enhancers (boosting agents) · table of fda-approved antivirals and regimens · tables of antiretroviral drug interactions

Healthcare Professional Site | Single-Tablet HIV Regimen
www.hiv1-treatment-hcp.com
Ad Register to Receive Info About A Single-Tablet **HIV** Regimen - Learn More.
Download Prescribing Info · Register Today · Support For Patients · View Clinical Trial Info

Dosing & Administration
Find Information About Dosing And Administration At The HCP Site

Clinical Efficacy
Clinical Trial Results For A Complete Regimen That Treats HIV-1

Safety & Tolerability
View Info About Contraindications, Drug Interactions & Boxed Warning

Prescribing Info
See Full Prescribing Information For An HIV-1 Treatment

Patient Support
Learn About A Program That Helps Eligible Patients Access Medication

Register For Updates
Register To Receive Updates About An HIV-1 Treatment

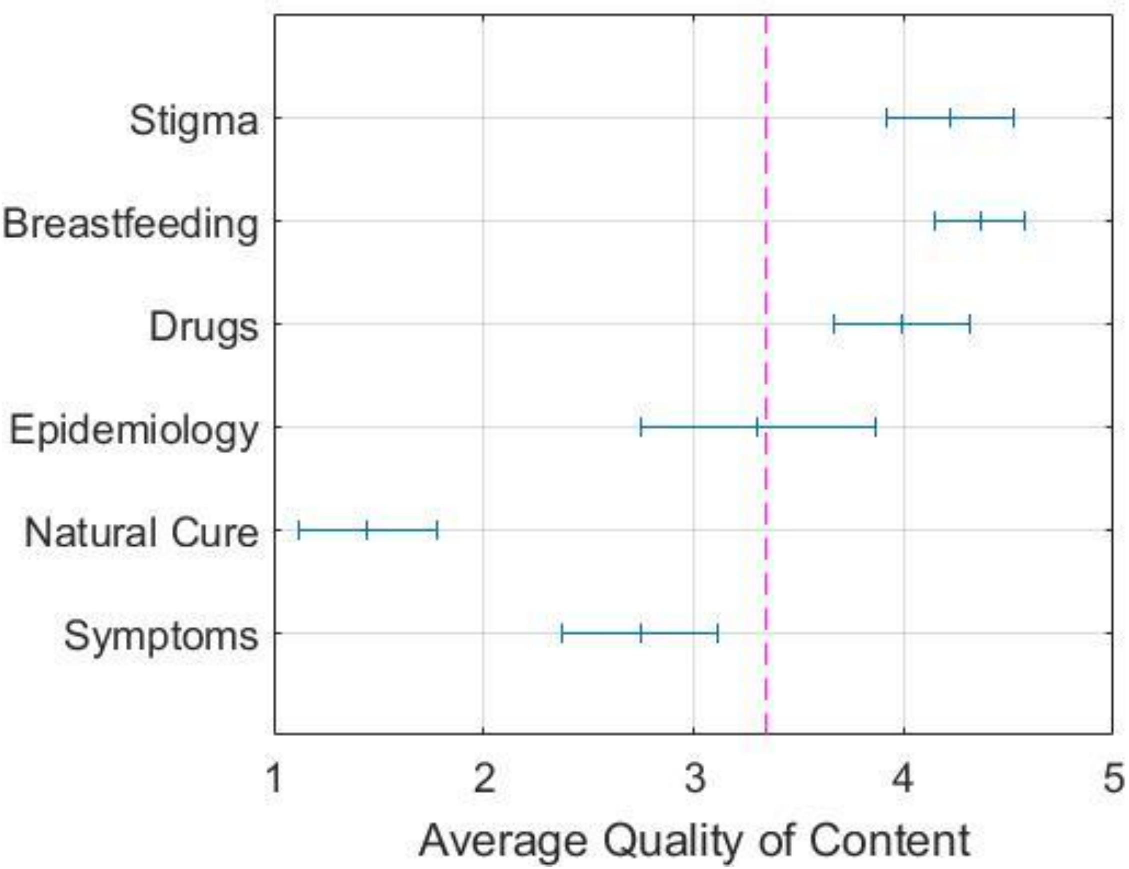
Management of HIV/AIDS

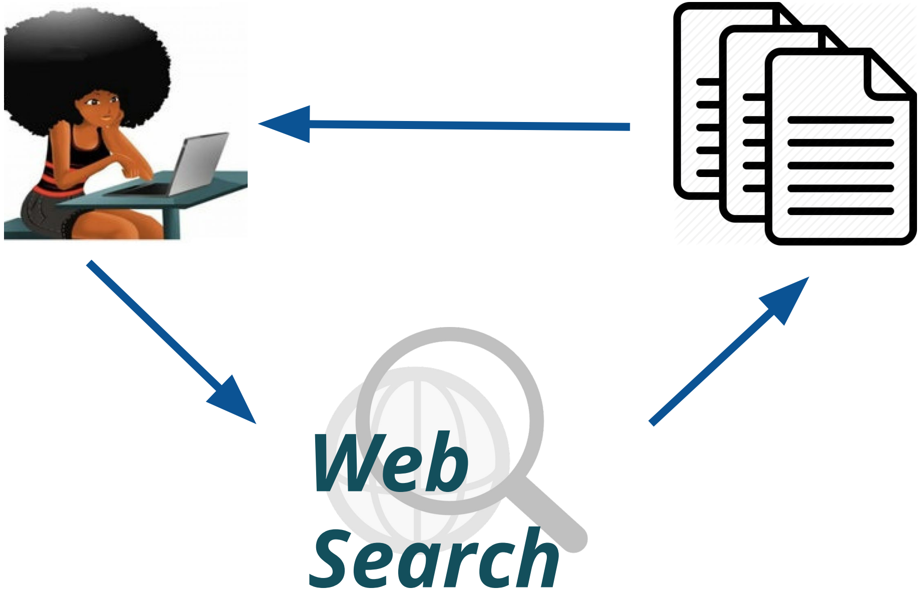
The management of HIV/AIDS normally includes the use of multiple antiretroviral drugs in an attempt to control HIV infection. There are several classes of antiretroviral agents that act on different stages of the HIV life-cycle. The use of multiple drugs that act on different viral targets is known as highly active antiretroviral therapy. HAART decreases the patient's total burden of HIV, maintains function of the immune system, and prevents opportunistic infections that often lead to death.

Wikipedia

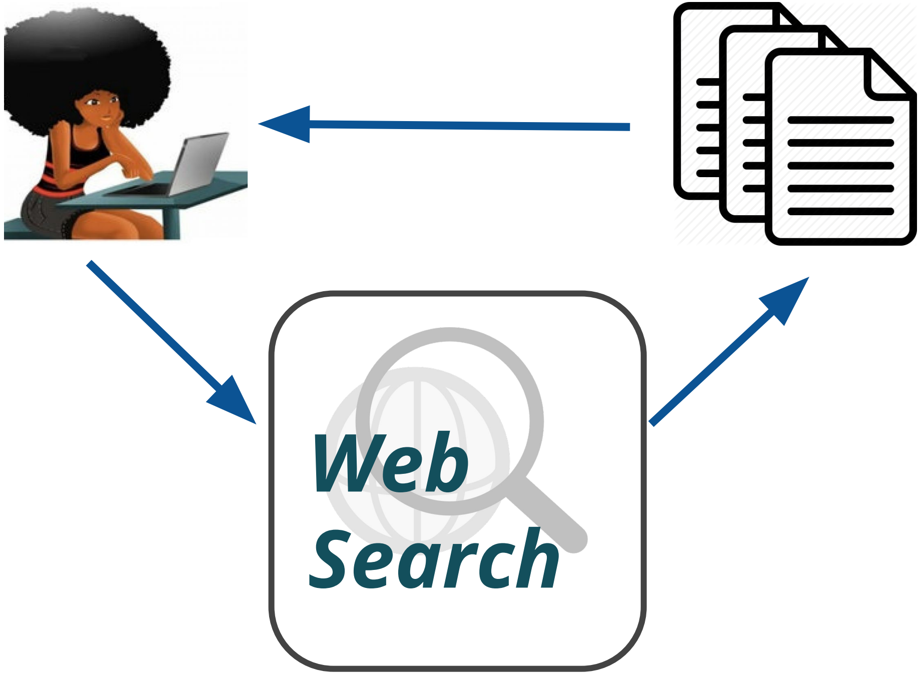
May treat: HIV/AIDS · Kaposi's sarcoma · HTLV-I Infections

