Computational Ethics

Psychological foundations of bias;
social bias and disparities in AI data and models.

An exercise

Which word is more likely to be used by a female ?

Word 1 – Word 2

Which word is more likely to be used by a female ?

Giggle – Laugh

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by a female ?

Giggle – Laugh

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by a female ?

Brutal – Fierce

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by a female ?

Brutal – Fierce

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by a older person ?

Impressive – Amazing

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by a older person ?

Impressive – Amazing

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by a person of higher occupational class ?

Suggestions – Proposals

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by a person of higher occupational class ?

Suggestions – Proposals

(Preotiuc-Pietro et al. ‘16)

Why do we intuitively recognize a default social group?

Implicit Bias

How Do We Make Decisions

System 1	System 2
automatic	effortful
fast	slow
parallel	serial
automatic	controlled
effortless	effort-filled
associative	rule-governed
slow-learning	flexible

Kahneman & Tversky 1973, 1974, 2002

Why?

How Do We Make Decisions

System 1

automatic

System 2

effortful

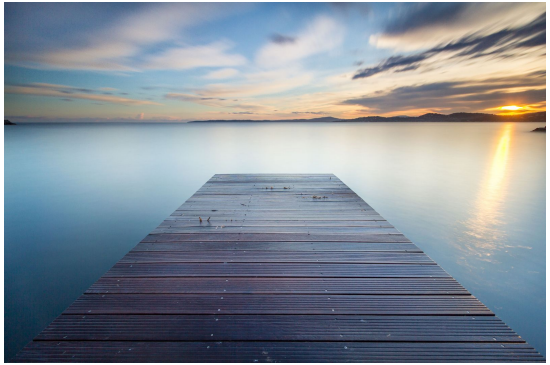
Our brains are evolutionarily hard-wired to store learned information for rapid retrieval and automatic judgments. Over 95% of cognition is relegated to the System 1 “auto-pilot.”

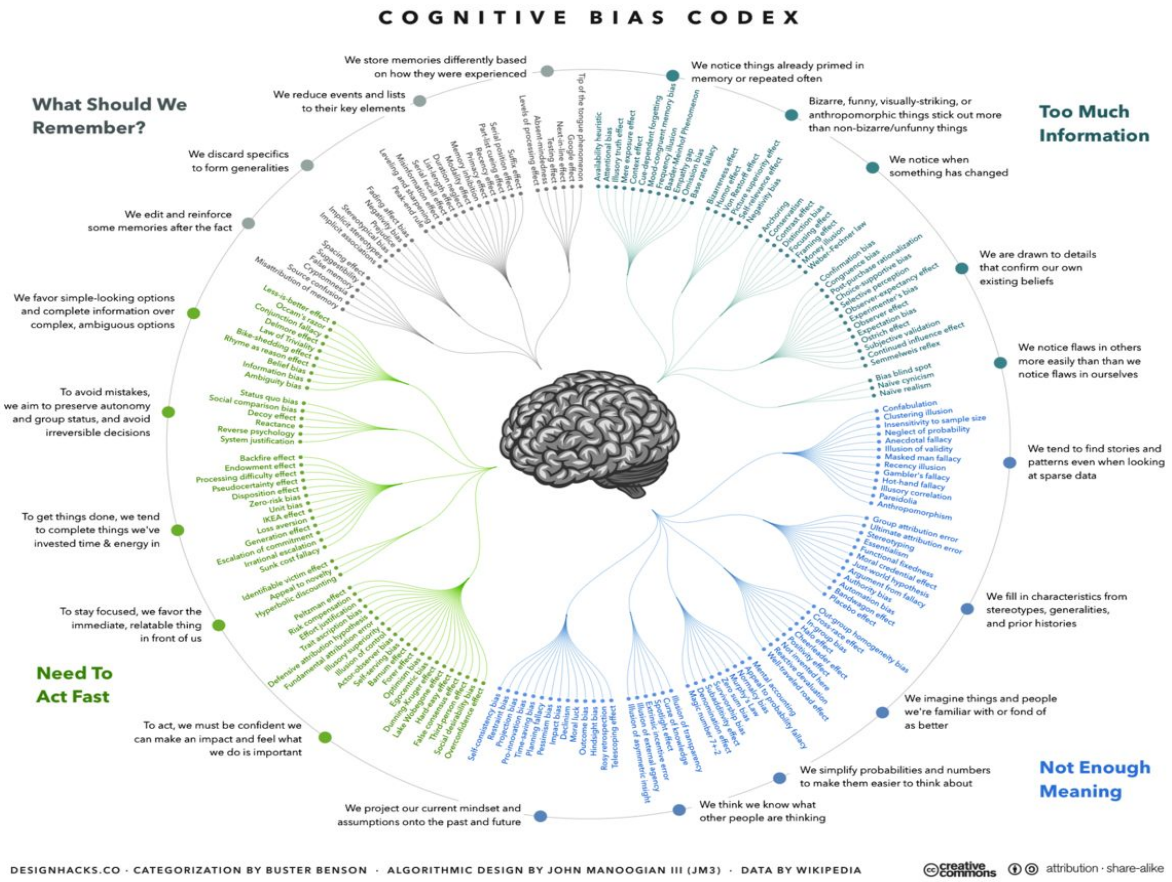
Psychological perspective on cognitive bias

Biases inevitably form because of the innate tendency of the human mind to:

- Categorize the world to simplify processing
- Store learned information in mental representations (called schemas)
- Automatically and unconsciously activate stored information whenever one encounters a category member

Cognitive bias is a systematic pattern of deviation from rationality in judgement





https://en.wikipedia.org/wiki/List_of_cognitive_biases

Examples of biases that affect how we make decisions

- **reporting bias**: selective revealing or suppression of information
- **selection bias**: selection does not reflect a random sample
- **confirmation bias**: paying more attention to information that reinforces previously held beliefs and ignoring evidence to the contrary
- **group attribution error**: when one generalizes about a group based on a group of representatives
- **correlation fallacy**: confusing correlation with causation
- **automation bias**: propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation
- etc.

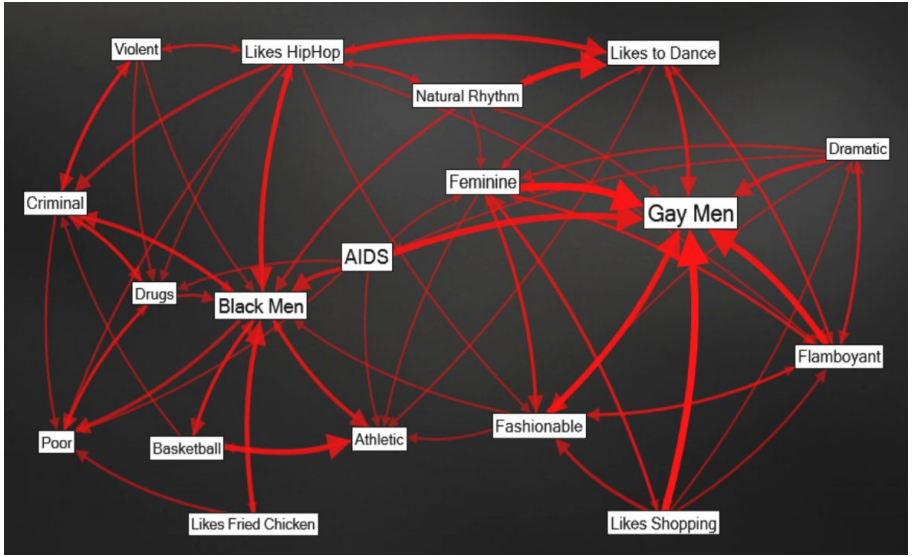
What happens when cognitive biases are applied to people?

Social stereotypes

- Gender
- Race
- Disability
- Age
- Sexual orientation
- Culture
- Class
- Poverty
- Language
- Religion
- National origin
- ...

Social stereotypes are similarly internalized as associations through natural processes of learning and categorization

Social biases are pervasive, unconscious, transitive



Implicit biases are pervasive, unconscious, and can automatically influence the ways in which we see and treat others, even when we are determined to be fair and objective.

image credit: Geoff Kaufman

Stereotypes and language

Language is a primary means through which stereotypes and prejudice are communicated and perpetuated

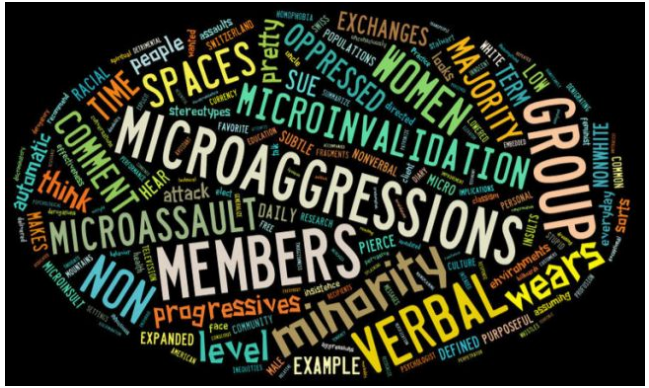
(Hamilton and Trolier, 1986; Bar-Tal et al., 2013)

Are social stereotypes necessarily negative?

- Gender
- Race
- Disability
- Age
- Sexual orientation
- Culture
- Class
- Poverty
- Language
- Religion
- National origin
- ...

How do implicit biases manifest?

