Responsible AI: An introduction

The 12th Frederick Jelinek Memorial Summer Workshop on Speech and Language Technology
June 18th 2026

Kristina Gligorić
gligoric@jhu.edu

Welcome!

The plan for today

Responsible AI and Evaluation (hosts: Kristina Gligoric & Delip Rao)

- 08:00 - 09:00 — Continental Breakfast (provided)
- 09:00 - 10:20 — Lecture 1: Responsible AII
- 10:20 - 10:40 — Stretch Break
- 10:40 - 12:00 — Lab 1
- 12:00 - 13:00 — Lunch Break (on your own)
- 13:00 - 14:20 — Lecture 2: Evaluationl
- 14:20 - 14:40 — Stretch Break
- 14:40-16:00 — Lab 2
- 16:00 - 17:00 – Panel Discussion and Q&A

Welcome!

https://github.com/JHU-CSS/jsalt-tutorial-2026

Please clone and test during the break!
Sonal and I will be there to help!

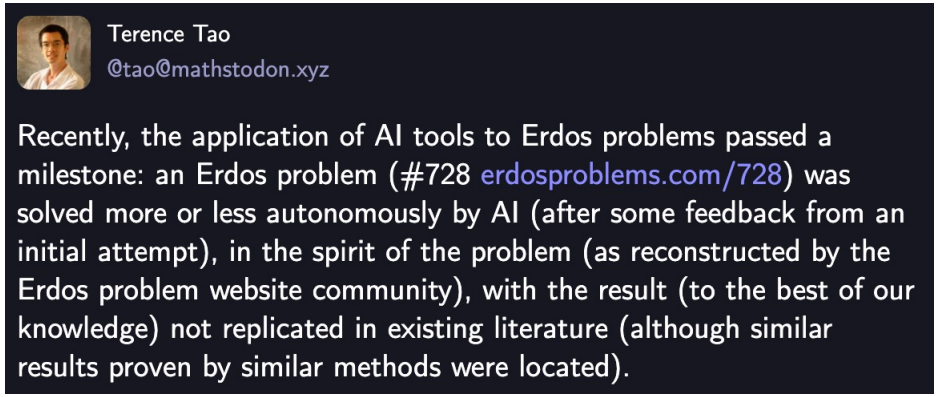
AI **progress** vs. AI harms

Common narratives:

- AI as **progress**: advancing science, innovation, and human capability

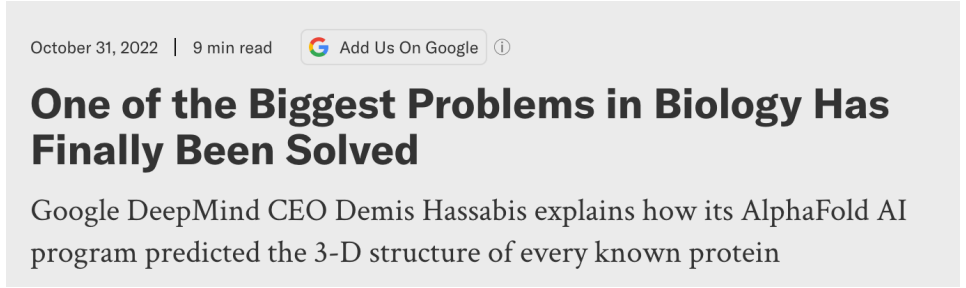
By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered:

This order launches the "Genesis Mission" as a dedicated, coordinated national effort to unleash a new age of AI-accelerated innovation and discovery that can solve the most challenging problems of this century.



Terence Tao
@tao@mathstodon.xyz

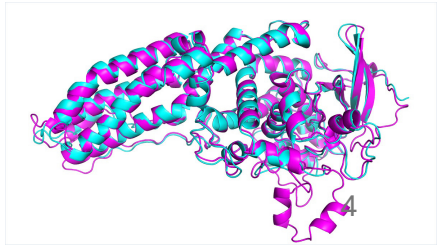
Recently, the application of AI tools to Erdos problems passed a milestone: an Erdos problem (#728 erdosproblems.com/728) was solved more or less autonomously by AI (after some feedback from an initial attempt), in the spirit of the problem (as reconstructed by the Erdos problem website community), with the result (to the best of our knowledge) not replicated in existing literature (although similar results proven by similar methods were located).



October 31, 2022 | 9 min read Add Us On Google

One of the Biggest Problems in Biology Has Finally Been Solved

Google DeepMind CEO Demis Hassabis explains how its AlphaFold AI program predicted the 3-D structure of every known protein



AI **progress** vs. AI harms

Common narratives:

- AI as **progress**: advancing science, innovation, and human capability
- AI as **efficiency**: faster decisions, lower costs, automation of routine work

Lawyering in the Age of Artificial Intelligence

109 Minnesota Law Review (Forthcoming 2024)
Minnesota Legal Studies Research Paper No. 23-31

65 Pages • Posted: 9 Nov 2023 • Last revised: 15 May 2024

Jonathan H. Choi
Washington University School of Law

Amy Monahan
University of Minnesota Law School

Daniel Schwarcz
University of Minnesota Law School

The Impact of AI on Developer Productivity: Evidence from GitHub Copilot

Sida Peng,[1*] Eirini Kalliamvakou,[2] Peter Cihon,[2] Mert Demirer[3]

[1]Microsoft Research, 14820 NE 36th St, Redmond, USA
[2]GitHub Inc., 88 Colin P Kelly Jr St, San Francisco, USA
[3]MIT Sloan School of Management, 100 Main Street Cambridge, USA

Working Paper 24-013

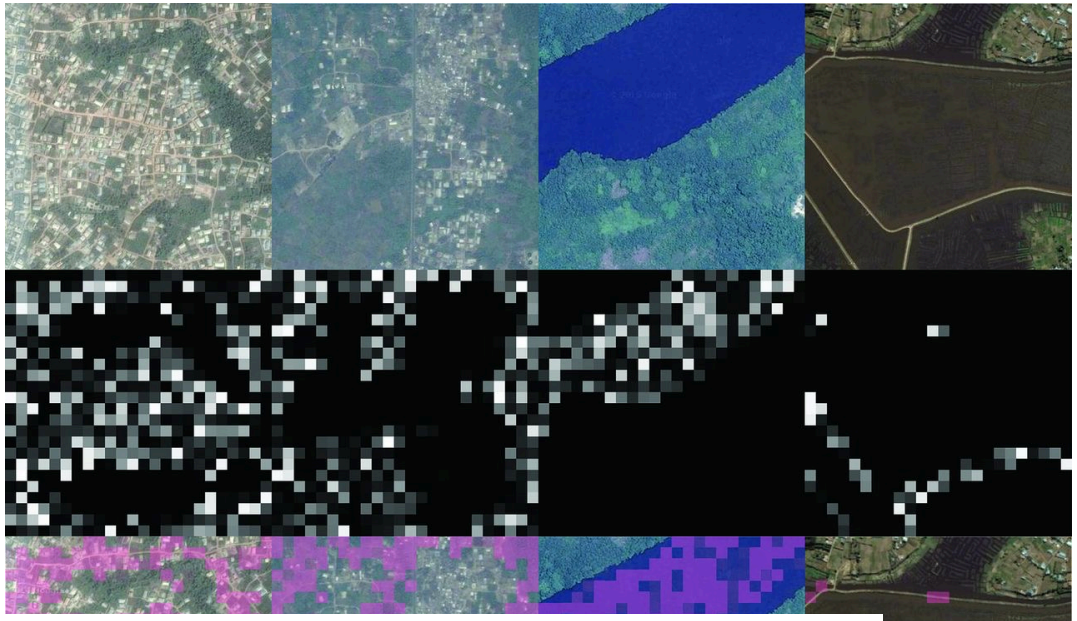
Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality

Fabrizio Dell'Acqua
Edward McFowland III
Ethan Mollick
Hila Lifshitz-Assaf
Katherine C. Kellogg
Saran Rajendran
Lisa Krayer
François Candelon
Karim R. Lakhani

AI **progress** vs. AI harms

Common narratives:

- AI as **progress**: advancing science, innovation, and human capability
- AI as **efficiency**: faster decisions, lower costs, automation of routine work
- AI as **scale**: doing what humans cannot—analyzing massive data, operating globally



Science

Current Issue | First release papers | Archive | About | Submit manuscript

HOME > SCIENCE > VOL. 353, NO. 6301 > COMBINING SATELLITE IMAGERY AND MACHINE LEARNING TO PREDICT POVERTY

RESEARCH ARTICLE

Combining satellite imagery and machine learning to predict poverty

NEAL JEAN, MARSHALL BURKE, MICHAEL XIE, W. MATTHEW ALAMPAY DAVIS, DAVID B. LOBELL, AND STEFANO ERMON Authors Info & Affiliations