People also search for See all (20+)


Stigma
Breastfeeding
Drugs
Epidemiology
Natural Cure
Symptoms
1
2
3
4
5
Average Quality of Content

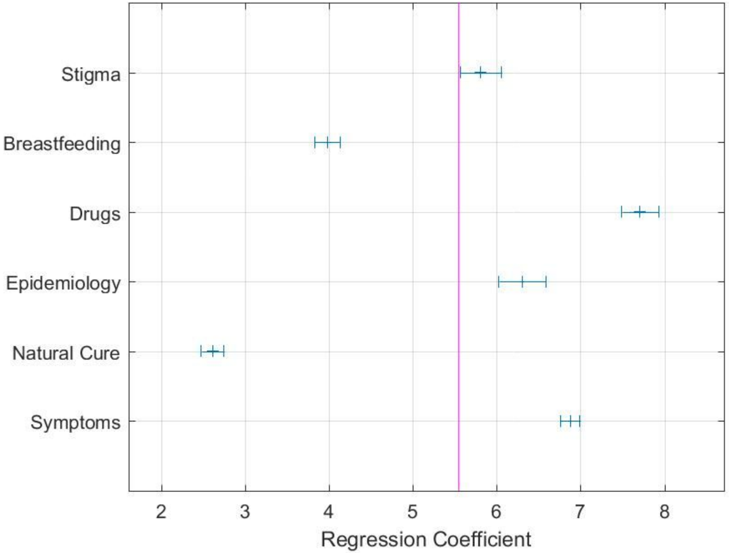


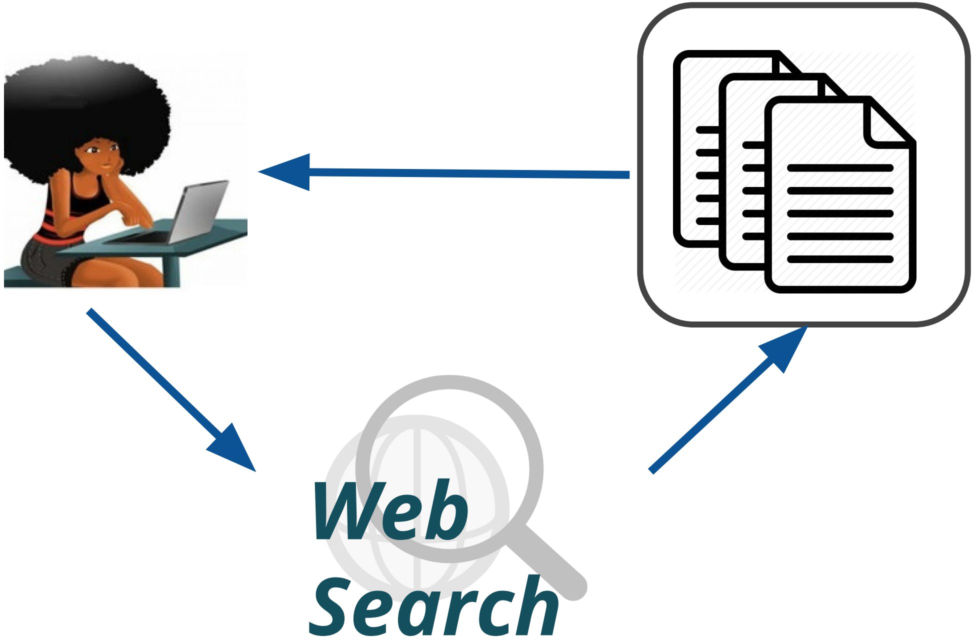
Causes of disparities in access to health information.



Causes of disparities in access to health information.

Discrepancies in “backend” processing by search engines.





Availability of Web pages on *drugs* vs. *natural cure*:

CDC	4.56x
NIH	4.97x
WHO	6.56x
UNAIDS	7.41x

Causes of disparities in access to health information.

Computing as diagnostic.

Computing can help us measure social problems and diagnose how they manifest in technical systems.

Risks: **diagnosis ≠ treatment**

Computing as diagnostic.

Computing can help us measure social problems and diagnose how they manifest in technical systems.

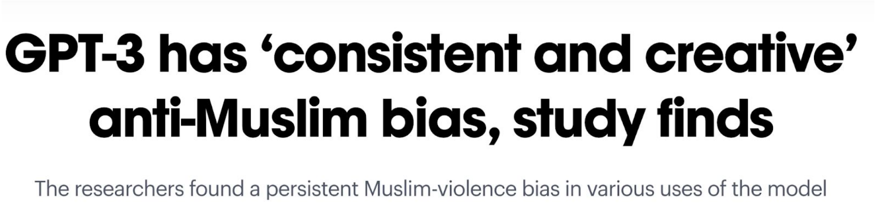
Risks: **diagnosis ≠ treatment**

“Data, in short, do not speak for
themselves and don’t always change
hearts and minds or policy.”
(Benjamin, 2019)

Ethics and Social Change:
NLP Perspectives

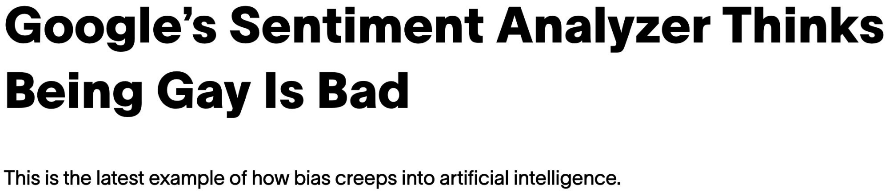
NLP Perspectives: Outline

- **Understanding the Problem**
  - **NLP Gone Wrong**
  - Sources of Bias
  - Bias Measurement
- Addressing Bias
  - Bias Mitigation
  - The Effects of Interventions
- Beyond Bias



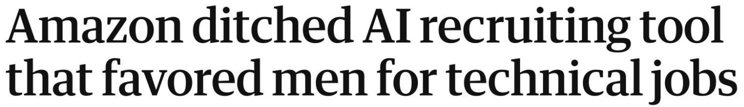
GPT-3 has 'consistent and creative' anti-Muslim bias, study finds

The researchers found a persistent Muslim-violence bias in various uses of the model

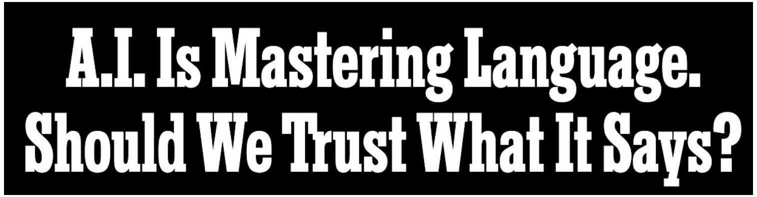


Google's Sentiment Analyzer Thinks Being Gay Is Bad

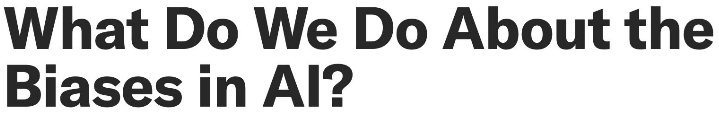
This is the latest example of how bias creeps into artificial intelligence.



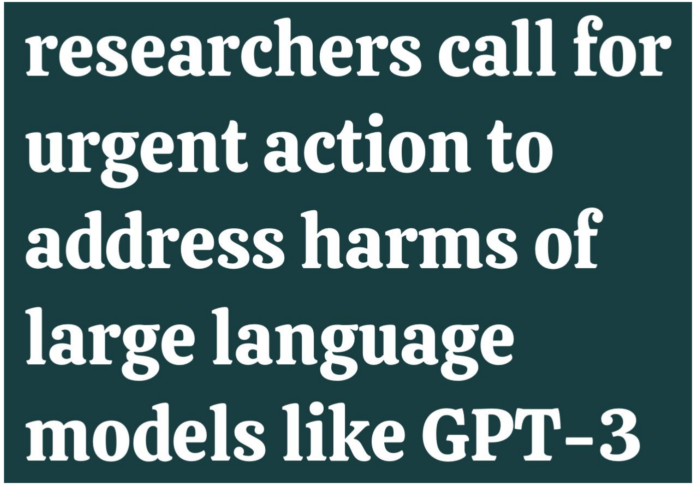
Amazon ditched AI recruiting tool that favored men for technical jobs



A.I. Is Mastering Language. Should We Trust What It Says?