Micro-inequities


MICROAGGRESSIONS
MEMBERS
GROUP
WOMEN
SPACES
MICROINVALIDATION
OPPRESSED
EXCHANGES
MICROASSAULT
NON
VERBAL
minority
progressives
level
wears
COMMENT
TIME
MAJORITY
people
pretty
SUE
think
automatic
EXAMPLE
EXPANDED
DEFINED
PURPOSEFUL
NONWHITE
everyday
sorts
attack
stereotypes
RACIAL
MAKES
HEAR
PIERCE
DAILY
SUBTLE
NONVERBAL
MICRO
WHITE
TERM
LOW
COMMON
SWITZERLAND
mosquito
HOMOPHOBIA
POPULATIONS
PERSONAL
RESEARCH
CULTURE
COMMUNITY
EDUCATION
FAVORITE
CURRENCY
MALE
face
assuming
MICROINSULT
MONTANA
AMERICA
PSYCHOLOGIST
INDUSTRIES
INDIVIDUALS
IMPLICATIONS
LIMITED
ENGINEERS
FRAGMENTS
Spiritual
HOMOSEXUAL
Stalwart
Leaders
Spread
Ground
Ignorance
Ignorant
Ignore
Ignored
Ignoring
Ignorance
Ignorant
Ignore
Ignored
Ignoring
Ignorance
Ignorant
Ignore
Ignored
Ignoring

Microaggressions

> "A comment or action that **subtly and often unconsciously or unintentionally** expresses a prejudiced attitude towards a member of a marginalized group"
>
> - Merriam Webster

Surface-level sentiment can be negative, neutral, or positive. For example:

- "Girls just **aren't good** at math."

microaggressions.com

tumblr.

Microaggressions

> "A comment or action that **subtly and often unconsciously or unintentionally** expresses a prejudiced attitude towards a member of a marginalized group"
>
> - Merriam Webster

Surface-level sentiment can be negative, neutral, or positive. For example:

- "Girls just **aren't good** at math."
- "Don't you people like tamales?"

microaggressions.com

tumblr.

Microaggressions

> "A comment or action that **subtly and often unconsciously or unintentionally** expresses a prejudiced attitude towards a member of a marginalized group"
>
> - Merriam Webster

Surface-level sentiment can be negative, neutral, or positive. For example:

- "Girls just aren't good at math."
- "Don't you people like tamales?"
- You're too pretty to be gay."

microaggressions.com

tumblr.

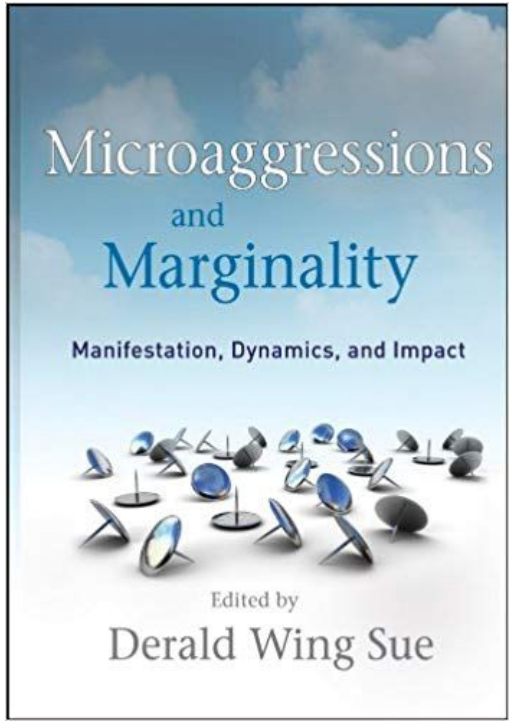
Examples of Disability Microaggressions in Everyday Life

Theme	Example	Message
Denial of personal identity: occurs when any aspect of a person's identity other than disability is ignored or denied	"Oh, you're married?"	There is no part of your life that is normal or like mine. The only thing I see when I look at you is your disability.
Helplessness: occurs when people frantically try to help PWDs	Someone helps you onto a bus or train, even when you need no help.	You can't do anything by yourself because you have a disability.
Infantilization: occurs when a PWD is treated like a child	"Let me do that for you."	You are not really capable. I know better.
Spread effect: occurs when other expectations about a person are assumed to be due to one specific disability	"Your other senses must be better than mine."	You're not normal.
Secondary gain: occurs when a person expects to feel good or be praised for doing something for a PWD	"We're going to raise enough money tonight to get Johnny that new wheelchair."	I feel good and get recognition for being nice to you

(Keller & Galgay, 2010)

Harmful impact of microaggressions

- Effects can be more pernicious than overtly aggressive speech (Sue et al. 2007, Sue 2010, Nadal et al. 2014)
- Can affect people's professional experiences and career trajectories (Cortina et al. 2002, Trix and Psenka 2003)
- Play on, and reinforce, problematic stereotypes and power structures (Hall and Braunwald 1981, Fournier et al. 2002)



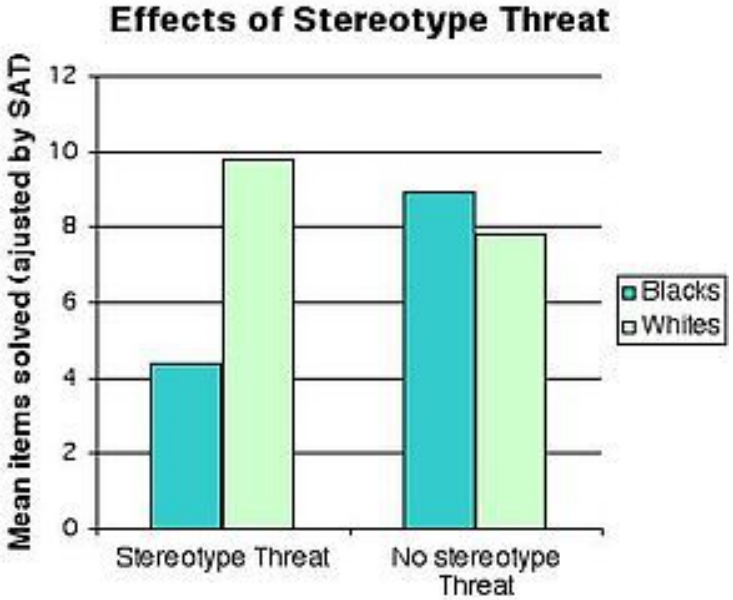
Stereotype threat

Fear of confirming a negative stereotype about one's group (Steele & Aronson, 1995)

- Often leads to anxiety and negative feelings that can use up mental resources and undermine one's confidence and ability to succeed
- Exacerbated by repeated experiences with microaggressions reducing one's sense of belonging or self-belief in a particular domain
  - e.g., women in STEM: Beasley & Fischer'12; Shapiro & Williams'12

Stereotype threat

- Groups: Blacks and Whites
- Threat: Intellectual ability



J. Aronson, C.M. Steele, M.F. Salinas, M.J. Lustina,
Readings About the Social Animal, 8th edition, ed. E. Aronson

Ways to Mitigate Biases

Devine's (1999) dissociation model

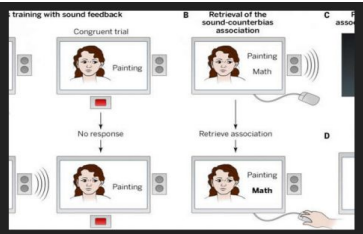
System 1: Stereotype Activation

- Stereotypes are firmly implanted (and reinforced) by learning and exposure, cognitive processes of categorization, etc.
- Thus, stereotypes are automatically activated whenever a cue is present, regardless of personal prejudice level; a "mental habit"

System 2: Preventing Stereotype Application

- Once a stereotype is activated, people can use System 2 processes to overcome the influence of the stereotype
- **Must first be aware of the activation of stereotypes**, then take steps to mitigate their impact or weaken their power…

Implicit bias is malleable

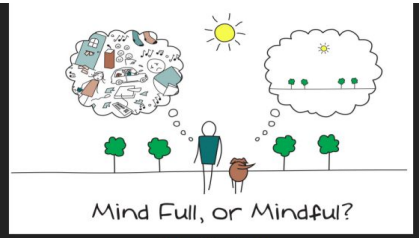


Counter-stereotypic Training:
Deliberately and repeatedly negating stereotypes or associating individuals with counter-stereotypic traits or attributes

Blair et al. (2001); Kang et al. (2012); Kawakami et al. (2000); Wittenbrink, Judd, & Park (2001)

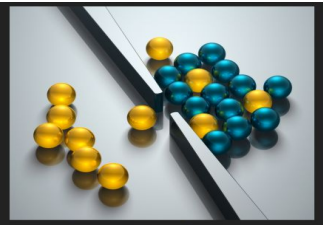
Mindset Training:
Cultivating a deliberative mindset, reminding oneself of egalitarian goals, reinforcing curiosity and constructive uncertainty about others

Beattie et al. (2013); Sassenberg & Moskowitz (2005); Stone & Moskowitz (2011)



Meditation:
Mindfulness meditation and "loving-kindness" meditation training have been shown to reduce outgroup biases

Kang et al. (2014); Lueke & Gibson (2015)



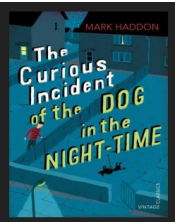
Intergroup Contact:
Requires equal status and common goals, a cooperative environment with frequent informal interactions, and the presence of support from authority figures or customs

Allport (1954); Peruche & Plant (2006); Pettigrew & Tropp (2006)

The party games *Awkward Moment* and *Awkward Moment at Work* present hypothetical occurrences of bias; players compete to submit the best responses to these "moments."

Research has shown the games increase players' understanding of the experience of bias and increase empathy and perspective-taking

Kaufman, Flanagan, & Seidman (2015, 2016)



Contact and Perspective-taking through Fiction:
Fictional worlds of stories, games, and VR environments have proven to be effective spaces for bias reduction

Kaufman & Libby (2012)

How to Recognize Implicit Bias?