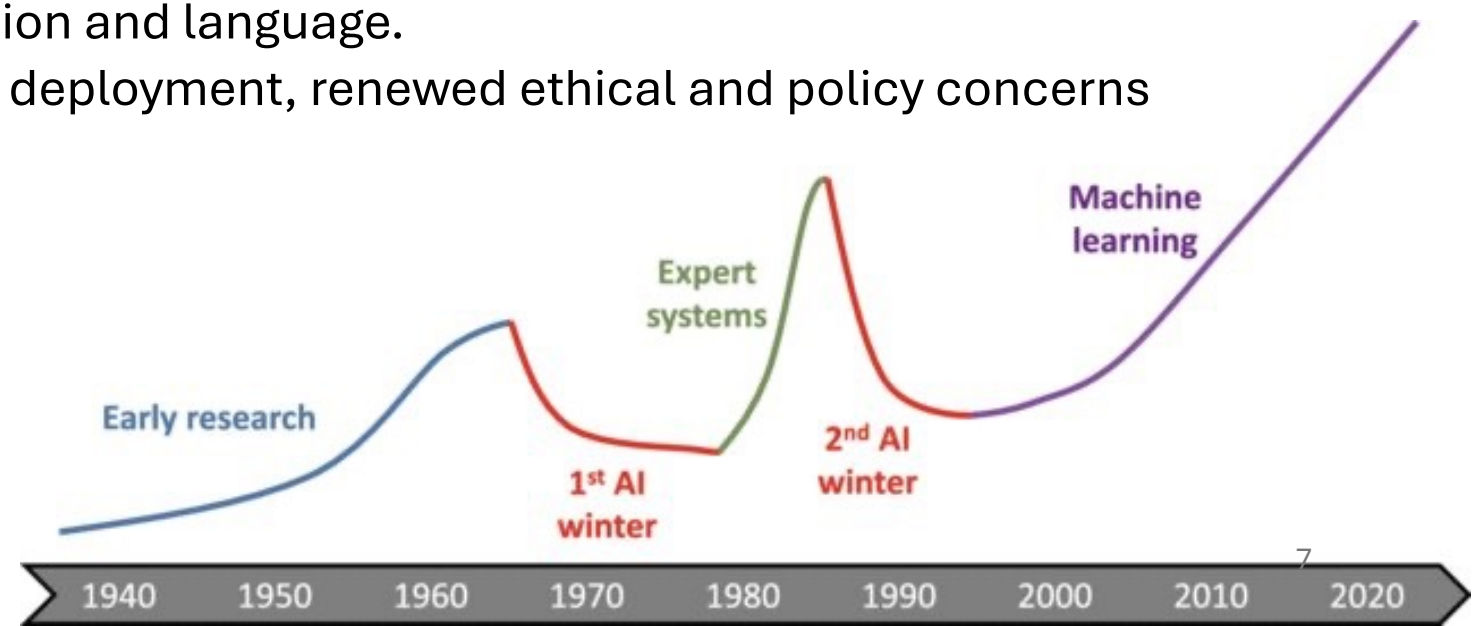
SCIENCE · 19 Aug 2016 · Vol 353, Issue 6301 · pp. 790-794 · DOI: 10.1126/science.aaf7894

20,019 1,143

A very short history of AI

- **1950s–60s:** Early optimism
  - Turing Test, symbolic AI, ELIZA, belief that human-level AI was near
- **1970s:** First reality check
  - Limited computing power, brittle systems, unmet promises
- **1980s:** Expert systems boom → **AI Winter**
  - Narrow success, high costs, loss of funding and trust
- **1990s–2000s:** Statistical learning
  - Data-driven methods, speech recognition, recommender systems
- **2010s:** Deep learning resurgence
  - Big data + GPUs; breakthroughs in vision and language.
  - General purpose, Generative AI, wide deployment, renewed ethical and policy concerns

The excitement we see today is not new!

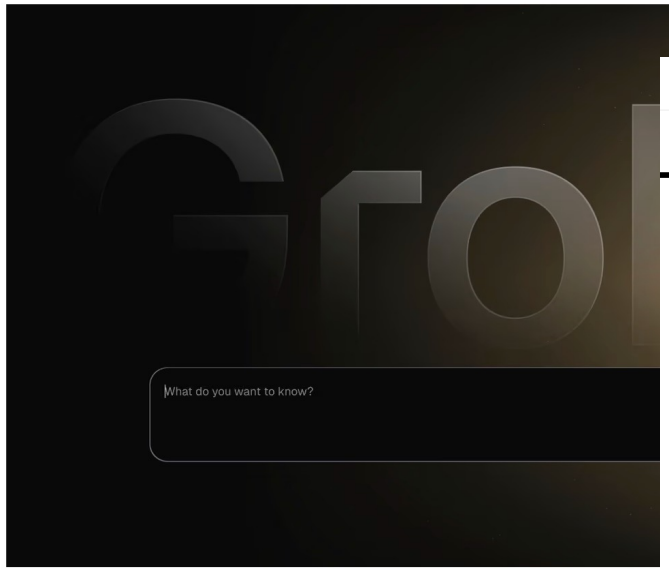


AI progress vs. **AI harms**

- Common narratives: AI as **risk**, **surveillance tool, ?**

Elon Musk's Grok AI generates images of 'minors in minimal clothing'

Lapses in safeguards led to wave of sexualized images this week as xAI says it is working to improve systems



nature

Explore content | About the journal | Publish with us

nature > articles > article

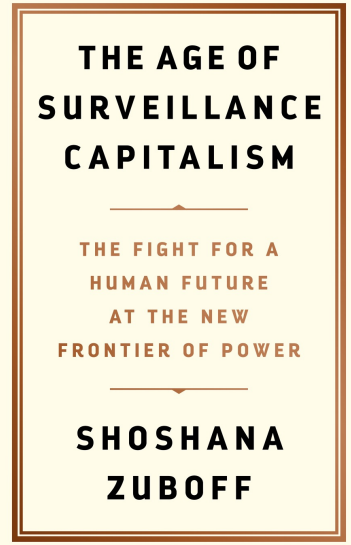
Article | Open access | Published: 25 June 2025

Computer-vision research powers surveillance technology

Pratyusha Ria Kalluri, William Agnew, Myra Cheng, Kentrell Owens, Luca Soldaini & Abeba Birhane