What Do We Do About the Biases in AI?

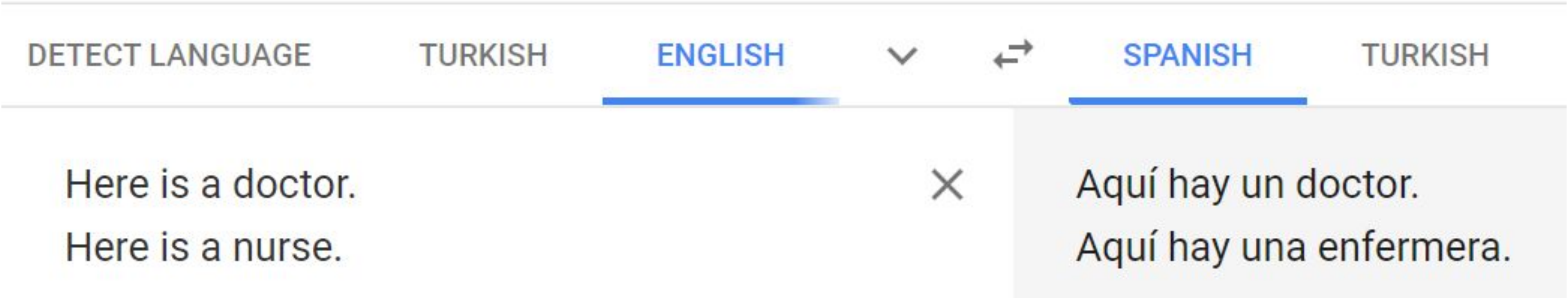


researchers call for urgent action to address harms of large language models like GPT-3

Outline

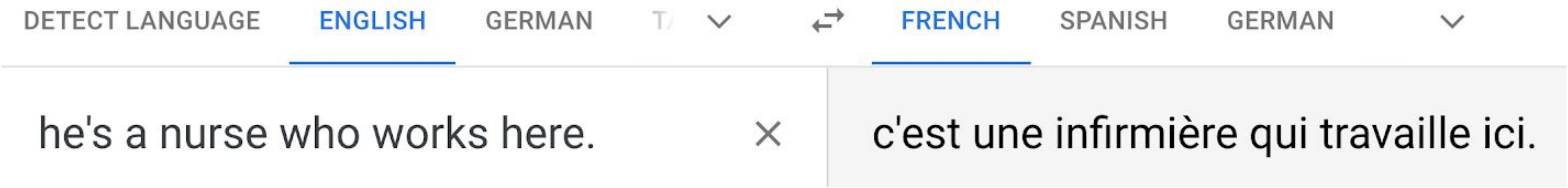
- Understanding the Problem
  - NLP Gone Wrong
  - **Sources of Bias**
  - Bias Measurement
- Addressing Bias
  - Bias Mitigation
  - The Effects of Interventions
- Beyond Bias

Bias in Machine Translation



Here is a doctor.
Here is a nurse.

Aquí hay un doctor.
Aquí hay una enfermera.

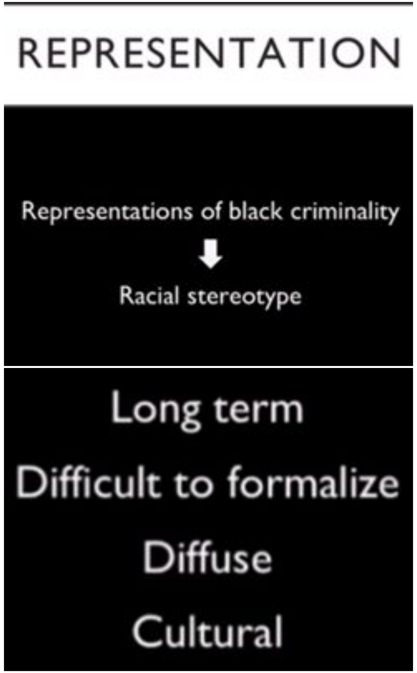


he's a nurse who works here.

c'est une infirmière qui travaille ici.

Types of Harm caused by AI (Crawford, 2017)

- Allocational harm: System performs worse on a group
- Representational harm: System perpetuates stereotypes about a group



REPRESENTATION

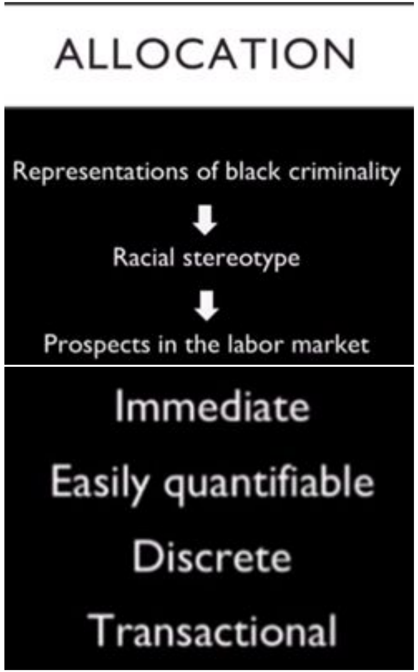
Representations of black criminality
↓
Racial stereotype

Long term

Difficult to formalize

Diffuse

Cultural



ALLOCATION

Representations of black criminality
↓
Racial stereotype
↓
Prospects in the labor market

Immediate

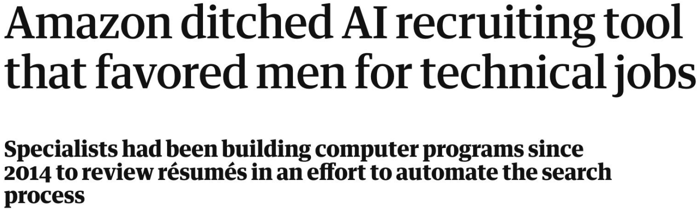
Easily quantifiable

Discrete

Transactional

Allocational harm

- Stereotype-based biases worsen model performance for groups already facing discrimination



Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

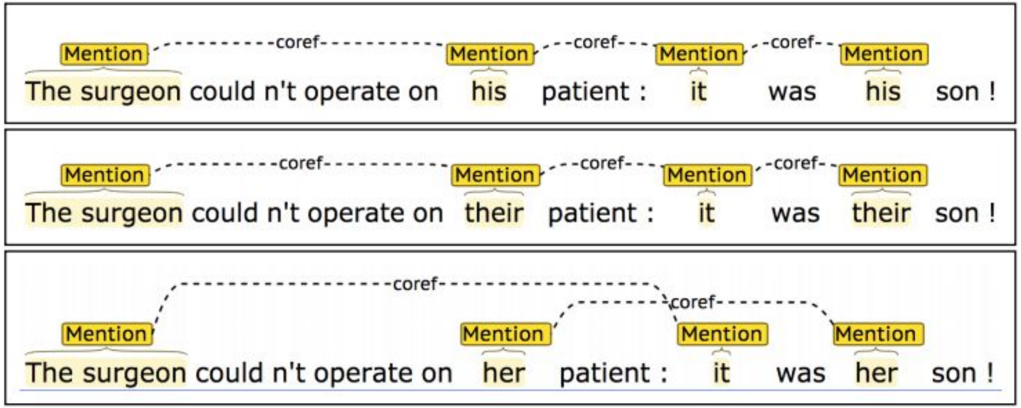
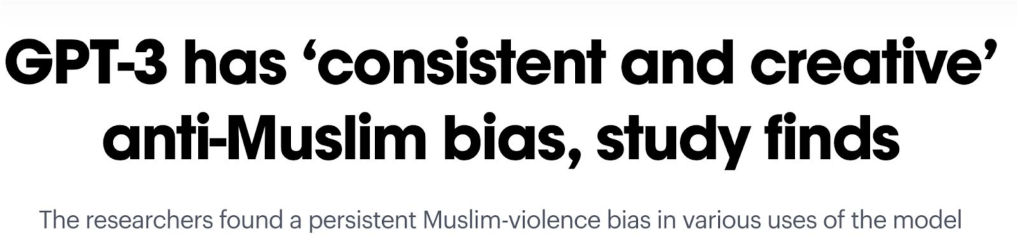

Mention
coref
Mention
coref
Mention
coref
Mention
The surgeon could n't operate on his patient : it was his son !
Mention
coref
Mention
coref
Mention
coref
Mention
The surgeon could n't operate on their patient : it was their son !
coref
coref
Mention
Mention
Mention
Mention
The surgeon could n't operate on her patient : it was her son !