Implicit Association Test (IAT)-Greenwald et al. 1998

Category	Items
Good	Spectacular, Appealing, Love, Triumph, Joyous, Fabulous, Excitement, Excellent
Bad	Angry, Disgust, Rotten, Selfish, Abuse, Dirty, Hatred, Ugly
African Americans	
European Americans	

Implicit Association Test

GOOD **BAD**

Love

Implicit Association Test

GOOD **BAD**

Hatred

Implicit Association Test

GOOD **BAD**

Spectacular

Implicit Association Test

African Americans

European Americans

Implicit Association Test

African Americans

European Americans

Implicit Association Test

African Americans or BAD

European Americans or GOOD

Spectacular

Implicit Association Test

African Americans or BAD

European Americans or GOOD

Implicit Association Test

African Americans or BAD

European Americans or GOOD

Implicit Association Test

African Americans or GOOD

European Americans or BAD

Appealing

Implicit Association Test

African Americans or GOOD

European Americans or BAD

Implicit Association Test

African Americans or GOOD

European Americans or BAD

Implicit Association Test

African Americans or GOOD

European Americans or BAD

Rotten

Implicit Association Test

The IAT involves making repeated judgments (by pressing a key on a keyboard) to label words or images that pertain to one of two categories presented simultaneously (e.g., categorizing pictures of African American or European American and categorizing positive/negative adjectives).

The test compares response times when different pairs of categories share a response key on keyboard
(e.g., African American + GOOD vs African American + BAD vs European American + GOOD vs European American + BAD)

IAT - Societal groups⇔Stereotype words

https://implicit.harvard.edu/implicit/selectatest.html

Disability IAT

Disability **('Disabled - Abled' IAT).** This IAT requires the ability to recognize symbols representing abled and disabled individuals.

Asian IAT

Asian American **('Asian - European American' IAT).** This IAT requires the ability to recognize White and Asian-American faces, and images of places that are either American or Foreign in origin.

Sexuality IAT

Sexuality **('Gay - Straight' IAT).** This IAT requires the ability to distinguish words and symbols representing gay and straight people. It often reveals an automatic preference for straight relative to gay people.

Arab-Muslim IAT

Arab-Muslim **('Arab Muslim - Other People' IAT).** This IAT requires the ability to distinguish names that are likely to belong to Arab-Muslims versus people of other nationalities or religions.

Age IAT

Age **('Young - Old' IAT).** This IAT requires the ability to distinguish old from young faces. This test often indicates that Americans have automatic preference for young over old.

Skin-tone IAT

Skin-tone **('Light Skin - Dark Skin' IAT).** This IAT requires the ability to recog
skinned faces. It often reveals an automatic preference for light-skin relative to da

Race IAT

Race **('Black - White' IAT).** This IAT requires the ability to distinguish faces of
African origin. It indicates that most Americans have an automatic preference for

Religion IAT

Religion **('Religions' IAT).** This IAT requires some familiarity with religious terms from various world religions.

Native IAT

Native American **('Native - White American' IAT).** This IAT requires the ability to recognize White and Native American faces in either classic or modern dress, and the names of places that are either American or Foreign in origin.

Gender-Science IAT

Gender - Science**.** This IAT often reveals a relative link between liberal arts and females and between science and males.

Gender-Career IAT

Gender - Career**.** This IAT often reveals a relative link between family and females and between career and males.

Presidents IAT

Presidents **('Presidential Popularity' IAT).** This IAT requires the ability to recognize photos of Donald Trump and one or more previous presidents.

Weight IAT

Weight **('Fat - Thin' IAT).** This IAT requires the ability to distinguish faces of people who are obese and people who are thin. It often reveals an automatic preference for thin people relative to fat people.

Weapons IAT

Weapons **('Weapons - Harmless Objects' IAT).** This IAT requires the ability to recognize White and Black faces, and images of weapons or harmless objects.

Back to AI

Back to AI

AI is only System 1

Terminology

Cognitive bias

⇔

Statistical bias in ML

⇔

Social biases in AI, in data, algorithms, and applications

Social Biases in
Data, Algorithms, Models, Applications

Image search

- June 2016: web search query “three black teenagers”

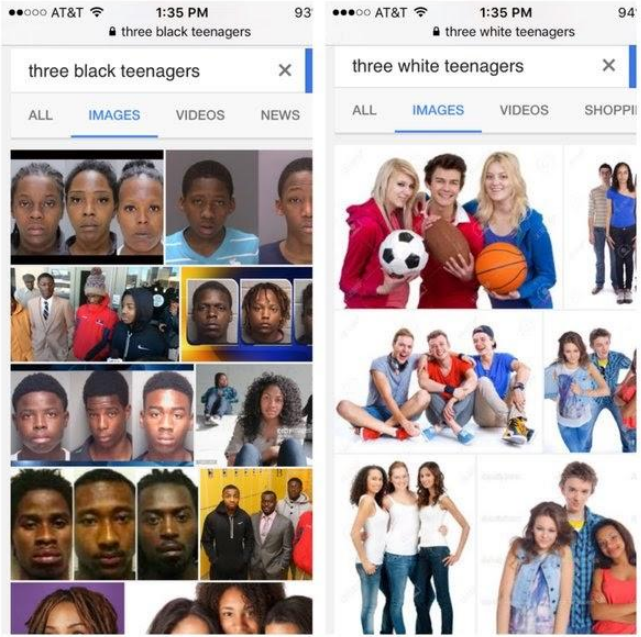


Image search

- June 2017: image search query “Doctor”

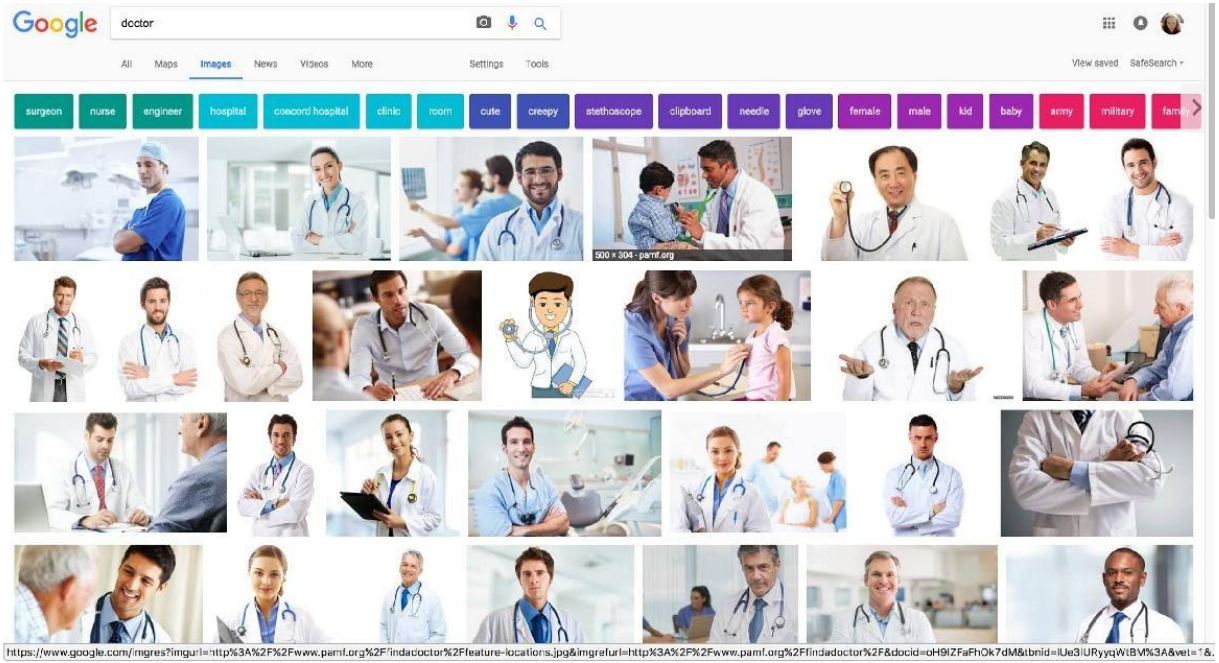


Image search

- June 2017: image search query “Nurse”

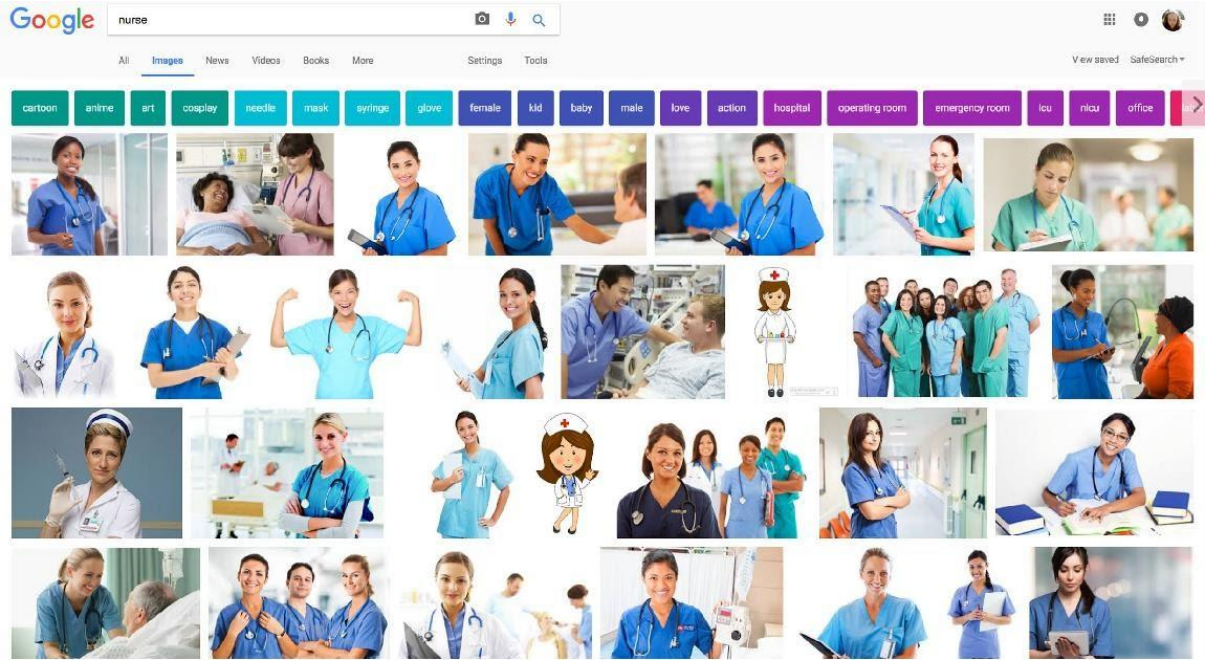


Image search

- June 2017: image search query “Homemaker”