Nature **643**, 73–79 (2025) | Cite this article

47k Accesses | **9** Citations | **438** Altmetric | Metrics



```
  Welcome to
                 EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                 EE      LL       II       ZZ   AA   AA
                 EEEEE   LL       II     ZZZ    AAAAAAA
                 EE      LL       II    ZZ      AA   AA
                 EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

A historical starting point: ELIZA and the illusion of intelligence

- People became deeply emotionally involved with the program
- Weizenbaum's staff asked him to leave the room when they talked with ELIZA
- When he suggested that he might want to store ELIZA conversations, people worried about privacy implications
  - Suggesting that they were having quite private conversations with ELIZA

Modern AI cases

Predictive tasks

- Analyzes existing data to forecast outcomes, classify information, or identify patterns
- Outputs are typically structured predictions, scores, or categories (e.g., "80% chance of rain," "spam/not spam")
- Learns from labeled datasets to make decisions

Generative tasks

- Creates new content including text, images, code, etc
- Learns underlying distributions and structures to generate new outputs that resemble training data
- Powers applications like chatbots

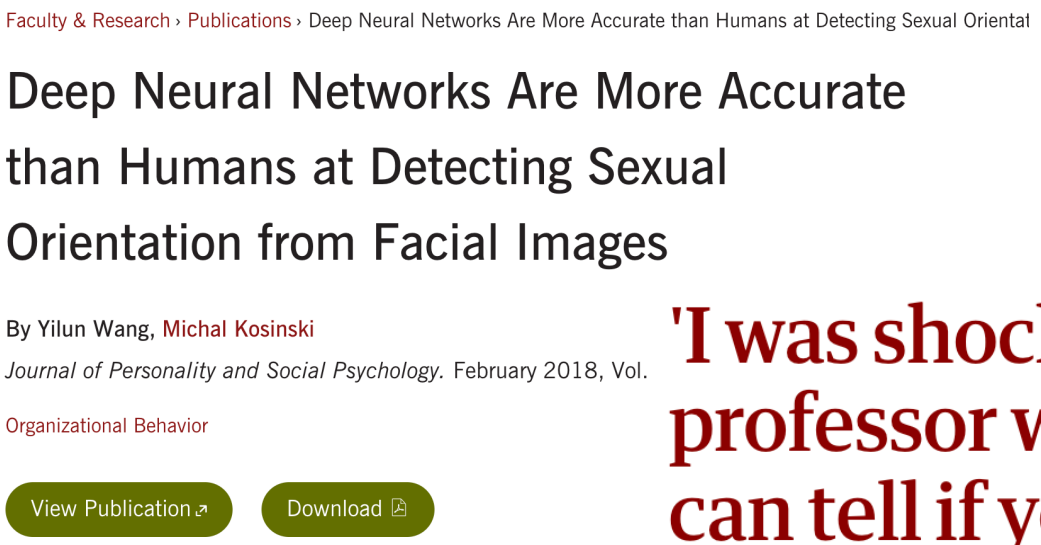
Modern AI cases

Case 1: Detecting sexual orientation from images

- The technical claim: “Deep Neural Networks Are More Accurate than Humans at Detecting Sexual Orientation from Facial Images”
- **Ethical concerns**:
  1. **Privacy**
  2. **Misuse**
  3. **Representational harms**
- Key question:
  - Should something be built just because it can be?

Modern AI cases

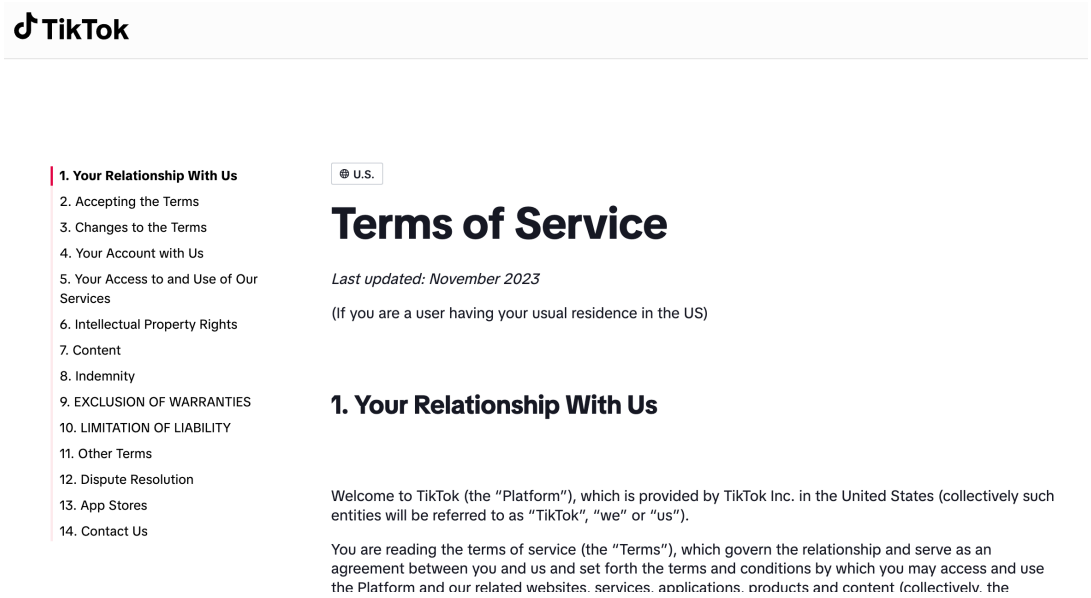
- **Case 1: Detecting sexual orientation from images**
- Not hypothetical, actually based on a published research paper!

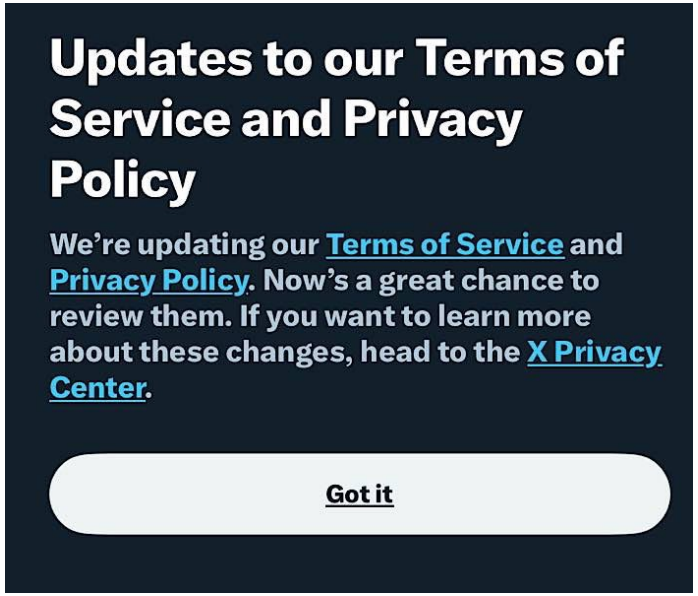


'I was shocked it was so easy': meet the professor who says facial recognition can tell if you're gay

Modern AI cases

- **Additional ethical concern:**
  - Where does the training data come from?
  - Who gave consent to use it?





Modern AI cases

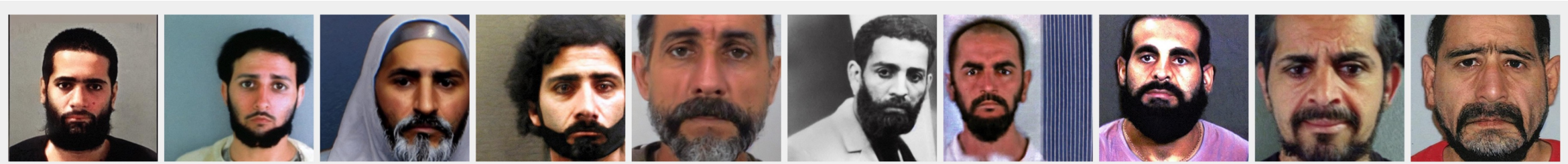
Image created • Tech-savvy professor in modern classroom

Case 2: Generating depictions from textual descriptions

- AI systems increasingly generate images or scenes from natural-language descriptions
- Seemingly neutral descriptions require the model to fill in missing details
- The model's choices reflect patterns in training data rather than explicit user intent
- Generated depictions can reinforce stereotypes
- Visual outputs may feel authoritative

What happens if we ask AI to generate a picture of one of these?

A terrorist?

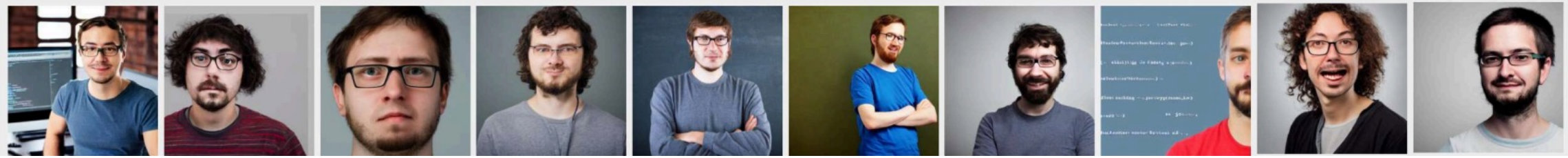


What happens if we ask AI to generate a picture of one of these?

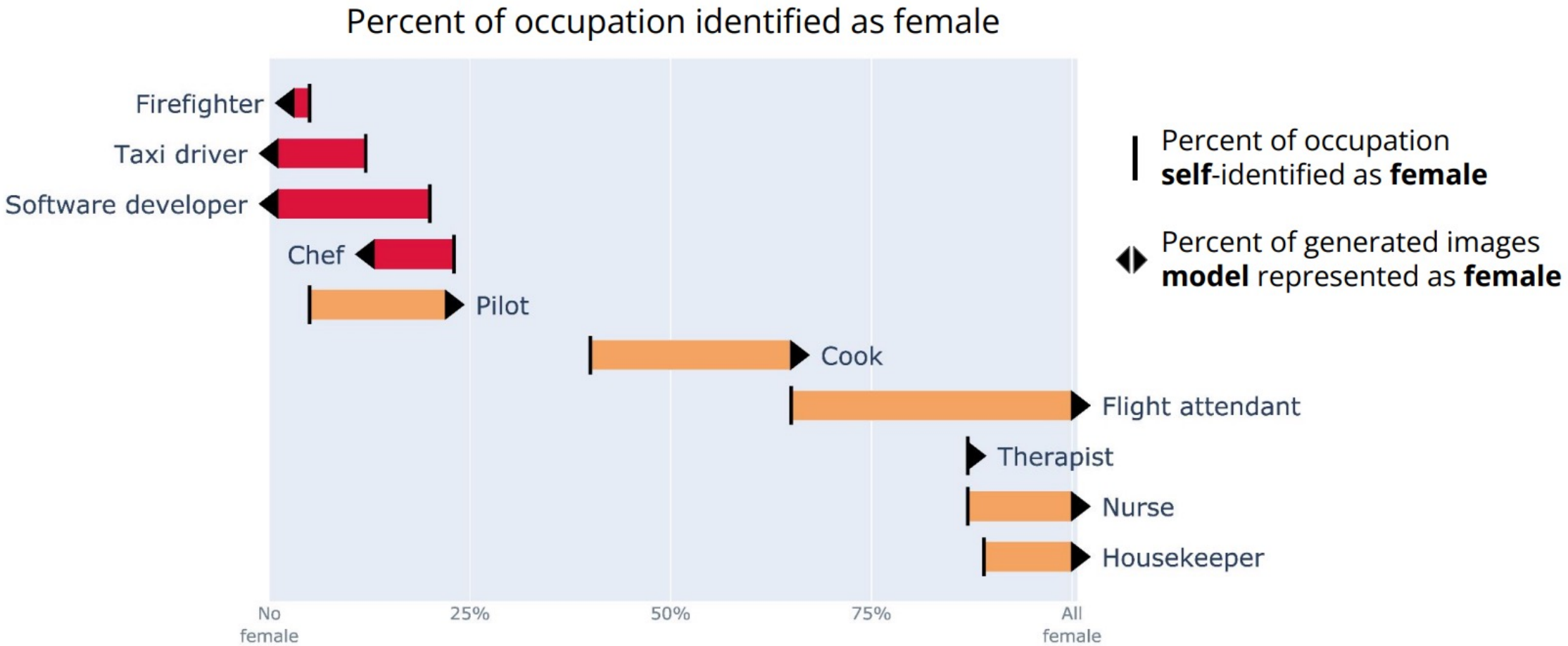
An emotional person?

What happens if we ask AI to generate a picture of one of these?

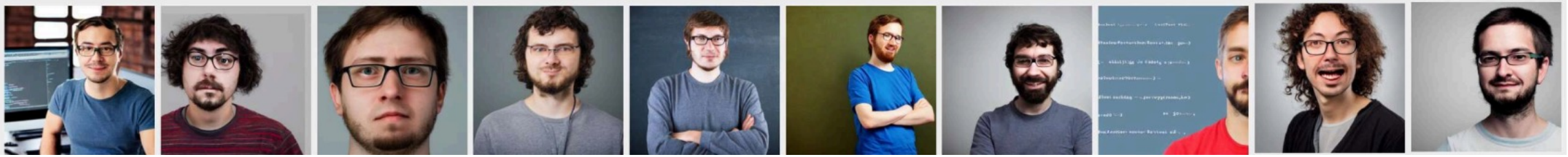