Figure 1: Stanford CoreNLP rule-based coreference system resolves a male and neutral pronoun as coreferent with "The surgeon," but does not for the corresponding female pronoun.

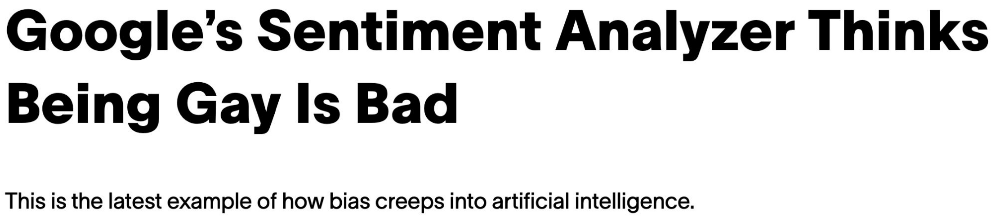
Representational harm

- Models represent groups in ways that perpetuate stereotypes



GPT-3 has ‘consistent and creative’ anti-Muslim bias, study finds

The researchers found a persistent Muslim-violence bias in various uses of the model



Google’s Sentiment Analyzer Thinks Being Gay Is Bad

This is the latest example of how bias creeps into artificial intelligence.

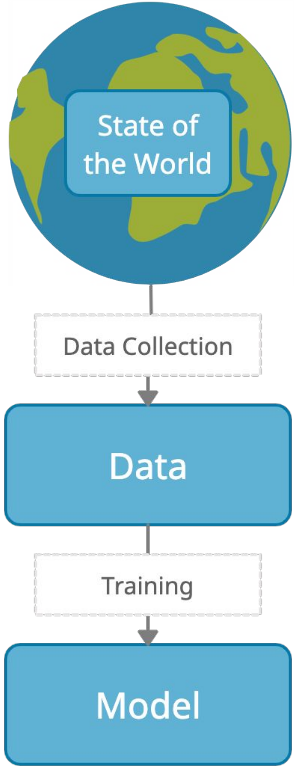
Evidence of Bias

- Gender & racial bias in translation and word embeddings (Caliskan et al., 2017)
- Gender bias:
  - Sentence encoding (May et al., 2019)
  - Image captioning (Zhao et al., 2017)
  - Coreference resolution (Rudinger et al., 2018)
- Islamophobia in large language modeling (Abid et al., 2021)
- Racial bias in hate speech detection (Sap et al., 2019)

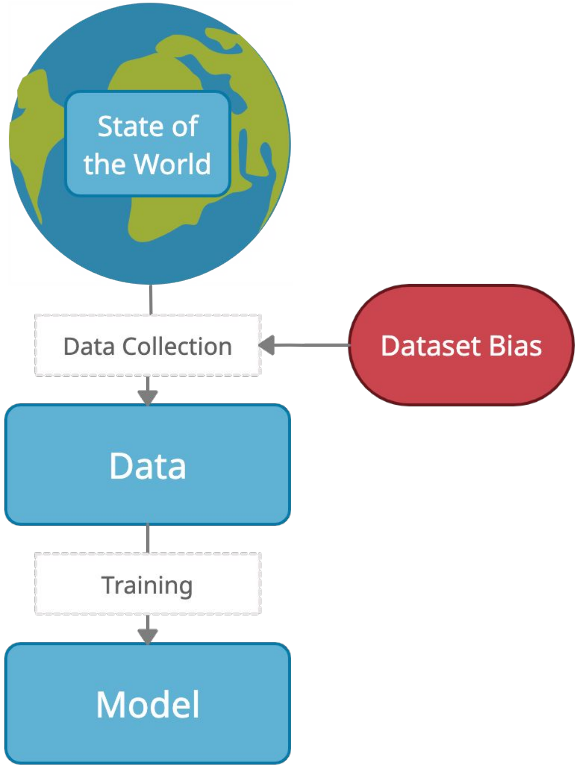
Outline

- Understanding the Problem
    - NLP Gone Wrong
    - **Sources of Bias**
    - Bias Measurement
- Addressing Bias
    - Bias Mitigation
    - The Effects of Interventions
- Beyond Bias

What Causes these Problems?



What Causes these Problems?



We're using AI instead of biased humans
What did you train the AI on?
What did you train the AI on?
imgflip.com

Dataset Issues: Collecting Data

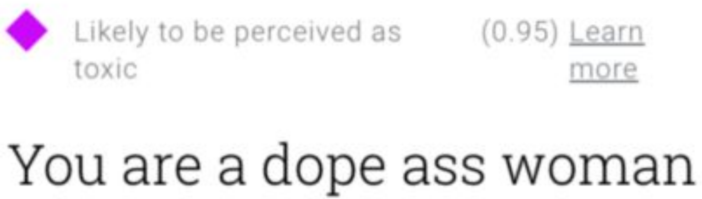
- Newer, larger models require large amounts of data
- NLP corpora are often scraped from uncurated web text
  - What text is on the web that we might want a dataset to exclude?

Dataset Issues: Collecting Data

- Newer, larger models require large amounts of data
- NLP corpora are often scraped from uncurated web text
    - What text is on the web that we might want a dataset to exclude?
        - Hate speech, stereotypical language
        - Spam
        - Adult content
        - Machine-generated text
    - Careful: filters for excluding this content can be “biased,” too!

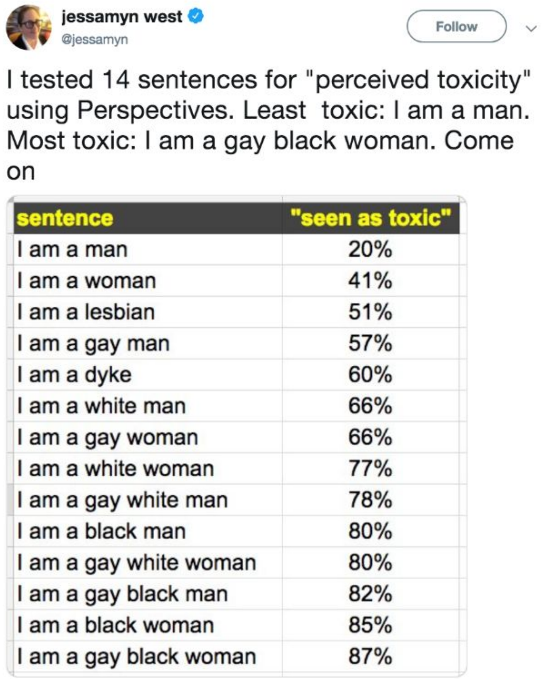
Dataset Issues: Filtering Data

- Careful: filters for excluding this content can be “biased,” too!
- Blocklists disproportionately exclude text by minority groups (Dodge et al., 2021)
- Hate speech/offensive speech detectors penalize identity terms and African-American English (Sap et al., 2019; Park & Fung, 2017)



Likely to be perceived as toxic (0.95) Learn more

You are a dope ass woman



jessamyn west
@jessamyn

I tested 14 sentences for "perceived toxicity" using Perspectives. Least toxic: I am a man. Most toxic: I am a gay black woman. Come on

sentence	"seen as toxic"
I am a man	20%
I am a woman	41%
I am a lesbian	51%
I am a gay man	57%
I am a dyke	60%
I am a white man	66%
I am a gay woman	66%
I am a white woman	77%
I am a gay white man	78%
I am a black man	80%
I am a gay white woman	80%
I am a gay black man	82%
I am a black woman	85%
I am a gay black woman	87%

Dataset Issues: Collecting Data

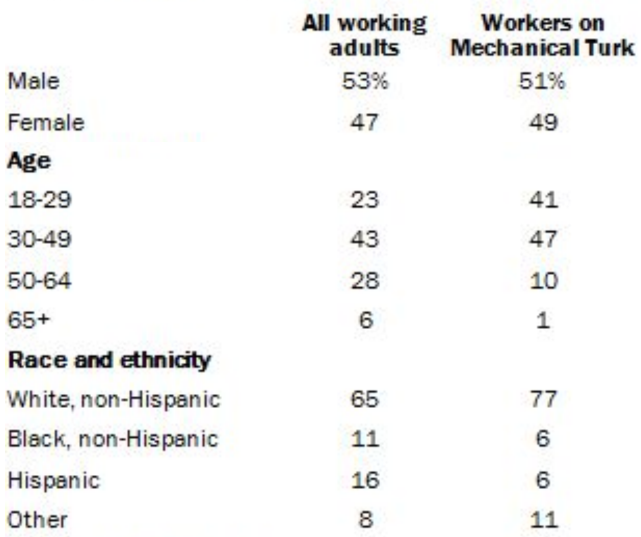
- What text *isn’t* as common on the web that we might want a dataset to include?