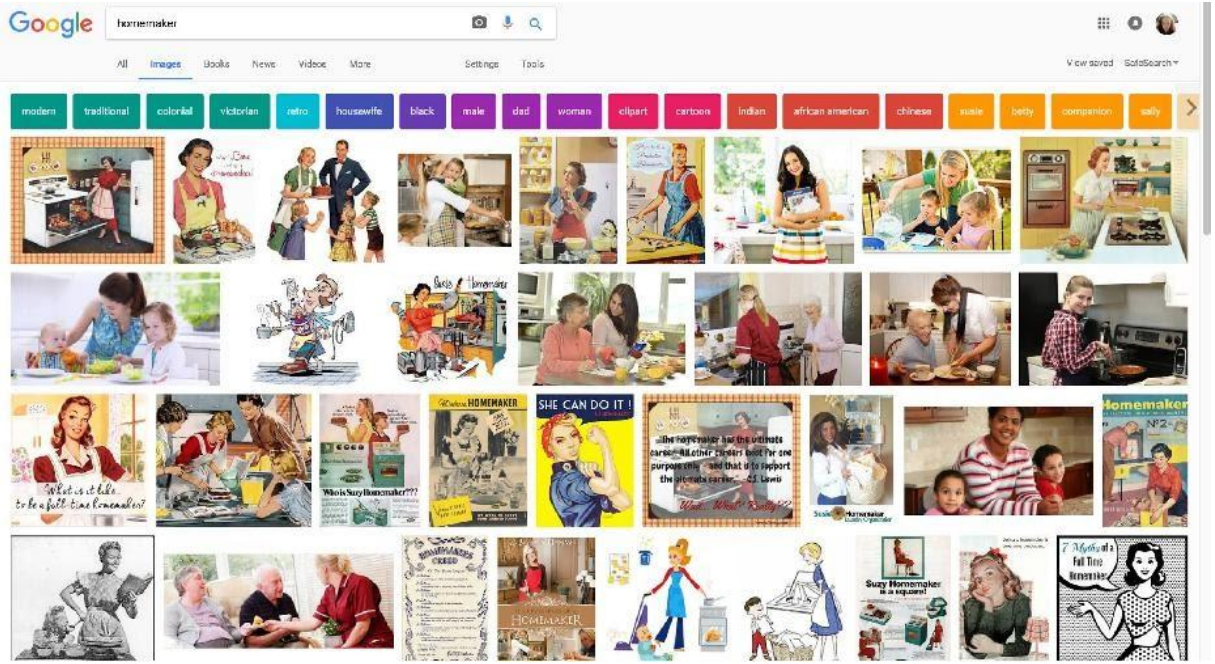


Image search

- June 2017: image search query “CEO”

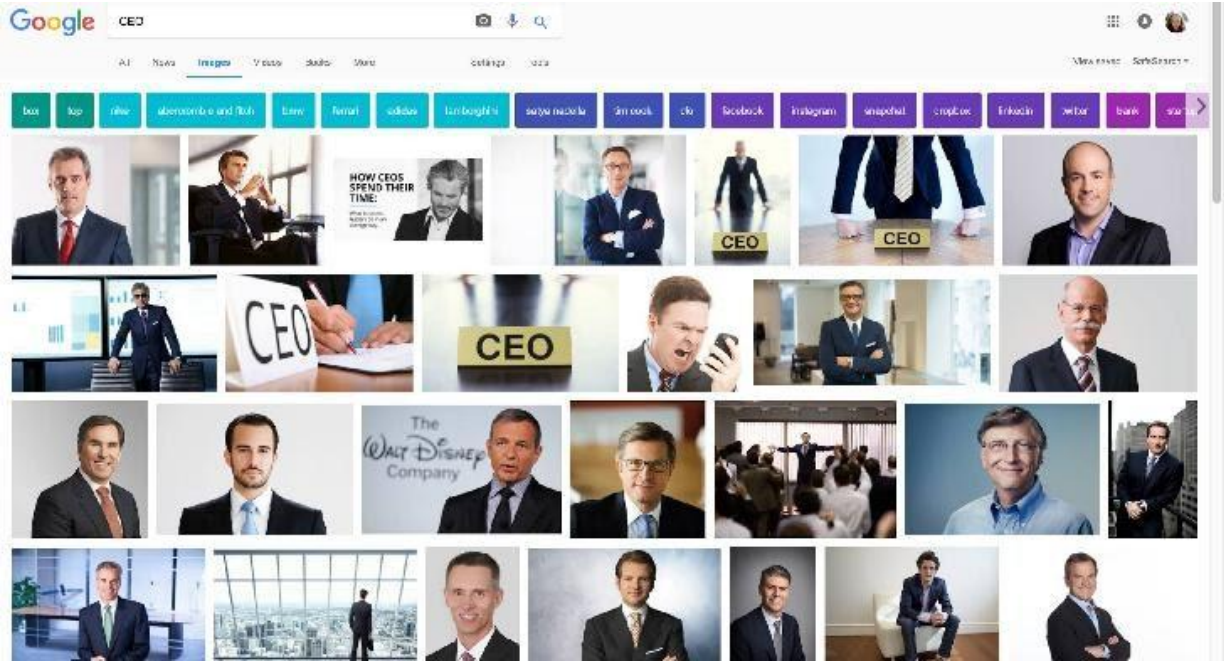
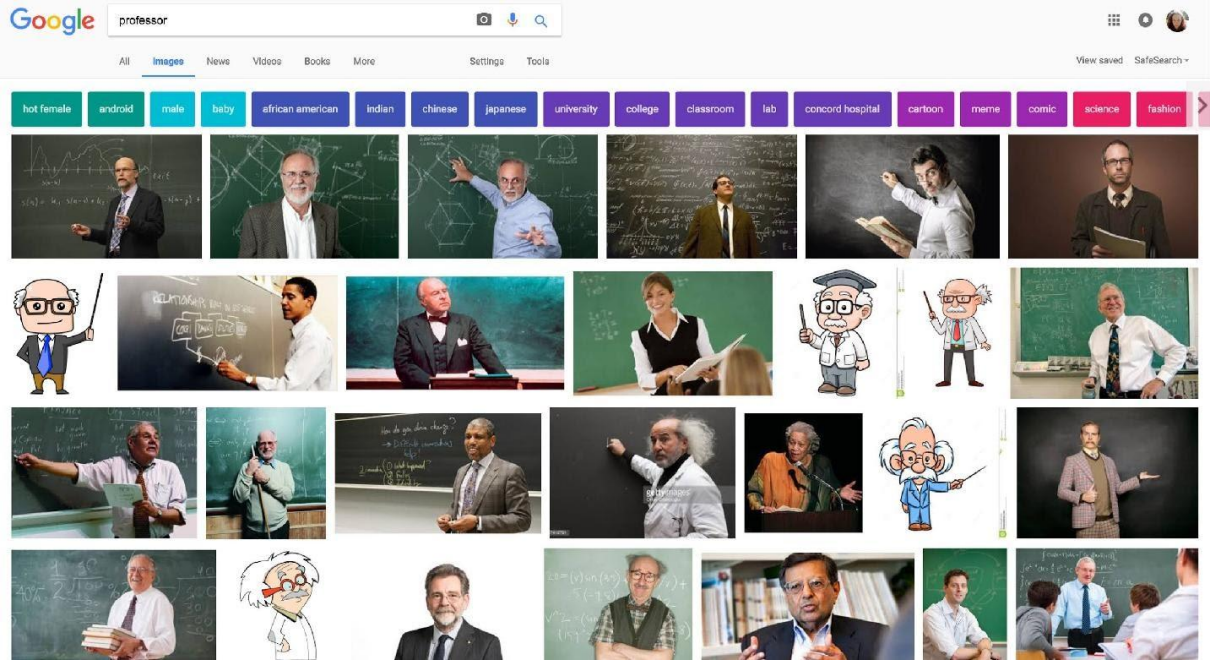


Image search

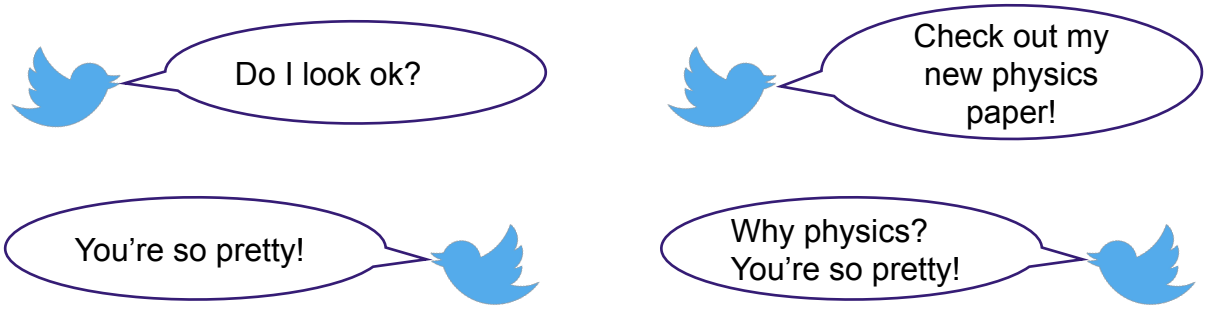
- June 2017: image search query “Professor”



Face recognition

Positive or negative?