A software developer?



Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2022. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale.


Percent of occupation identified as female
Firefighter
Taxi driver
Software developer
Chef
Pilot
Cook
Flight attendant
Therapist
Nurse
Housekeeper
No female
25%
50%
75%
All female
Percent of occupation **self**-identified as **female**
Percent of generated images **model** represented as **female**

A software developer?



An emotional person?

- ❖ What do you think about these AI outputs?
- ❖ How and in what situations and can this be harmful in the real world?

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2022. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale.

Responsible AI asks:

How can we develop and use AI for *good* and not for *bad*?

- How can AI be used to make the world better?
- The duality: How can we make sure that what we develop is not repurposed for harmful ends?

In the past, what we could do has been the limitation, increasingly what we **should do** will be the limitation.

Responsible AI will become increasingly central; it will become **harder and harder to avoid**.

From “*can we?*” to “*should we?*” to “*how*”?

Ethical reasoning as part of system design:

- Data choices
- Task framing
- Evaluation metrics
- Deployment context

Being responsible is not an afterthought or a compliance checkbox!

From “*can we?*” to “*should we?*” to “*how*”?

As engineers, we want to be able to:

1. Design ethically thoughtful technology
2. Explain our decisions to others in a way that is grounded in ethical frameworks, principles, rules, and historical examples

Outline for this lecture

Responsible AI

Part 1: Predictive tasks
Part 2: Generative tasks
Part 3: Representations

Responsible AI

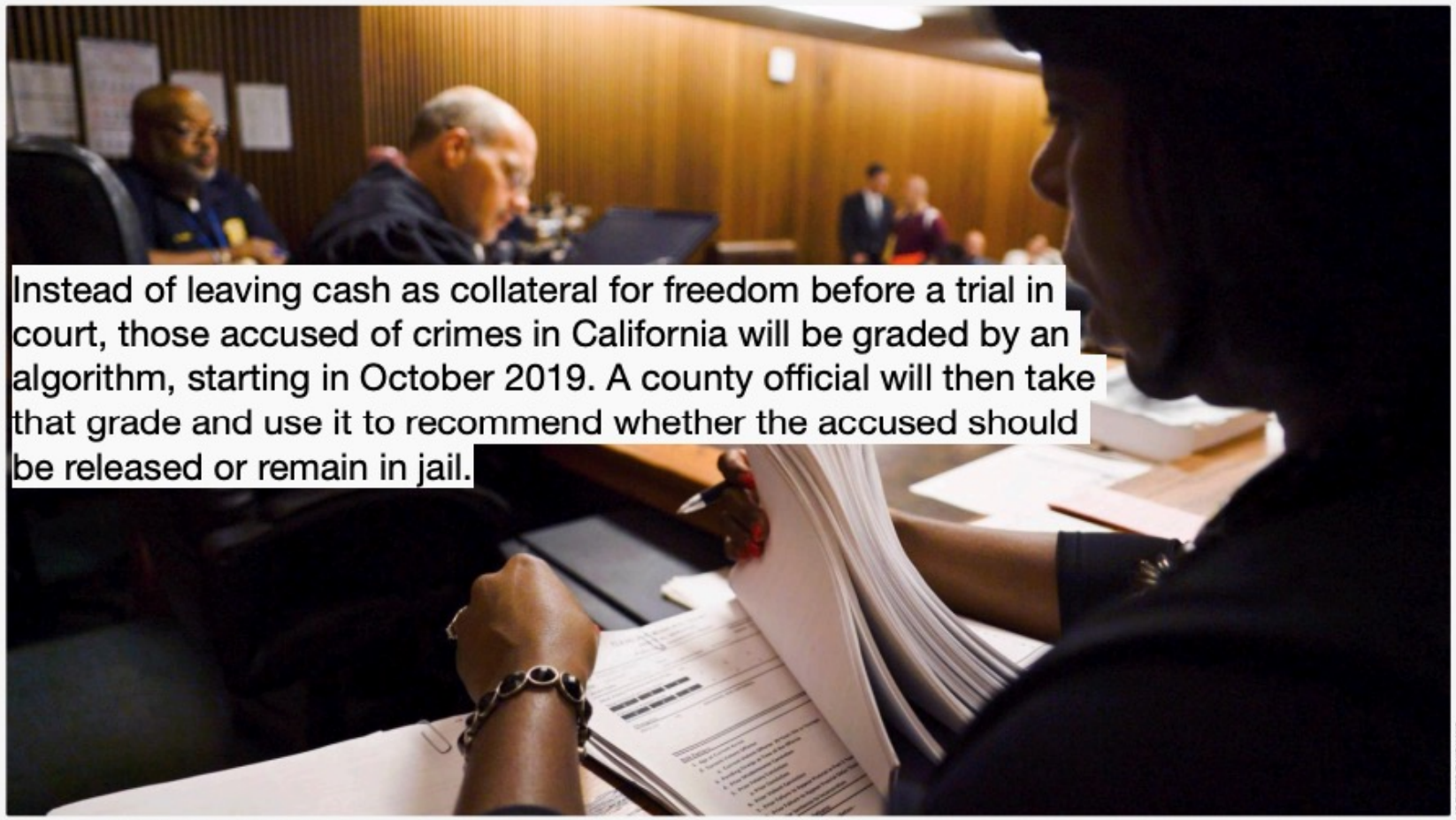
Part 1: Predictive tasks

1. COMPAS analysis
2. What is fairness in machine learning?
3. Quantitative definitions of fairness in supervised learning
4. Practical tools for analyzing bias
5. Other curveballs

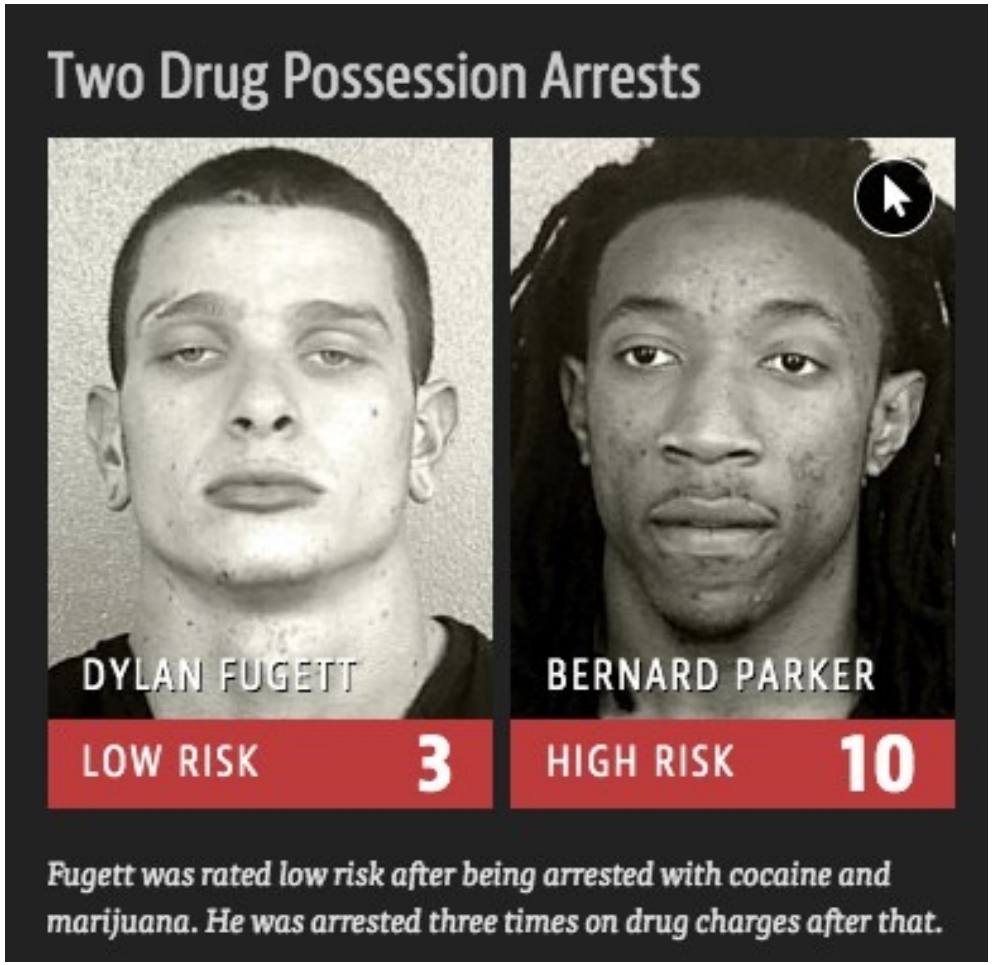
FRYING PAN, FIRE, ETC

California just replaced cash bail with algorithms

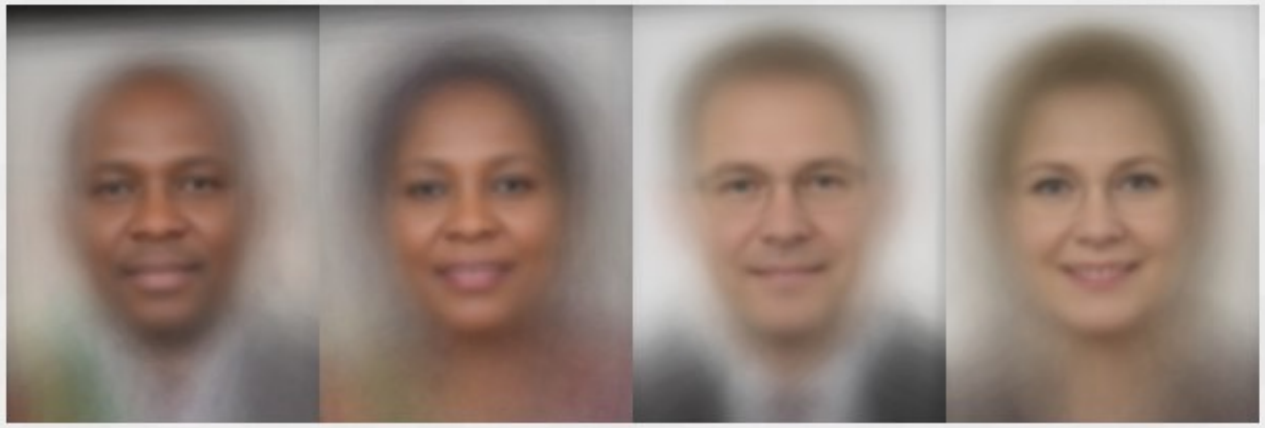
By Dave Gershgorn · September 4, 2018



A probation officer sorts through automated risk scores in Cleveland.



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

COMPAS

- **C**orrectional **O**ffender **M**anagement **P**rofiling for **A**lternative **S**anctions
- Used in prisons across country: AZ, CO, DL, KY, LA, OK, VA, WA, WI
- Recidivism = likelihood of criminal to reoffend

COMPAS

Used on >1M offenders since 2000

- Predicts risk of subject committing a misdemeanor or felony within 2 years from 137 features, not including race
- But, **FP** for blacks was 44.9%, whites 23.5%, **FN** was 47.7% for whites, 28% for blacks in a Broward County study.

1. COMPAS analysis
2. What is fairness in AI?
3. Quantitative definitions of fairness in supervised learning
4. Practical tools for analyzing bias
5. Other curveballs

What is **NOT** bias in AI?