Dataset Issues: Collecting Data

- What text *isn’t* as common on the web that we might want a dataset to include?
    - Low-resource languages
    - Dialects with fewer speakers (e.g., African-American English)
    - Non-written languages
    - Older people’s language
    - Text by people without Internet access (often dependent on socioeconomic status & country where located)
- Oftentimes, people already facing disadvantages are further marginalized in datasets

Dataset Issues: Annotating and Filtering Data

- NLP corpora are often annotated by inexperienced labelers on platforms like Amazon Mechanical Turk
- Who annotates on Mechanical Turk?
  - Disproportionately white and young
  - Turkers from different countries may not be informed about relevant local issues in other countries
- Dataset quality measures can further suppress minority voices

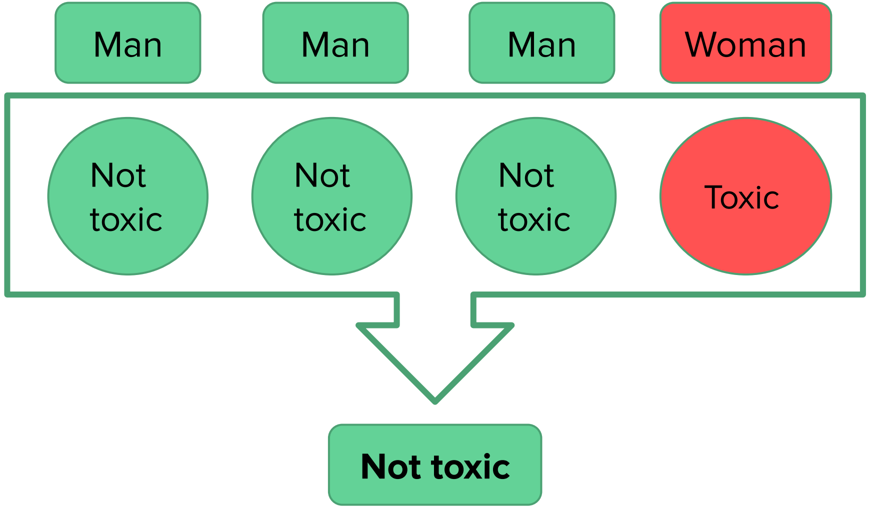


	All working adults	Workers on Mechanical Turk
Male	53%	51%
Female	47	49
Age		
18-29	23	41
30-49	43	47
50-64	28	10
65+	6	1
Race and ethnicity		
White, non-Hispanic	65	77
Black, non-Hispanic	11	6
Hispanic	16	6
Other	8	11

Dataset Issues: Annotating and Filtering Data

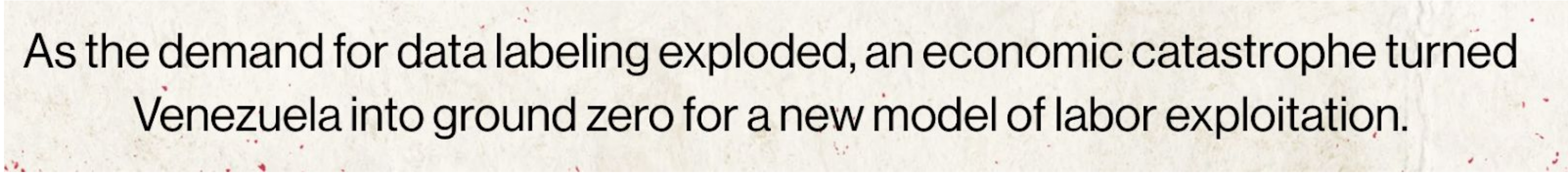
Is this sentence toxic?

“I’m not sexist, but a Ferrari just isn’t the sort of car that a woman should drive.”



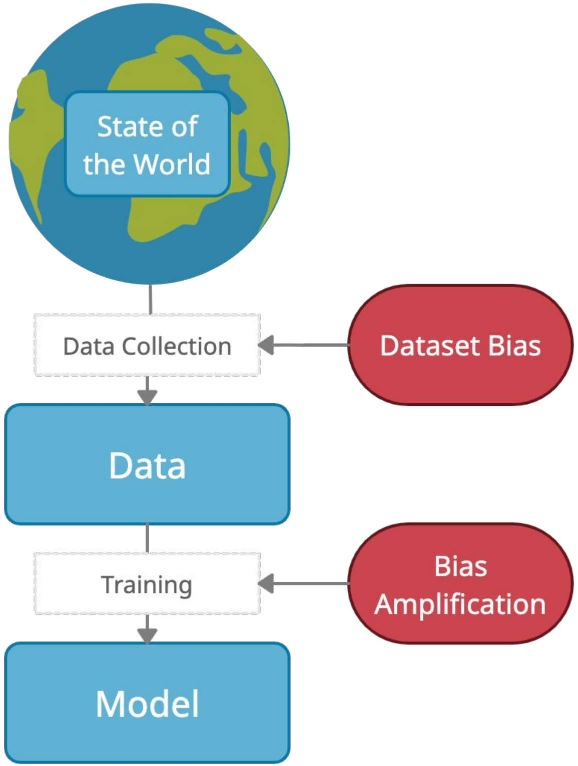
Dataset Issues: Beyond Bias

- Data labelers are disproportionately low-income and not always adequately compensated for their work
- For some tasks, data labelers are increasingly drawn from countries that permit lower pay or worse working conditions (Perrigo, 2022; Hao & Hernandez, 2022)
- Ensure your labelers get paid enough and question where your data comes from!



As the demand for data labeling exploded, an economic catastrophe turned Venezuela into ground zero for a new model of labor exploitation.

What Causes these Problems?



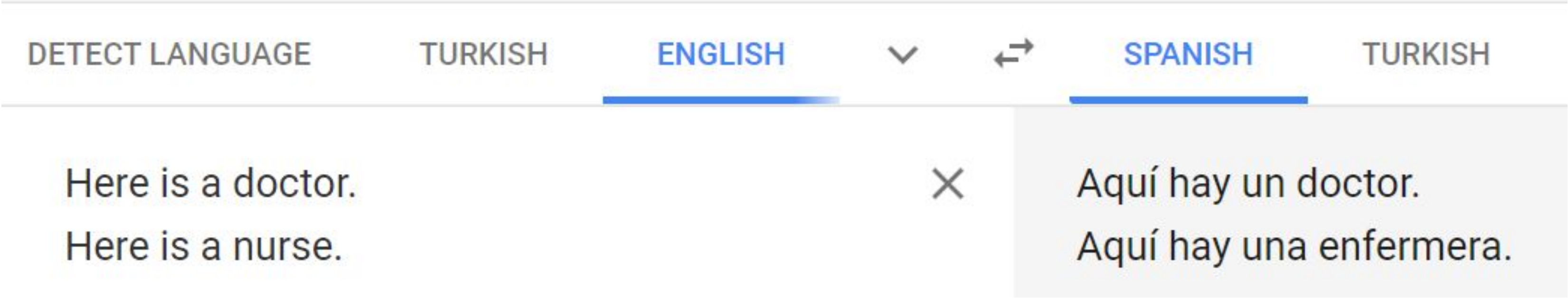
- Combination of **dataset bias** and **bias amplification** results in highly biased output