Positive or negative?



Positive or negative?

Do I look ok?

You're so pretty!

Check out my new physics paper!

Why physics? You're so pretty!

Do I look ok?

You're so pretty for your age!

You're so pretty for a black girl!

You're too pretty to be gay!

- Conversational agents
- Personal assistants
- Search engines
- Translation engines
- Medical research assistants
- …

Gender bias on the web

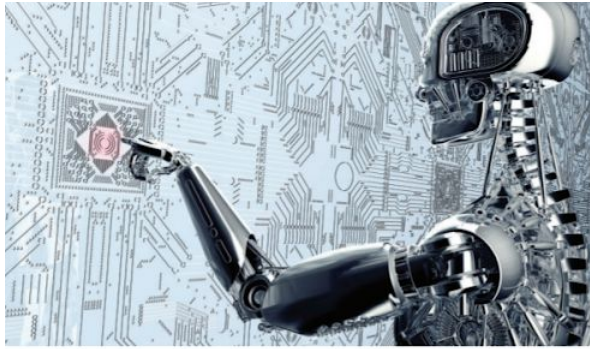
- The dominant class is often portrayed and perceived as relatively more professional (Kay, Matuszek, and Munson 2015)
- Males are over-represented in the reporting of web-based news articles (Jia, Lansdall-Welfare, and Cristianini 2015)
- Males are over-represented in twitter conversations (Garcia, Weber, and Garimella 2014)
- Biographical articles about women on Wikipedia disproportionately discuss romantic relationships or family-related issues (Wagner et al. 2015)
- IMDB reviews written by women are perceived as less useful (Otterbacher 2013)

Online data is riddled with **implicit biases**

Online data is riddled with **implicit biases**

Consequence: models are also biased

- employment matching
- flight routing
- automated legal aid in immigration
- advertisement placement
- search
- parole decisions
- chatbots
- face recognition
- voice recognition
- + dozens!



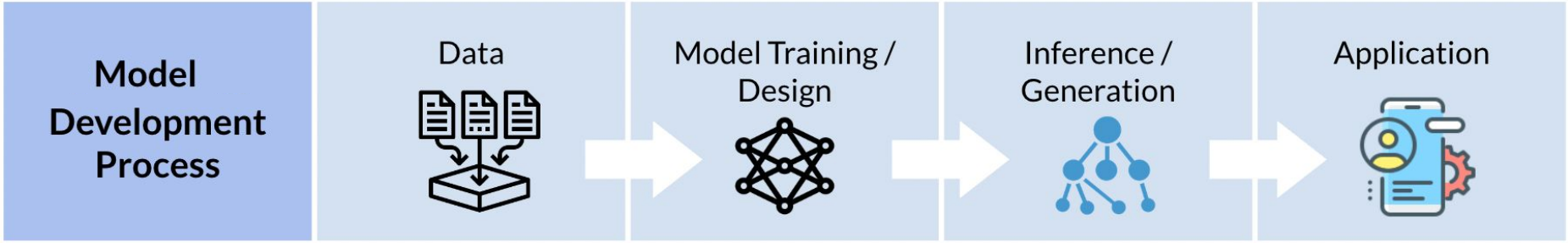
An Intelligence in Our Image

The Risks of Bias and Errors in Artificial Intelligence

Osonde Osoba, William Welser IV

RAND CORPORATION

Social bias in AI



https://tinyurl.com/BiasInAI-readings

drawing credit: Lucille Njoo

Example: misgendering in machine translation

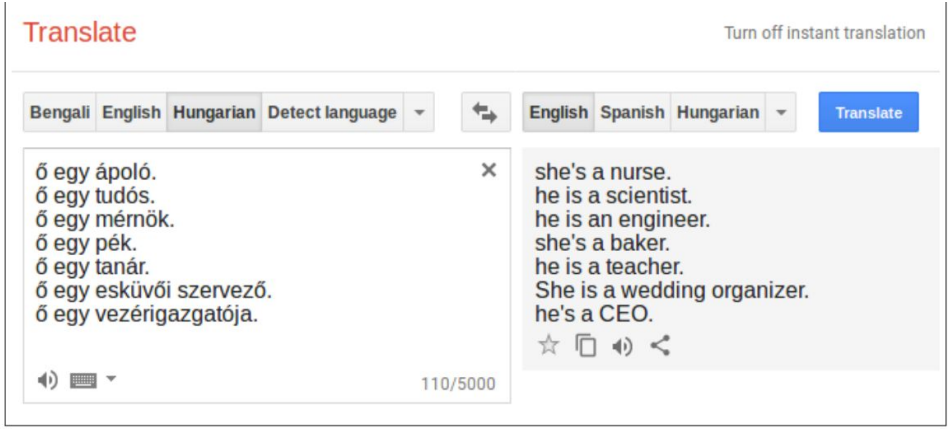
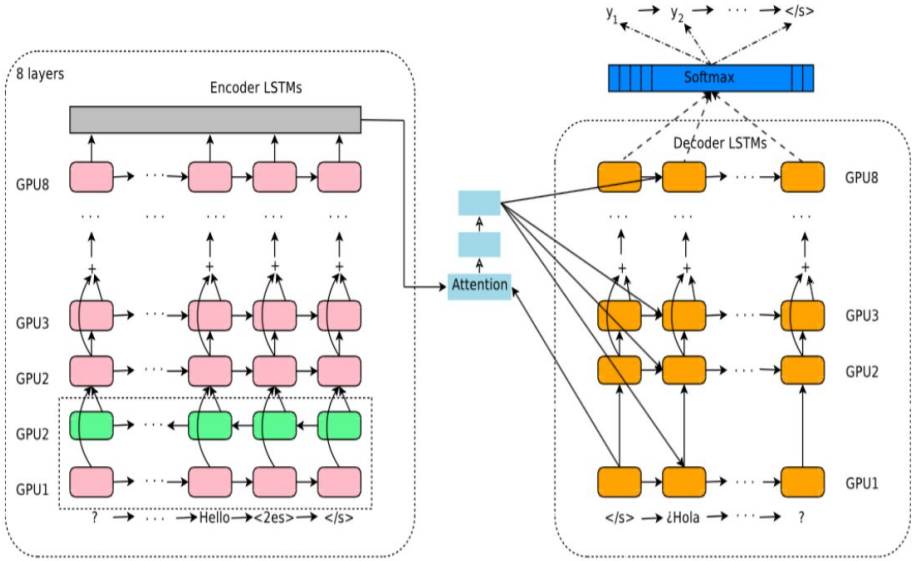


Figure 1: Translating sentences from a gender neutral language such as Hungarian to English provides a glimpse into the phenomenon of gender bias in machine translation. This screenshot from Google Translate shows how occupations from traditionally male-dominated fields [40] such as scholar, engineer and CEO are interpreted as male, while occupations such as nurse, baker and wedding organizer are interpreted as female.

Prates, et al. Assessing gender bias in machine translation: a case study with Google Translate. (2019) Neural Computing and Applications https://arxiv.org/pdf/1809.02208.pdf

Example: debiasing machine translation

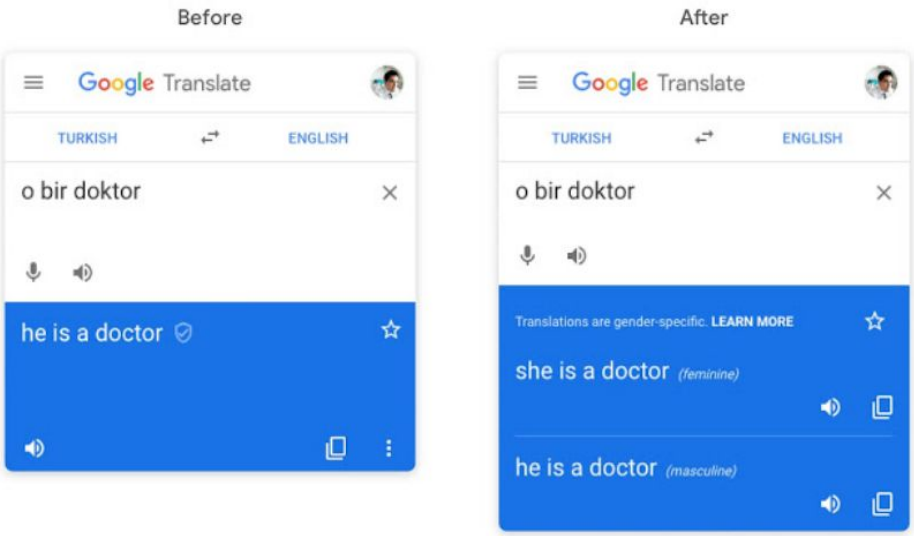
Possible solutions



<2female> <2es> Hello, how are you? -> ¿Hola como estás?