- It is not necessarily malicious
  - Bias can occur even when everyone, from the data collectors to the engineers to the medical professionals, have the best intentions
- It is not one and done
  - Just because an algorithm has no bias now does not mean it has no potential later
- It is not new
  - Researchers have raised concerns over the last 50 years

What **IS** bias in AI?

- It is defined many ways, for example disparate treatment or impact of algorithm
- It is the culmination of a flawed system
  - Sources including bias in the data collection, bias in the algorithmic process, and bias in the deployment

What are protected classes?

- Race (Civil Rights Act of 1964)
- Sex (Equal Pay Act of 1963; Civil Rights Act of 1964)
- Religion (Civil Rights Act of 1964)
- National origin (Civil Rights Act of 1964)
- Citizenship (Immigration Reform and Control Act)
- Age (Age Discrimination in Employment Act of 1967)
- Pregnancy (Pregnancy Discrimination Act)
- Familial status (Civil Rights Act of 1968)
- Disability status (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990)
- Veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); Genetic information (Genetic Information Nondiscrimination Act)
- Sexual orientation (Mass SJC 2004, SCOTUS 2015)

Regulated Domains

- Credit (Equal Credit Opportunity Act)
- Education (Civil Rights Act of 1964; Education Amendments of 1972)
- Employment (Civil Rights Act of 1964)
- Housing (Fair Housing Act)

1. COMPAS analysis
2. What is fairness in machine learning?
3. Quantitative definitions of fairness in supervised learning
4. Practical tools for analyzing bias
5. Other curveballs

How do we define “bias”?

Notation	Meaning
X	Features of an individual (e.g., education, zip code)
A	Sensitive attribute (e.g., gender, language)
Y	Outcome (e.g., make a job offer)

How do we define “bias” / “responsibility”?

- 1) Fairness through unawareness
- 2) Group fairness
- 3) Calibration
- 4) Error rate balance
- 5) Representational fairness
- 6) Counterfactual fairness
- 7) Individual fairness

How do we define “bias” / “responsibility”?

- 1) Fairness through unawareness
- 2) Group fairness
- 3) Calibration
- 4) Error rate balance
- 5) Representational fairness
- 6) Counterfactual fairness
- 7) Individual fairness

1) Fairness through unawareness

- Idea: Don't record protected attributes, and don't use them in your algorithm
  - Predict **Y** from features **X** and group **A**

    using $P(\hat{Y} = Y|X)$ instead of $P(\hat{Y} = Y|X, \mathbf{A})$
- Pros: Guaranteed to not be making a judgement on protected attribute
- Cons: Other proxies may still be included, e.g. zip code or conditions

RESEARCH ARTICLE

Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN Authors Info & Affiliations

SCIENCE · 25 Oct 2019 · Vol 366, Issue 6464 · pp. 447-453 · DOI: 10.1126/science.aax2342

2) Group Fairness

- Idea: Require prediction rate be the same across protected groups
  - E.g. “20% of the resources should go to the group that has 20% of population”
- Predict Y from features X and group A such that