Compounding Sources of Bias: Coreference Resolution

- Bureau of Labor Statistics: 39% of managers are female
- Corpus used for coreference resolution training: 5% of managers are female
- Coreference systems: No managers predicted female (Rudinger et al., 2018)
- Systems overgeneralize gender

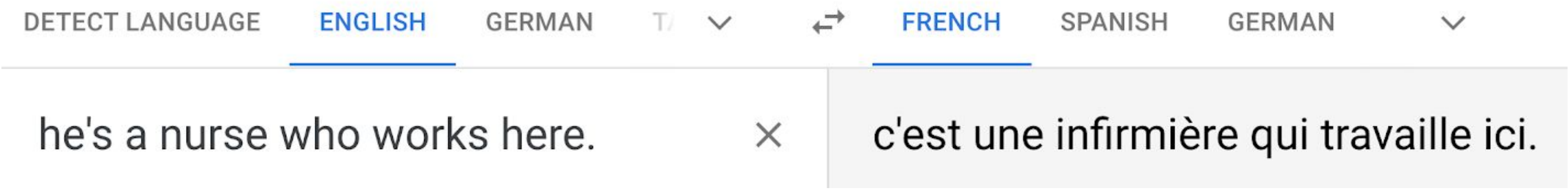
Bias in Machine Translation

- Dataset bias + bias amplification => stereotypically gendered translations



Here is a doctor.
Here is a nurse.

Aquí hay un doctor.
Aquí hay una enfermera.



he's a nurse who works here.

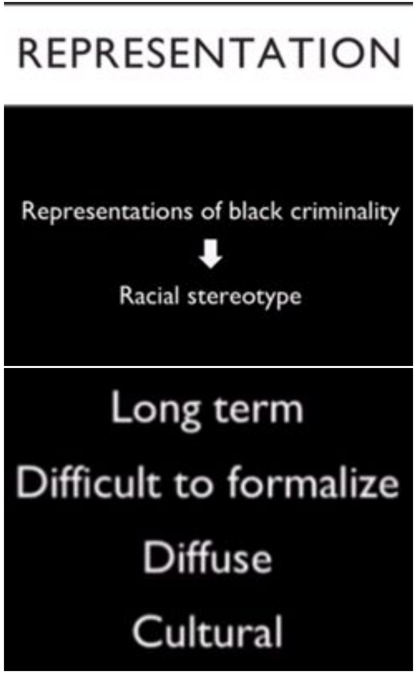
c'est une infirmière qui travaille ici.

Outline

- Understanding the Problem
    - NLP Gone Wrong
    - Sources of Bias
    - **Bias Measurement**
- Addressing Bias
    - Bias Mitigation
    - The Effects of Interventions
- Beyond Bias

Types of AI Harm (Crawford, 2017)

- Representational harm is harder to measure, but very common in language tasks



REPRESENTATION

Representations of black criminality

⬇

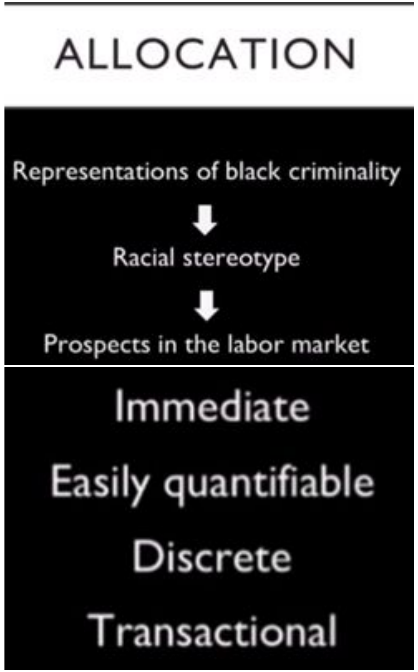
Racial stereotype

Long term

Difficult to formalize

Diffuse

Cultural



ALLOCATION

Representations of black criminality

⬇

Racial stereotype

⬇

Prospects in the labor market

Immediate

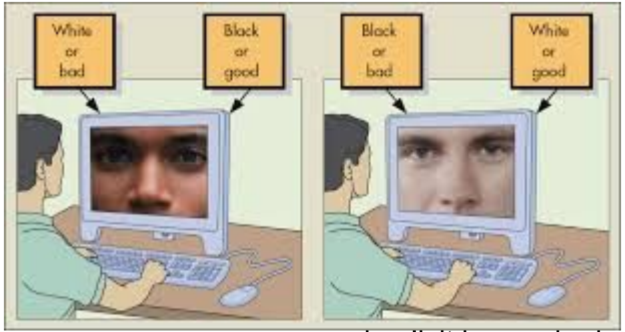
Easily quantifiable

Discrete

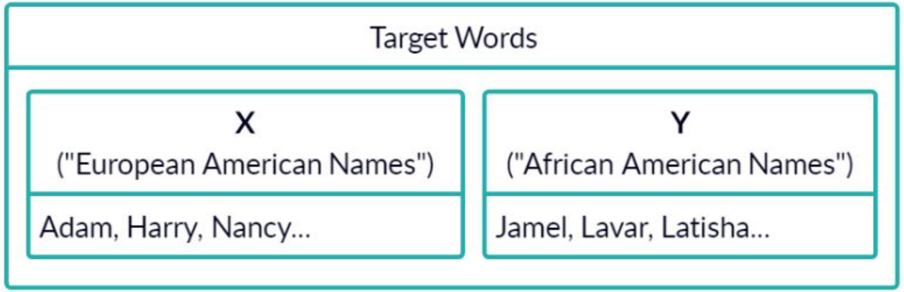
Transactional

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

- Measure bias in word embeddings (GloVe and word2vec)
- Based on Implicit Association Test
- Measure association between **target words** and **attribute words**

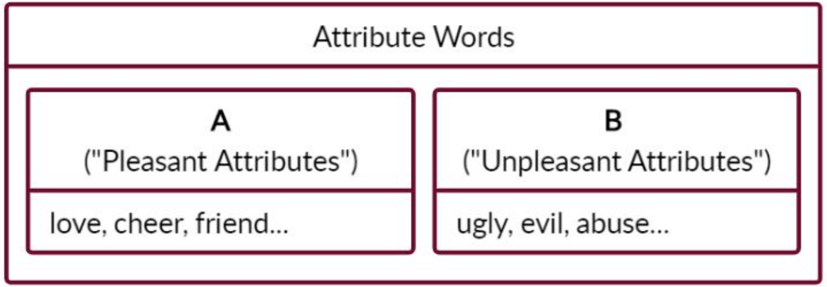


implicit.harvard.edu



Target Words

X ("European American Names")	Y ("African American Names")
Adam, Harry, Nancy...	Jamel, Lavar, Latisha...

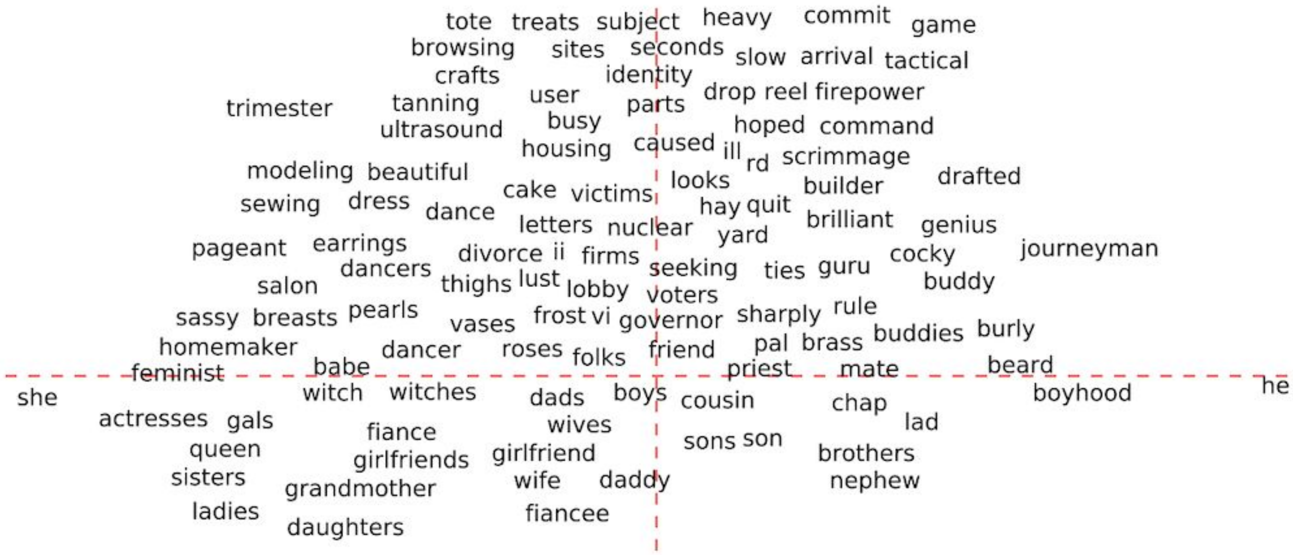


Attribute Words

A ("Pleasant Attributes")	B ("Unpleasant Attributes")
love, cheer, friend...	ugly, evil, abuse...