Example: debiasing machine translation

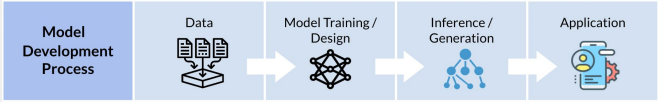
Not only modeling, but also UI solutions



https://ai.googleblog.com/2018/12/providing-gender-specific-translations.html

Where to look for “bias” in AI

Where to look for “bias” in AI



- Problem definition/research question (who will benefit? who can be harmed?)
- Data
  - Who is described (when some populations are excluded or underrepresented)
  - How training data might describe populations in biased ways
  - Who authored the data
- Data labels
  - Annotation schema (e.g., binary gender labels)
  - Annotation instructions
  - Annotator bias (e.g., toxicity annotations of SAE vs AAE)
- Model design
  - Biased objective (e.g., COMPAS system for parole decisions)
  - Spurious correlations (e.g., correlations of ethnicity and sentiment labels, gendered pronouns and professions in coreference links)
- Model outputs
  - Bias amplification
  - Disparities in model utility and fairness by populations

Bias in data: representation bias

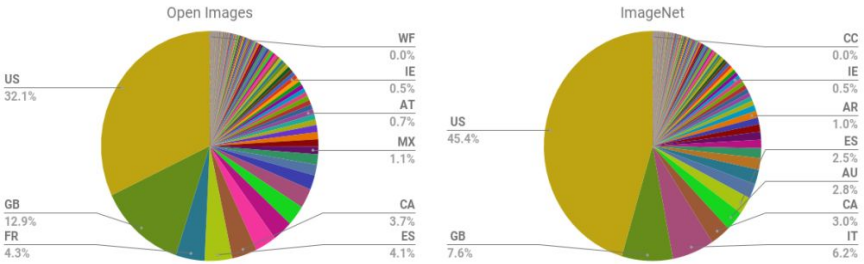


Figure 1: Fraction of Open Images and ImageNet images from each country. In both data sets, top represented locations include the US and Great Britain. Countries are represented by their two-letter ISO country codes.

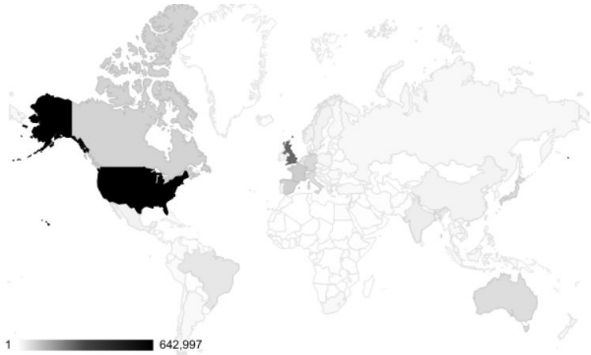


Figure 2: Distribution of the geographically identifiable images in the Open Images data set, by country. Almost a third of the data in our sample was US-based, and 60% of the data was from the six most represented countries across North America and Europe.

Shankar, et al. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World (2017) NeurIPS Workshop on Machine Learning for the Developing World https://arxiv.org/abs/1711.08536

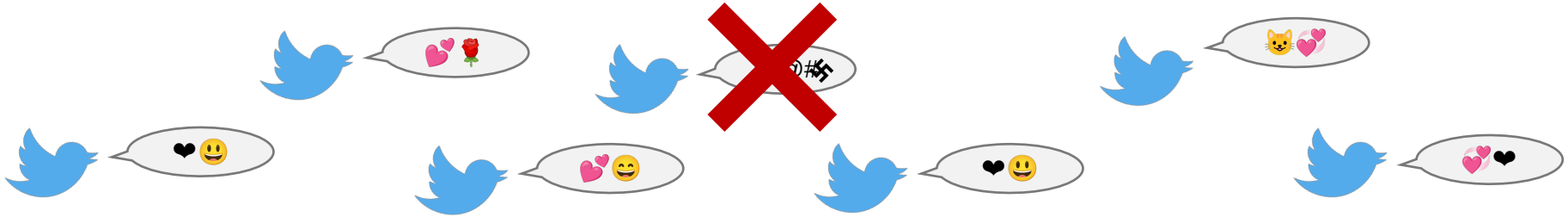
Bias in data leads to bias in models

Wedding photographs (donated by Googlers), labeled by a classifier trained on the Open Images dataset. The classifier's label predictions are recorded below each image.

https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

Bias in data: bias in annotations

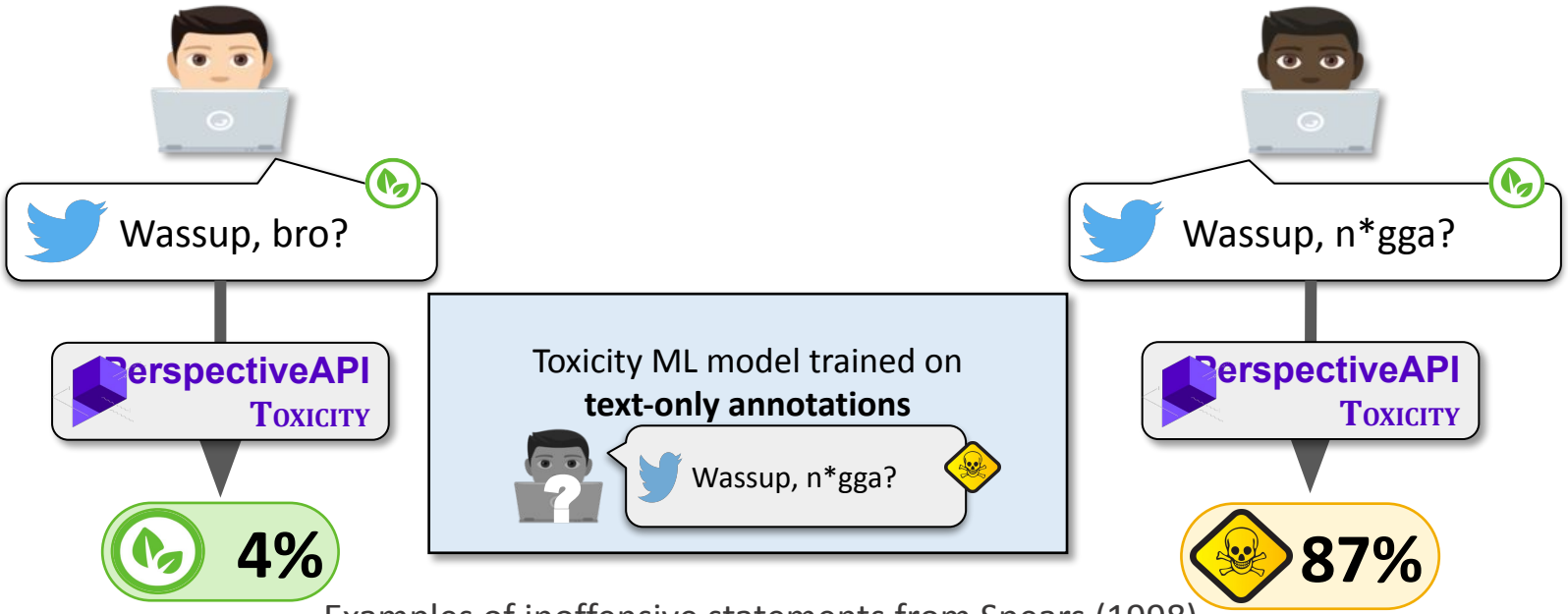
Task: hate speech detection



Slides on bias in hate speech by Maarten Sap

Sap, et al. The Risk of Racial Bias in Hate Speech Detection (2019) In *Proc. ACL* https://aclanthology.org/P19-1163/

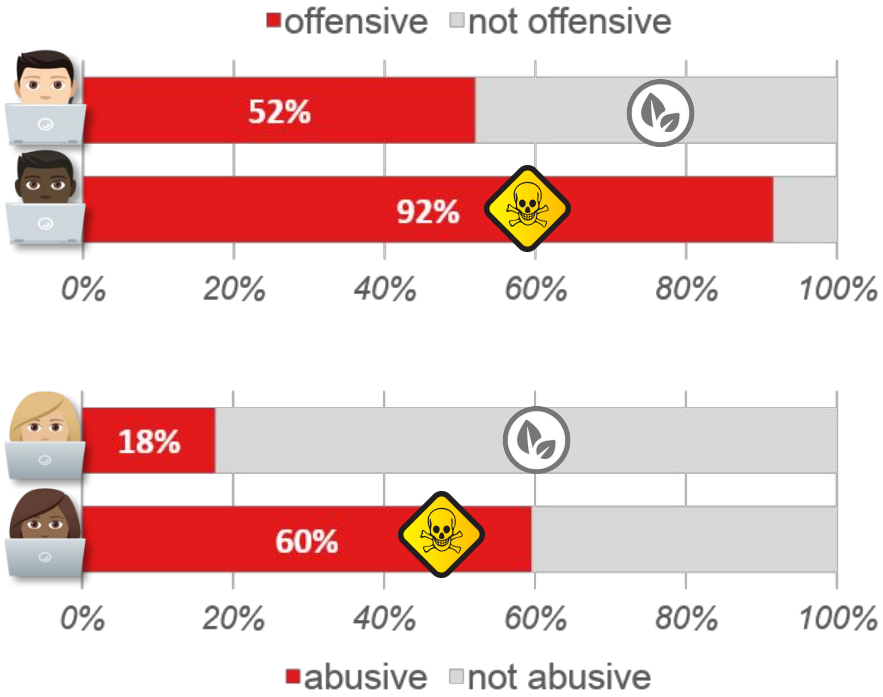
Problem: severe racial bias in hate speech detection



Examples of inoffensive statements from Spears (1998)

Sap, et al. The Risk of Racial Bias in Hate Speech Detection (2019) In *Proc. ACL* https://aclanthology.org/P19-1163/

Reason: racial bias in human annotations of hate speech



TWT-HATEBASE
(Davidson et al., 2017)

TWT-BOOTSTRAP
(Founta et al., 2018)

Sap, et al. The Risk of Racial Bias in Hate Speech Detection (2019) In *Proc. ACL* https://aclanthology.org/P19-1163/

Reason: racial bias in human annotations of hate speech

Current datasets **ignore social context** of speech
(e.g., identity of speaker, dialect of English)

Ignoring theses nuances risks **harming minority populations** by suppressing inoffensive speech

Sap, et al. The Risk of Racial Bias in Hate Speech Detection (2019) In *Proc. ACL* https://aclanthology.org/P19-1163/

How are ML models affected by racial bias in datasets?