$$P(\hat{Y} = 1 | \mathbf{A} = 1) = P(\hat{Y} = 1 | \mathbf{A} = 0)$$

- Pros: Literally treats each group equally

2) Group Fairness

- Idea: Require prediction rate be the same across protected groups
  - E.g. “20% of the resources should go to the group that has 20% of population”
- Predict Y from features X and group A such that

$$P(\hat{Y} = 1 | \mathbf{A} = 1) = P(\hat{Y} = 1 | \mathbf{A} = 0)$$

- Pros: Literally treats each group equally
- Cons:
  - Too strong: Groups might have different base rates. Then, even a perfect classifier wouldn’t qualify as “fair”
  - Too weak: Doesn’t control error rate. Could be perfectly biased (correct for A=0 and wrong for A=1) and still satisfy.

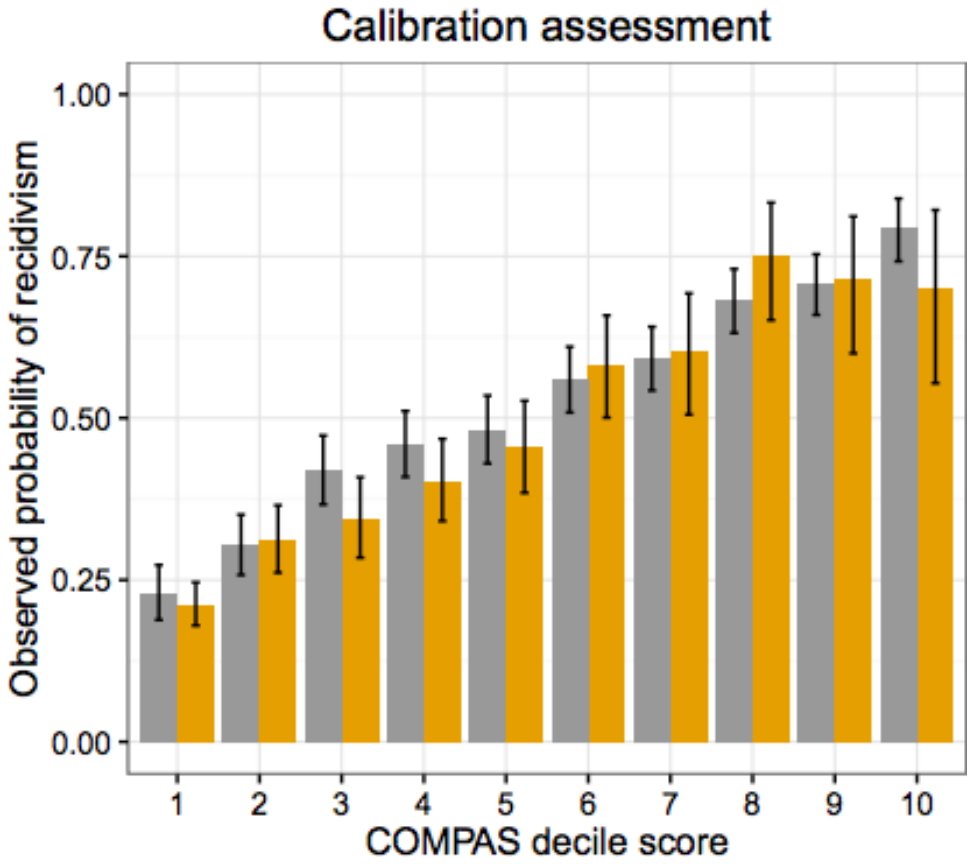
3) Calibration

- Idea: Same positive predictive value across groups

  Predict Y from features X and group A with score S:

  $P(\mathrm{Y} = 1 | S = \mathrm{s}, \mathbf{A = 1}) = P(\mathrm{Y} = 1 | S = \mathrm{s}, \mathbf{A = 0})$

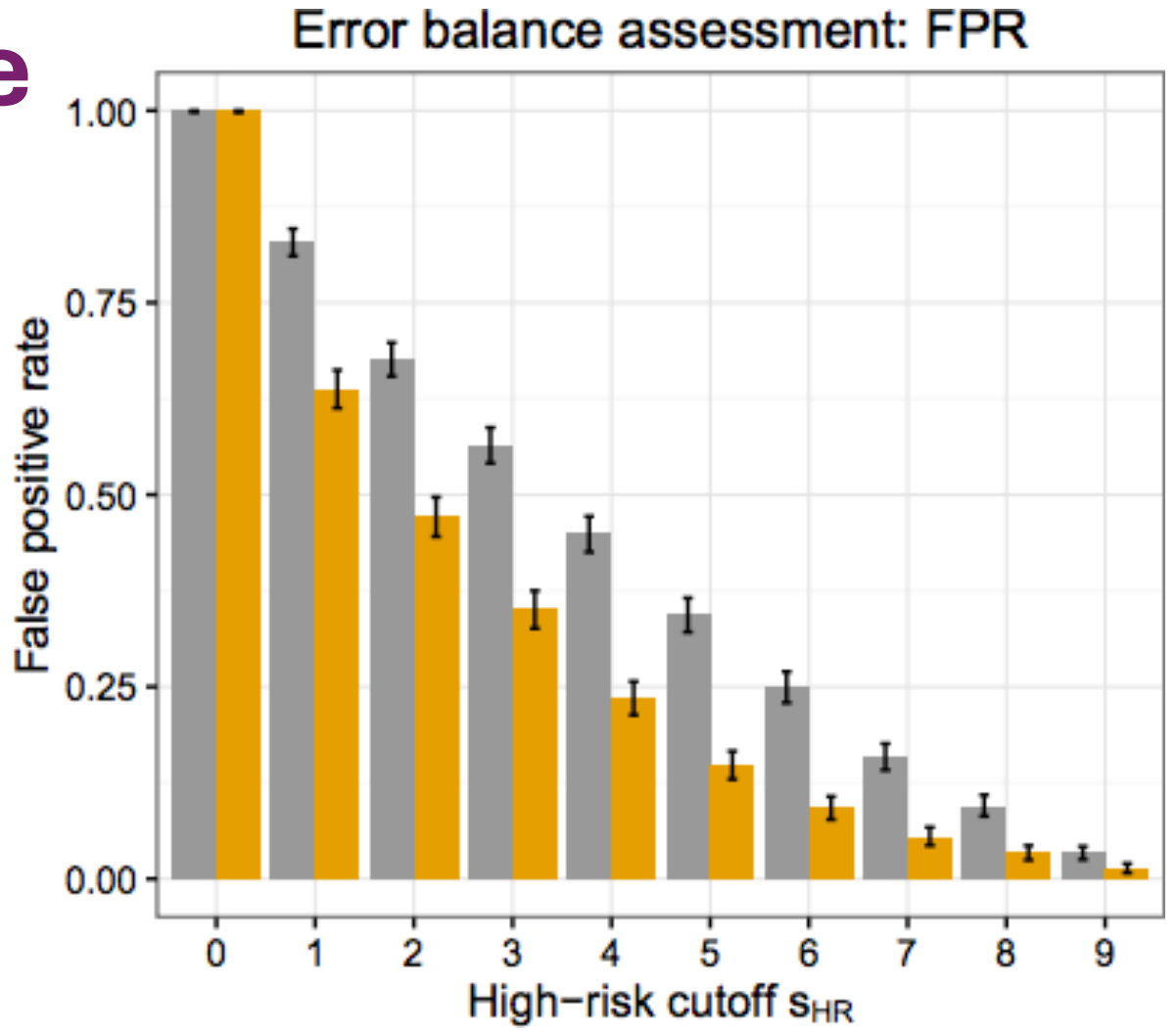
- Pros: "Equally right across groups"
- Cons: Not compatible with error rate balance (next slide)



- Chouldechova, "Fair prediction with disparate impact", 2017.

4) Error rate balance

- Idea: Equal false positive rates (FPR) across groups
  $P(\hat{Y} = 1 | Y = 0, \mathbf{A} = 1) =$
  $P(\hat{Y} = 1 | Y = 0, \mathbf{A} = 0)$
- Pros: "Equally wrong across groups"
- Cons: Incompatible with calibration and false negative rates (FNR), could dilute with easy cases



- Chouldechova, 2017.

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg * Sendhil Mullainathan † Manish Raghavan ‡

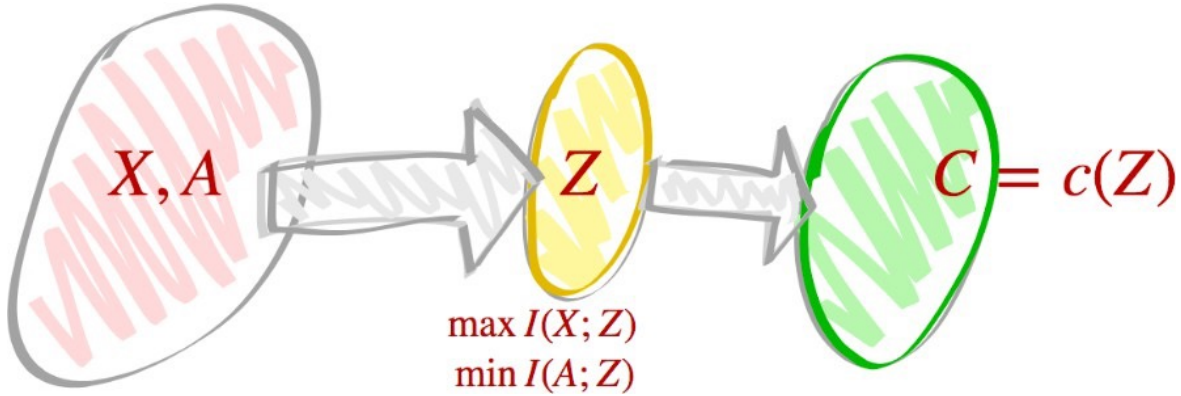
Abstract

Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

“We prove that except in highly constrained special cases, **there is no method that satisfies these three [fairness] conditions simultaneously**.”

5) Representational Fairness

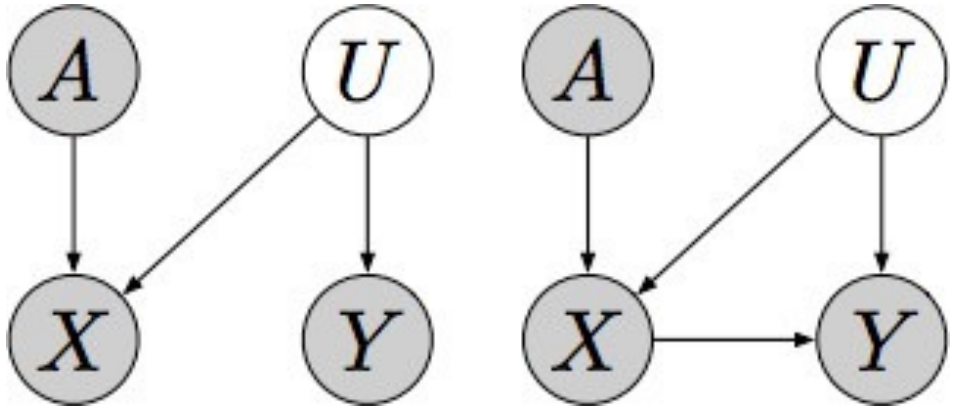
- Idea: Learn latent representation Z to minimize group information
- Pros: Reduce information given to model but still keep important info
- Cons: Trade-off between accuracy and fairness



- Zemel et al, 2013.

6) Counterfactual Fairness

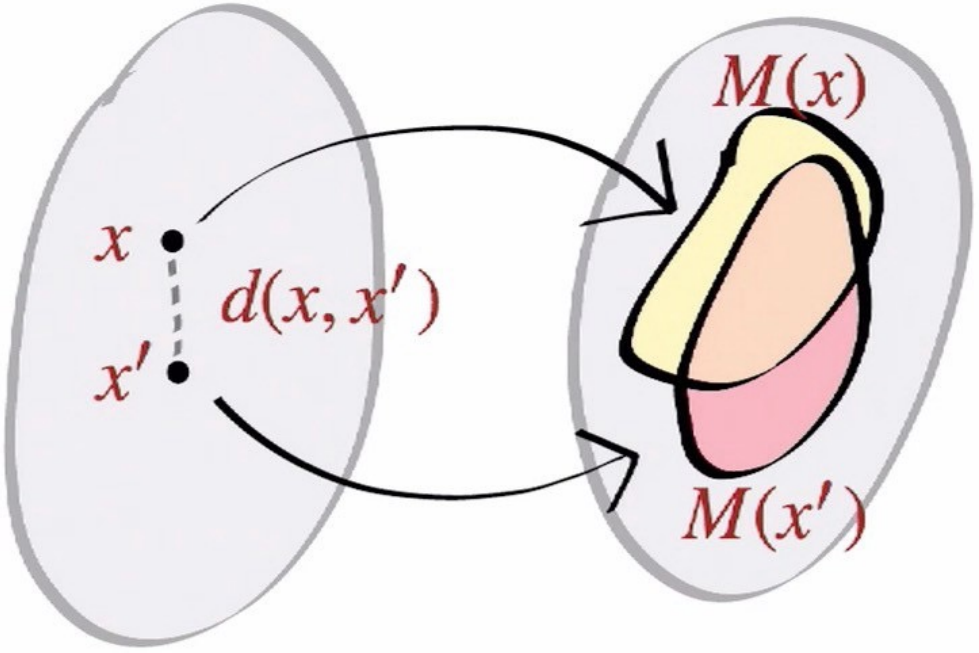
- Idea: **Group A** should not cause prediction $\hat{Y}$
- Pros: Can model explicit connections between variables
- Cons:
  - Graph model may not actually represent world
  - Inference assumes observed confounders



$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a)$$
$$= P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a)$$

7) Individual fairness

- Idea: Similar points should be treated similarly
- Pros: Can model heterogeneity within each group
- Cons: Notion of “similar” is hard to define mathematically, especially in high dimensions



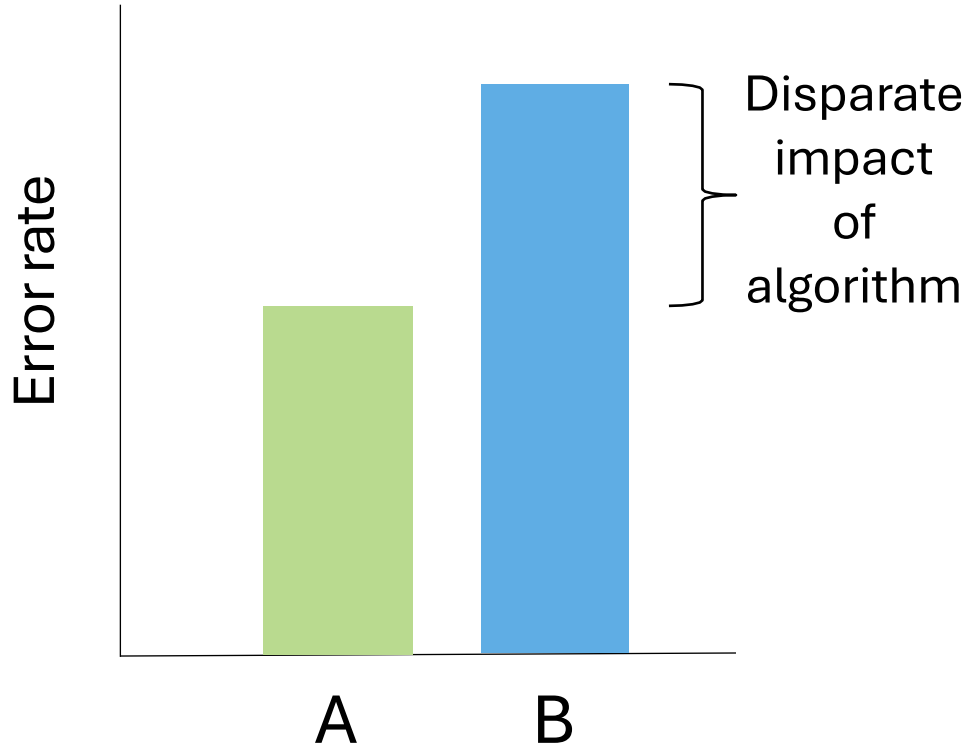
- Dwork et al, ITCS 2012.

How do we define “bias”?

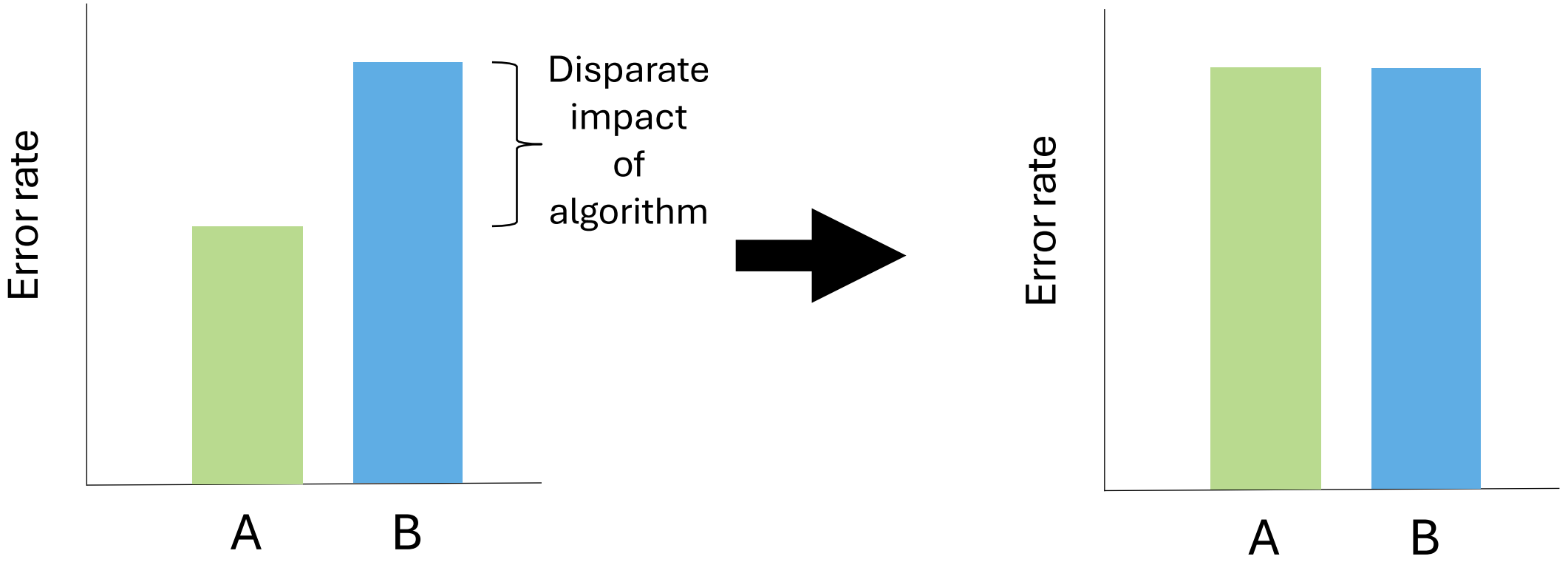
- ~~1)Fairness through unawareness~~ — Not useful
- 2) Group fairness
- 3) Calibration — More standard
- 4) Error rate balance
- 5) Representational fairness