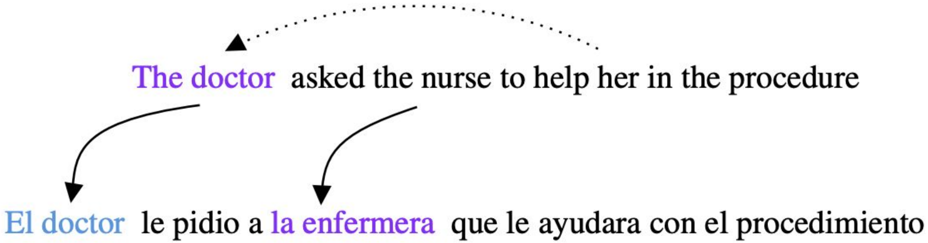
Quantifying gender bias

- Bolukbasi et al. (2016): bias in word embeddings
- Introduce idea of **gender direction**



Evaluating Bias in Language: Challenge Datasets

- Challenge datasets for bias in coreference resolution, machine translation, sentiment analysis
    - WinoMT, WinoGender, Equity Evaluation Corpus
    - E.g., sentences balanced between male/female genders and male/female role assignment
    - Measure difference in accuracy between sentences involving male/female genders or stereotypical and anti-stereotypical role assignment



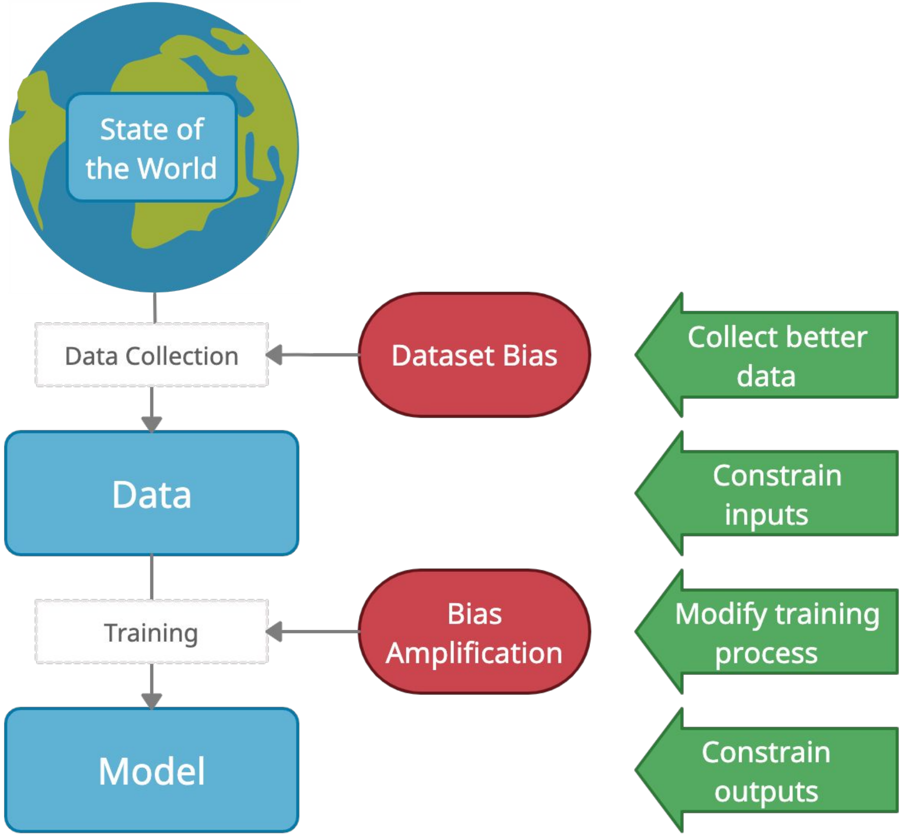
The doctor asked the nurse to help her in the procedure

El doctor le pidio a la enfermera que le ayudara con el procedimiento

The NLP Perspective

- Understanding the Problem
  - NLP Gone Wrong
  - Sources of Bias
  - Bias Measurement
- **Addressing Bias**
  - **Bias Mitigation**
  - The Effects of Interventions
- Beyond Bias

Bias Mitigation



Bias Mitigation: Improving Data Collection

- Tag protected attributes in corpora (Vanmassenhove et al., 2019)
- Fine-tune with a smaller, unbiased dataset (Saunders and Byrne, 2020)
- Pros: Often the most effective available method!
- Cons:
  - Data collection is costly and sometimes infeasible
  - How do you “balance” a dataset across many attributes?

Bias Mitigation: Constraining Inputs, Loss, or Outputs

- Constraining inputs
  - Adjust word embeddings (Bolukbasi et al., 2016)
- During training
  - Penalties, adversaries, or rewards (Zhang et al., 2017; Xia et al., 2019)
- Constraining outputs
  - Perturb model during decoding (He et al., 2021)

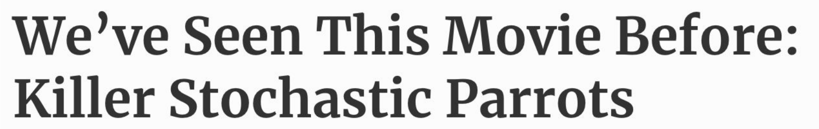
Effects of Bias Mitigation

- Some interventions aren’t very effective
  - “Lipstick on a Pig”: word embedding debiasing easily avoided
- But most are fairly effective…

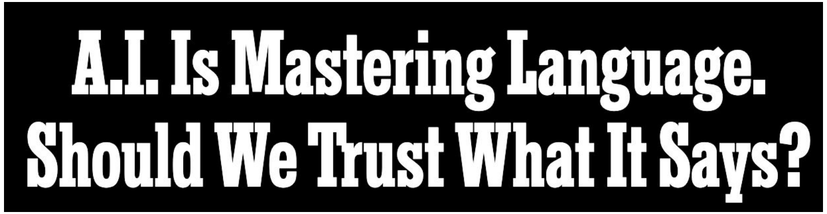
Effects of Bias Mitigation

- Some interventions aren’t very effective
  - “Lipstick on a Pig”: word embedding debiasing easily avoided
- But most are fairly effective…

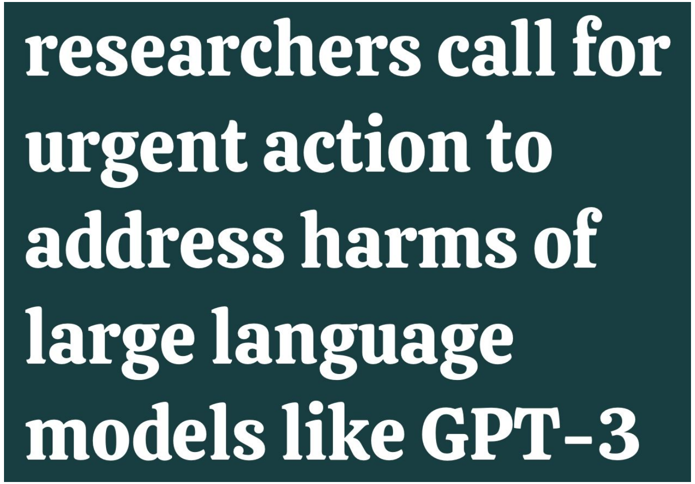
  …so why do models still cause harm?