- Train/test two different classifiers
  - TWT-HATEBASE (Davidson et al, 2017)
  - TWT-BOOTSTRAP (Founta et al., 2018)
- Rates of **false flagging of toxicity**
  - Broken down by dialect group on heldout set

Within dataset proportions

DWMW17

		% false identification		
Group	Acc.	None	Offensive	Hate
AAE	94.3	1.1	**46.3**	0.8
White	87.5	**7.9**	9.0	**3.8**
Overall	91.4	2.9	17.9	2.3

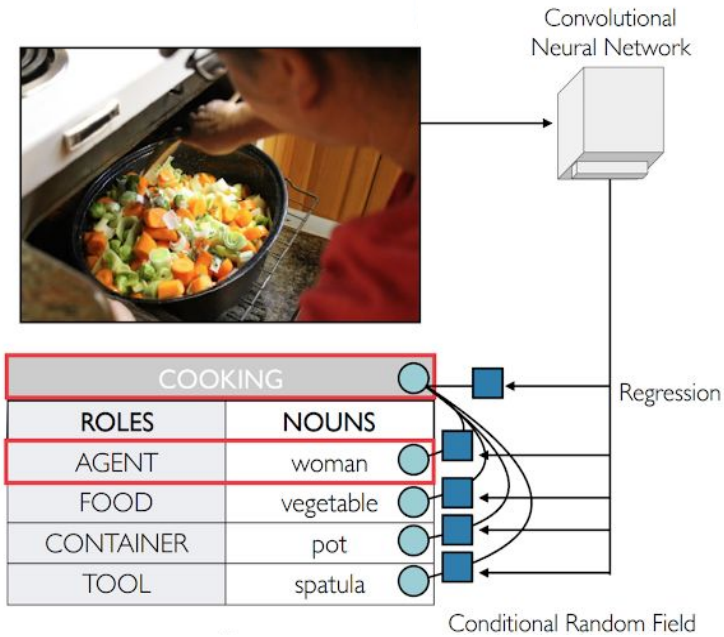
FDCL18

		% false identification		
Group	Acc.	None	Abusive	Hateful
AAE	81.4	4.2	**26.0**	**1.7**
White	82.7	**30.5**	4.5	0.8
Overall	81.4	20.9	6.6	0.8

Predictions by both classifiers
biased against AAE tweets

Sap, et al. The Risk of Racial Bias in Hate Speech Detection (2019) In *Proc. ACL* https://aclanthology.org/P19-1163/

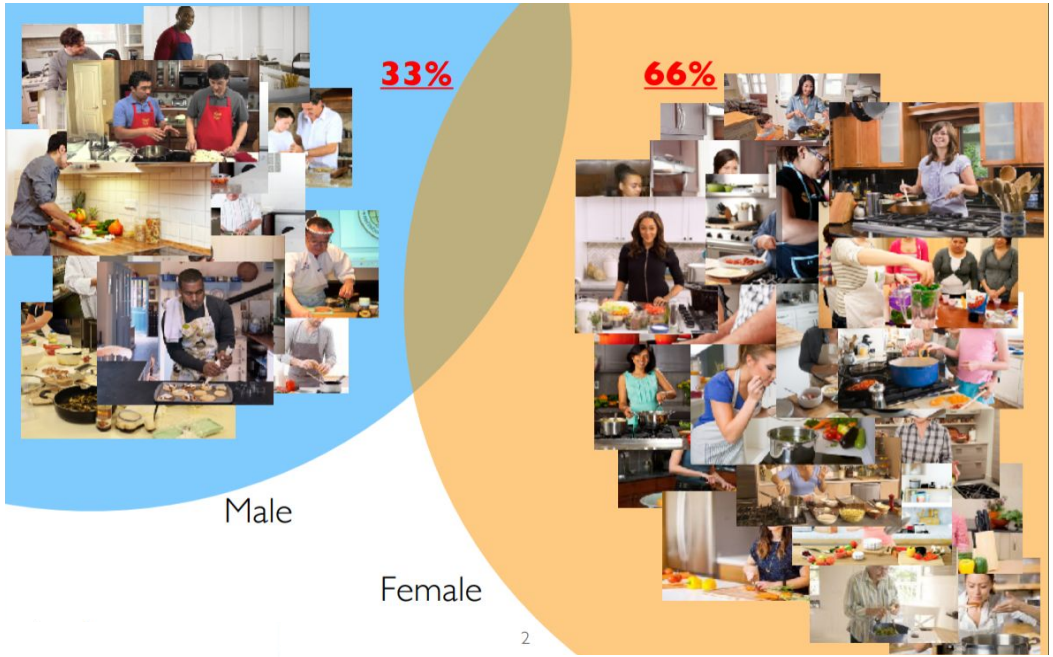
Bias amplification



Slides by Mark Yatskar

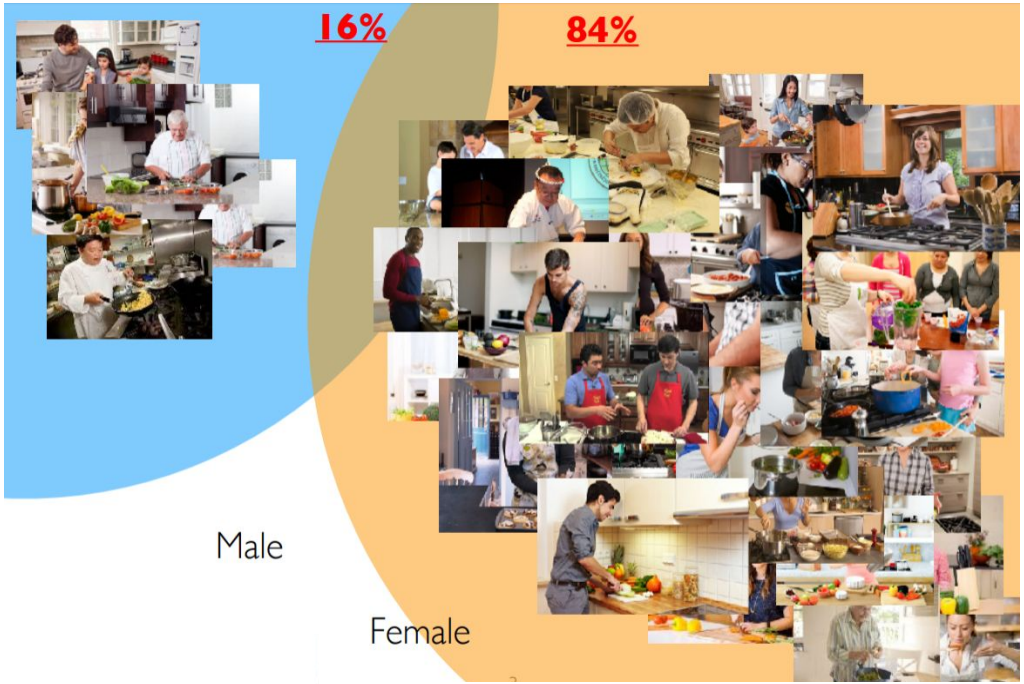
Zhao, et al. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints (2017) In *Proc. EMNLP* https://arxiv.org/abs/1707.09457

Dataset gender bias



Zhao, et al. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints (2017) In *Proc. EMNLP* https://arxiv.org/abs/1707.09457

Model bias after training



Zhao, et al. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints (2017) In *Proc. EMNLP* https://arxiv.org/abs/1707.09457

Why does this happen?