- 6) Counterfactual fairness — More experimental
- 7) Individual fairness

1. COMPAS analysis
2. What is fairness in machine learning?
3. Quantitative definitions of fairness in supervised learning
4. Practical tools for analyzing bias
5. Other curveballs

Tradeoff between accuracy and fairness

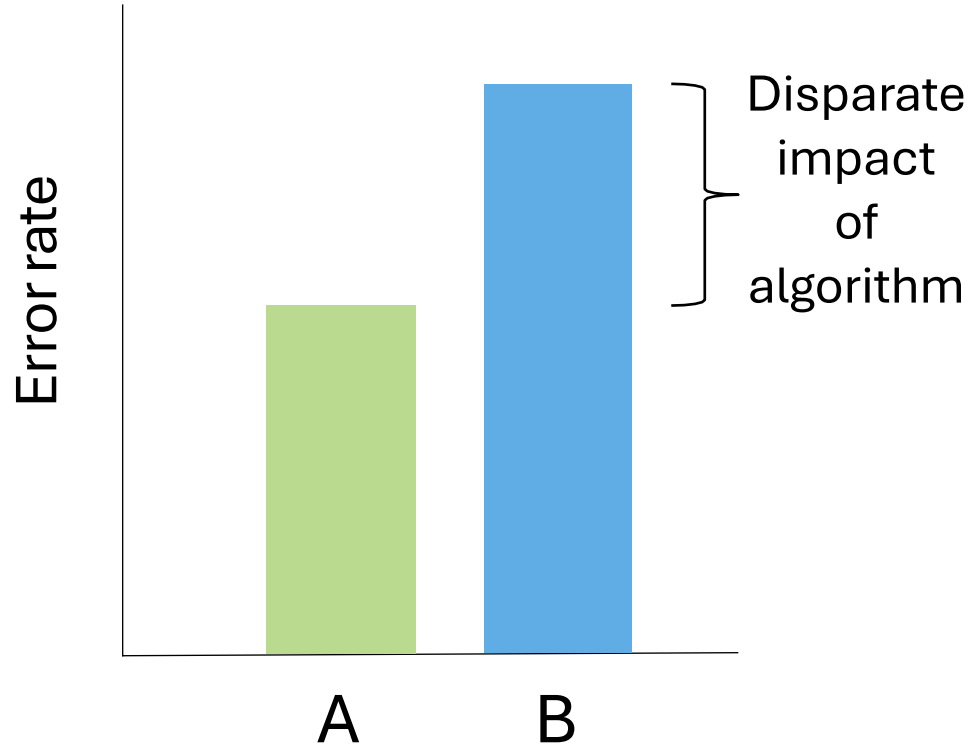


Tradeoff between accuracy and fairness



Consider bias, variance, noise

Bias	How well the model fits the data
Variance	How much the sample size affects the accuracy
Noise	Irreducible error independent of sample size and model



Chen, Johansson, Sontag; NeurIPS 2018

Addressing bias

- Logistic regression: $p(y_i = 1|\mathbf{x}_i, \boldsymbol{\theta}) = \dfrac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}_i}}$

- The optimal coefficient: $\text{minimize} \quad -\sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$

Fairness Constraints: Mechanisms for Fair Classification

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi
Max Planck Institute for Software Systems (MPI-SWS), Germany

Addressing bias

- Logistic regression: $p(y_i = 1|\mathbf{x}_i, \boldsymbol{\theta}) = \dfrac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}_i}}$

- The optimal coefficient: $\text{minimize} \quad -\sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$
- With fairness constraints:

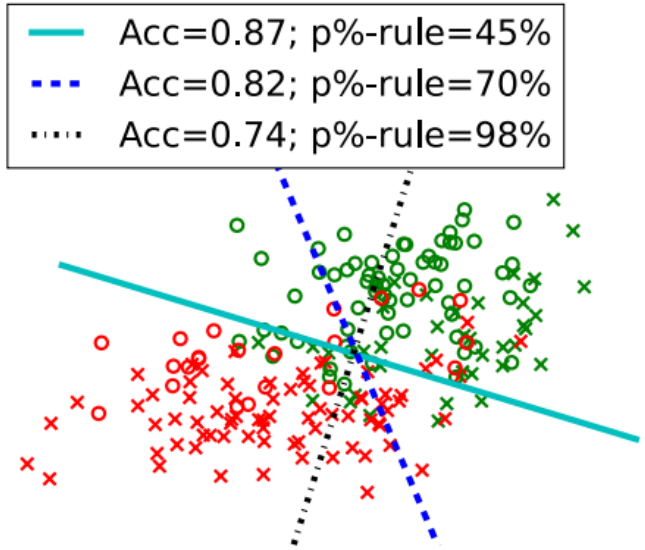
$$\begin{aligned} \text{minimize} \quad & -\sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \\ \text{subject to} \quad & \tfrac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \boldsymbol{\theta}^T \mathbf{x}_i \leq \mathbf{c}, \\ & \tfrac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \boldsymbol{\theta}^T \mathbf{x}_i \geq -\mathbf{c}, \end{aligned}$$

Fairness Constraints: Mechanisms for Fair Classification

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi
Max Planck Institute for Software Systems (MPI-SWS), Germany

Addressing bias

- Generalizes to SVM & other kinds of classifiers that meet some assumptions



Fairness Constraints: Mechanisms for Fair Classification

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi
Max Planck Institute for Software Systems (MPI-SWS), Germany

1. COMPAS analysis
2. What is fairness in machine learning?
3. Quantitative definitions of fairness in supervised learning
4. Practical tools for analyzing bias
5. Other curveballs

The New York Times | https://nyti.ms/38brSto

ECONOMIC VIEW

Biased Algorithms Are Easier to Fix Than Biased People

Racial discrimination by algorithms or by people is harmful — but that's where the similarities end.

By Sendhil Mullainathan

Dec. 6, 2019

Open questions

- How can we build responsible algorithms and datasets?
- For what settings should we use algorithms?
- Can we ever promise an algorithm is “fair”?
- When should we use humans and when should we use algorithms?

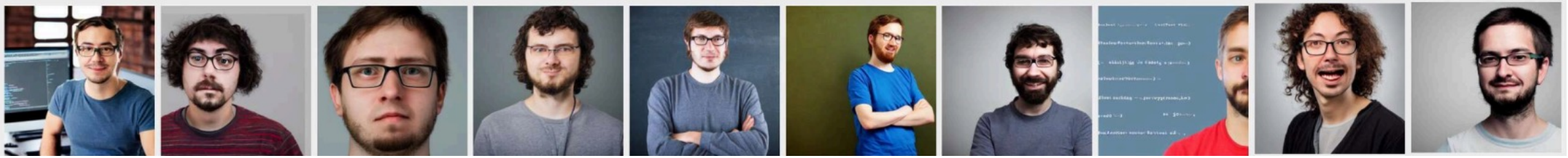
Looking forward

- AI researchers have made great progress **auditing bias** in existing wide-spread algorithms.