We’ve Seen This Movie Before: Killer Stochastic Parrots

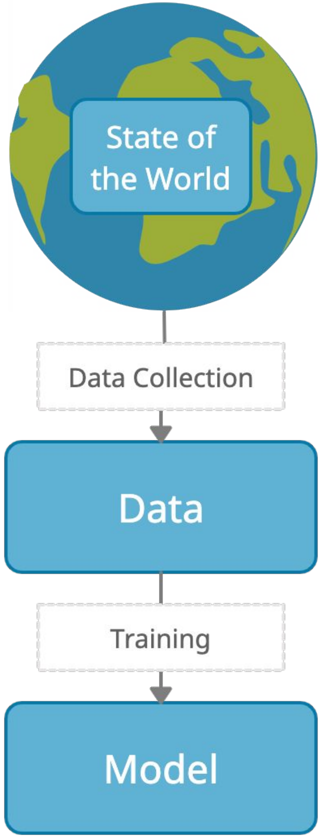


A.I. Is Mastering Language. Should We Trust What It Says?

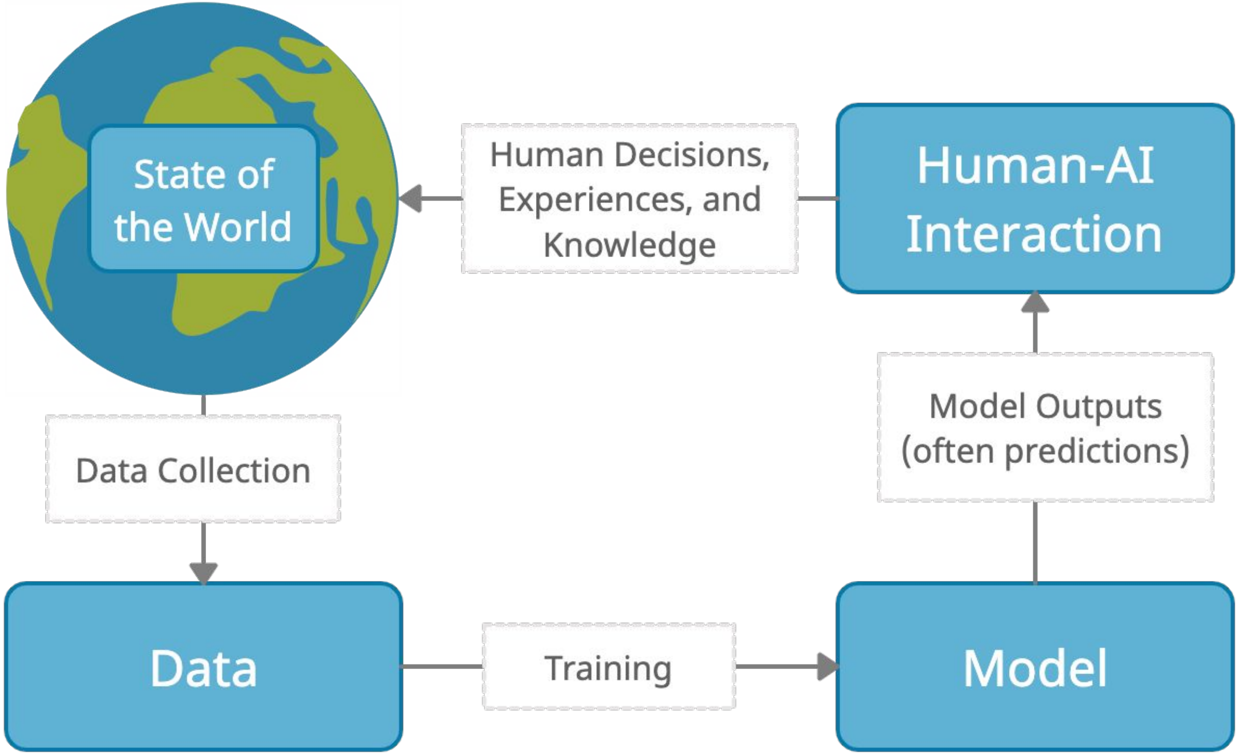


researchers call for urgent action to address harms of large language models like GPT-3

The Machine Learning Loop



The Machine Learning Loop


State of the World
Human Decisions, Experiences, and Knowledge
Human-AI Interaction
Model Outputs (often predictions)
Data Collection
Data
Training
Model

The Machine Learning Loop

- As computer scientists, it’s easy to assume that the answer always lies in changing the data or the algorithm
- But addressing deeper patterns of injustice requires adjusting interventions and the broader role of AI in society
- This requires the work of many–not just computer scientists
- In these efforts, algorithmic intervention isn’t the only role for computer science: we can also diagnose problems, formalize them, or work as part of broader interventions

Criticism of Bias Mitigation

- Language (Technology) is Power (Blodgett et al., 2019)
- Need to engage critically with what constitutes “bias”
  - Bias is inherently normative
  - Unstated assumptions about what systems should or shouldn’t do
    - These assumptions also reproduce harms
  - What makes a system’s behavior harmful?
- Some papers state system performance as the primary or only harm
- Research examines concerns from the dataset or model used, but rarely how the model is used in practice

Bias Mitigation

- Recommendations:
  - Ground work in the literature outside machine learning
    - HCI, sociology, linguistics
  - Explicitly lay out why system behaviors described as bias are harmful, how, and to whom
  - Work with people in affected communities to understand what they want and need
    - Change the balance of power

Beyond Bias: Assumptions and Oversimplification

- Language
  - Standard American English ≠ all language
  - Why prioritize languages with more speakers?
  - Does everyone speaks the same dialect of a language?
- What associations with gender/race/sexuality/etc. are “acceptable”?

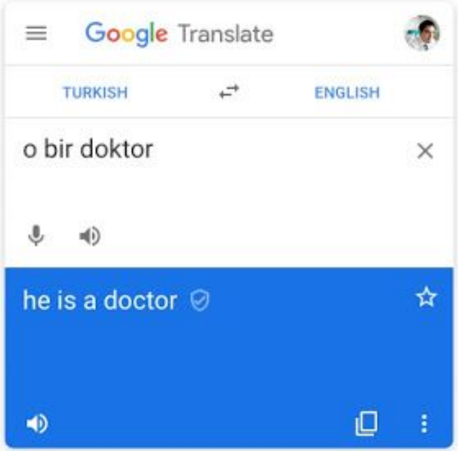
Complications in Bias Measurement and Evaluation

- Do existing bias measures cover all forms of discrimination?
    - Access
    - Intersectionality
    - Coverage
        - False negatives: misleading claims of fairness
    - Subtlety
        - Hate speech detection
    - Downstream effects
        - Machine learning loop
- How do we deal with more complex models?

Intervening outside the black box

- Technology as diagnostic
- Giving affected communities a voice
- User choice & human-AI interaction
- Change the problem, not the solution

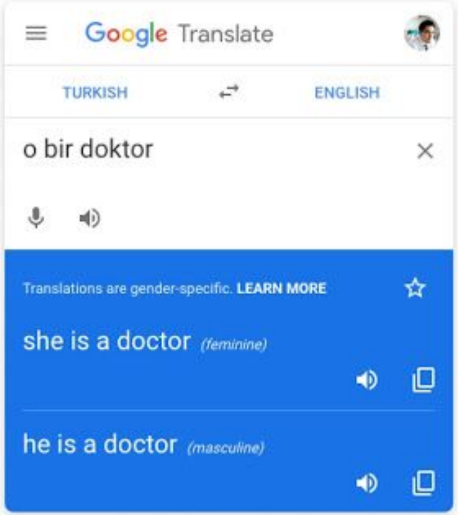
Before



o bir doktor

he is a doctor

After



o bir doktor

Translations are gender-specific. **LEARN MORE**

she is a doctor *(feminine)*

he is a doctor *(masculine)*

Thoughts for Discussion

- When should we intervene through algorithmic interventions vs. outside interventions?
- Do notions of “bias” cover all forms of harm?
  - What happens if our diagnostic tools allow for false negatives?
  - How would we capture or prevent more forms of harm?
- How do we incentivize researchers and companies to make less harmful models?