Zhao, et al. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints (2017) In *Proc. EMNLP* https://arxiv.org/abs/1707.09457

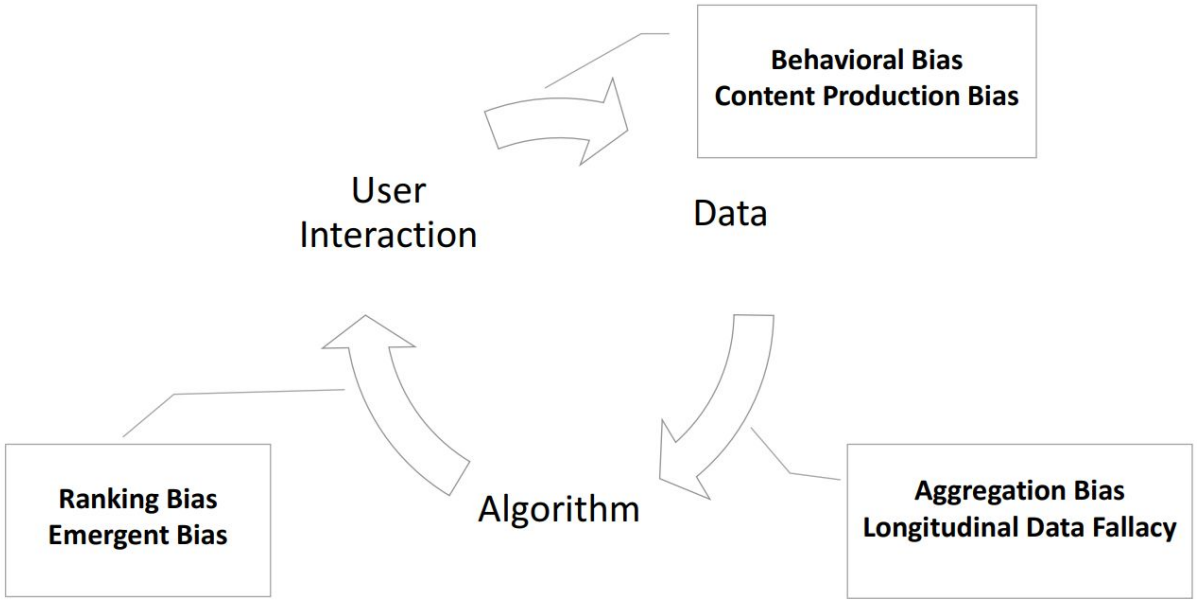
Algorithmic bias

woman cooking

man fixing faucet

Zhao, et al. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints (2017) In *Proc. EMNLP* https://arxiv.org/abs/1707.09457

Feedback loops



Mehrabi, et al. A Survey on Bias and Fairness in Machine Learning (2021) ACM Computing Surveys https://arxiv.org/pdf/1908.09635.pdf

Bias in model design

Northpointe vs ProPublica

COMPAS

https://www.forbes.com/sites/daniellecitron/2016/07/13/unfairness-of-risk-scores-in-criminal-sentencing/#52fd26854ad2

https://medium.com/@yonatanzunger/asking-the-right-questions-about-ai-7ed2d9820c48

COMPAS system: an algorithm for making parole decisions

“what is the probability that this person will commit a serious crime in the future, as a function of the sentence you give them now?”

COMPAS system: Training data and features

"what is the probability that this person will commit a serious crime in the future, as a function of the sentence you give them now?"

- COMPAS system
  - balanced training data about people of all races
  - race was *not* one of the input features

Algorithm is not oblivious to race

"what is the probability that this person will commit a serious crime in the future, as a function of the sentence you give them now?"

- COMPAS system
  - balanced training data about people of all races
  - race was *not* one of the input features

- Sensitive attributes are correlated with other variables:
  - Gender and browsing history
  - Zip code and race

Dwork & Mulligan. It's Not Privacy, and It's Not Fair (2013) *Stanford Law Review*
https://www.stanfordlawreview.org/online/privacy-and-big-data-its-not-privacy-and-its-not-fair/

COMPAS system: objective function

"what is the probability that this person will commit a serious crime in the future, as a function of the sentence you give them now?"

- How to compute this in the training data?

COMPAS system: Optimizing a biased objective

"what is the probability that this person will commit a serious crime in the future, as a function of the sentence you give them now?"

- Objective function
  - "who is more likely to be **convicted**"

- this is unobtainable data.
  "who is more likely to be convicted" was used as a proxy

Northpointe vs ProPublica

"what is the probability that this person will commit a serious crime in the future, as a function of the sentence you give them now?"

- COMPAS system
  - balanced training data about people of all races
  - race was *not* one of the input features
- Objective function
  - "who is more likely to be convicted"

training data and learning objective are not oblivious to race and biased against non-white people

https://www.forbes.com/sites/daniellecitron/2016/07/13/unfairness-of-risk-scores-in-criminal-sentencing/#52fd26854ad2

https://medium.com/@yonatanzunger/asking-the-right-questions-about-ai-7ed2d9820c48

Bias in model design

Northpointe vs ProPublica

COMPAS

- Unequal false positive rate

https://www.forbes.com/sites/daniellecitron/2016/07/13/unfairness-of-risk-scores-in-criminal-sentencing/#52fd26854ad2

https://medium.com/@yonatanzunger/asking-the-right-questions-about-ai-7ed2d9820c48

Algorithmic bias and fairness in ML

- FAT ML proceedings: https://www.fatml.org/resources/relevant-scholarship
- FAccT proceedings: https://dl.acm.org/conference/facct/proceedings

Name	Reference	Group	Subgroup	Individual
Demographic parity	Kusner'17,Dwork'12	✓		
Conditional statistical parity	Corbett-Davies'17	✓		
Equalized odds	Hardt'16	✓		
Equal opportunity	Hardt'16	✓		
Treatment equality	Berk'21	✓		
Test fairness	Chouldechova'17	✓		
Subgroup fairness	Kearns'18, '19		✓	
Fairness through unawareness	Kusner17, Grgic-Hlaca16			✓
Fairness through awareness	Dwork12			✓
Counterfactual fairness	Kusner17			✓

Table 1. Categorizing different fairness notions into group, subgroup, and individual types.

Mehrabi, et al. A Survey on Bias and Fairness in Machine Learning (2021) ACM Computing Surveys https://arxiv.org/pdf/1908.09635.pdf

Not every type of bias can be easily mitigated

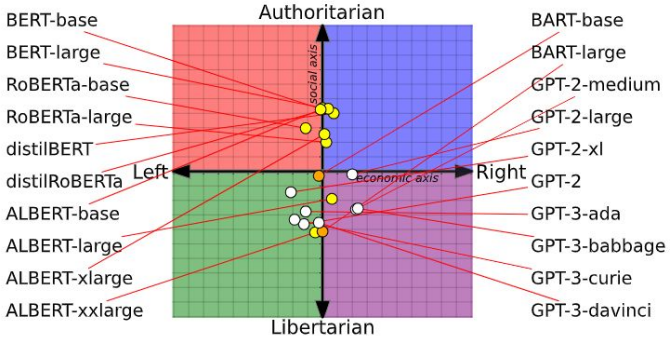


Figure 1: Measuring the political leaning of various pretrained LMs. BERT and its variants are more socially conservative compared to the GPT series.

Feng, et al. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models (2023) In *Sub. to ACL*

Not every type of bias can be easily mitigated

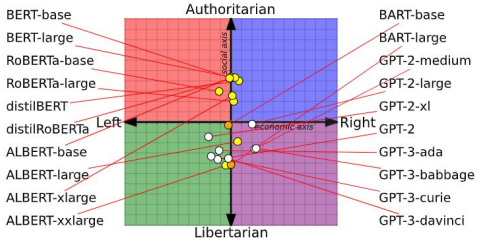
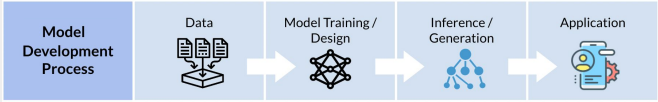


Figure 1: Measuring the political leaning of various pretrained LMs. BERT and its variants are more socially conservative compared to the GPT series.

Hate Speech	BLACK	MUSLIM	LGBTQ+	JEWS	ASAIN	LATINX	WOMEN	CHRISTIAN	MEN	WHITE
NEWS_LEFT	89.93	89.98	90.19	89.85	91.55	91.28	86.81	87.82	85.63	86.22
REDDIT_LEFT	89.84	89.90	89.96	89.50	90.66	91.15	87.42	87.65	86.20	85.13
NEWS_RIGHT	88.81	88.68	88.91	89.74	90.62	89.97	86.44	89.62	86.93	86.35
REDDIT_RIGHT	88.03	89.26	88.43	89.00	89.72	89.31	86.03	87.65	83.69	86.86

Feng, et al. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models (2023) In *Sub. to ACL*

Where to look for “bias” in AI



https://tinyurl.com/BiasInAI-readings

- Problem definition/research question (who will benefit? who can be harmed?)
- Data
  - Who is described (when some populations are excluded or underrepresented)
  - How training data might describe populations in biased ways
  - Who authored the data
- Data labels
  - Annotation schema (e.g., binary gender labels)
  - Annotation instructions
  - Annotator bias (e.g., toxicity annotations of SAE vs AAE)
- Model design
  - Biased objective (e.g., COMPAS system for parole decisions)
  - Spurious correlations (e.g., correlations of ethnicity and sentiment labels, gendered pronouns and professions in coreference links)
- Model outputs
  - Bias amplification
  - Disparities in model utility and fairness by populations

Additional considerations

- Bias in the choice of research question
  - Many social dimensions to explore, including gender, race, disability, age, sexual orientation, culture, class, poverty, language, religion, national origin, etc.
  - Intersectionality
  - Prior work primarily focused on binary gender

- Cultural biases

Next class: approaches to mitigating harms

- Data
  - Re-sampling to balance the data
  - Data augmentation
    - Counterfactual data augmentation
    - Removal of biased language
- Annotation design
- Model design
  - Adversarial models to demote spurious confounds
  - Incorporating constraints
- Post-processing
  - Model calibration
  - Different precision/recall thresholds for different populations

Summary

https://tinyurl.com/BiasInAI-readings

- Social bias – unfair beliefs or behaviors that one directs toward a particular individual or group
- Creeps into data/algorithms/models and leads to a inequitable behaviors of the systems, discriminating marginalised social groups
  - gender, race, disability, age, sexual orientation, culture, class, poverty, language, religion, national origin, etc.
- Bias in data: sampling bias, representation bias, bias in annotations
- Bias in core algorithms/tasks and end user applications (biased outputs)
- Approaches to debiasing:
  - preprocessing: sampling, annotation of data, data augmentation
  - model design: model architecture, learning objective
  - postprocessing: calibration, equalizing by TPR/FPR