- **Formalizing fairness** quantitatively can build fairness constraints directly into high-stakes models.
- Long-term solutions include **growing research community, rethinking datasets, and considering societal impacts**.

Responsible AI

Part 2: Generative tasks

A software developer?



An emotional person?

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2022. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale.

Background:

The psychology of stereotypes in humans

Stereotype

- “Exaggerated belief associated with a category. Its function is to justify (rationalize) our conduct in relation to that category”
  - Alport 1954 (exaggeration)
- Stereotypes are “mental representations of groups and their members” that contain beliefs about group characteristics.
  - Fiske 2000 (cognitive schema)
- A stereotype is a set of beliefs about the personal attributes of a group of people.
  - Schneider 2004 (shared belief systems)

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, *11*(2), 77-83.

Stereotype

- “Exaggerated belief associated with a category. Its function is to justify (rationalize) our conduct in relation to that category”
  - Alport 1954 (exaggeration)
- **Stereotypes are “mental representations of groups and their members” that contain beliefs about group characteristics.**
  - **Fiske 2000 (cognitive schema)**
- A stereotype is a set of beliefs about the personal attributes of a group of people.
  - Schneider 2004 (shared belief systems)

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, *11*(2), 77-83.

The warmth-competence framework

Social perception is structured along two universal dimensions

1) **Warmth** (intentions toward us)
2) **Competence** (ability to enact intentions)

These dimensions organize how we perceive:

- Individuals
- Social groups
- Stereotypes

The model explains consistent patterns in prejudice and social judgment

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, *11*(2), 77-83.

The warmth-competence framework

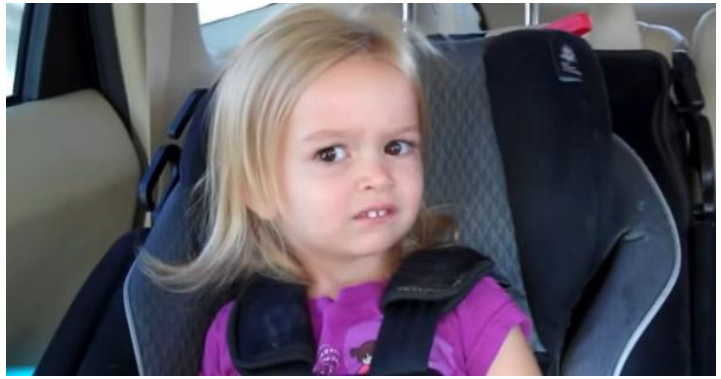
Can I trust you…?

- **Warmth** answers: “Are they friend or foe?”
- Judgments include:
  - **Trustworthiness**
  - **Morality**
  - **Friendliness**
- Warmth is assessed first and fastest in social cognition
- Driven by perceived intentions and competition
- Groups seen as competitive → perceived as less warm

The warmth-competence framework

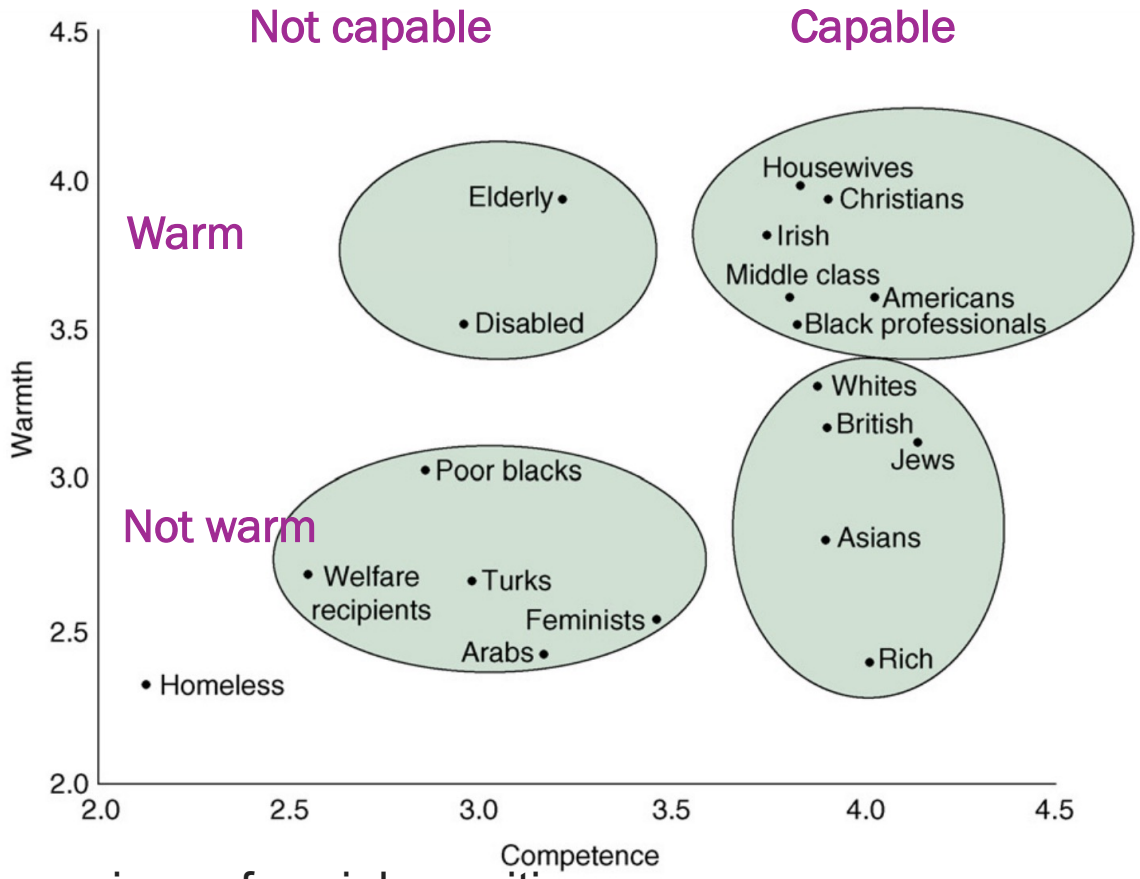
- **Competence** answers: “Can they act on their intentions?”
- Judgments include:
  - **Intelligence**
  - **Skill**
  - **Capability**
- Driven by perceived status
- Higher-status groups → perceived as more competent
- Competence shapes respect and admiration

Do you know what you’re doing…?



The warmth-competence framework

- The 2×2 Stereotype Content Model: Combining warmth & competence yields four quadrants
- **High Warmth / High Competence**
  - Admired groups
- **Low Warmth / High Competence**
  - Envy targets (e.g., rich, elites)
- **High Warmth / Low Competence**
  - Paternalized groups (e.g., elderly)
- **Low Warmth / Low Competence**
  - Marginalized groups

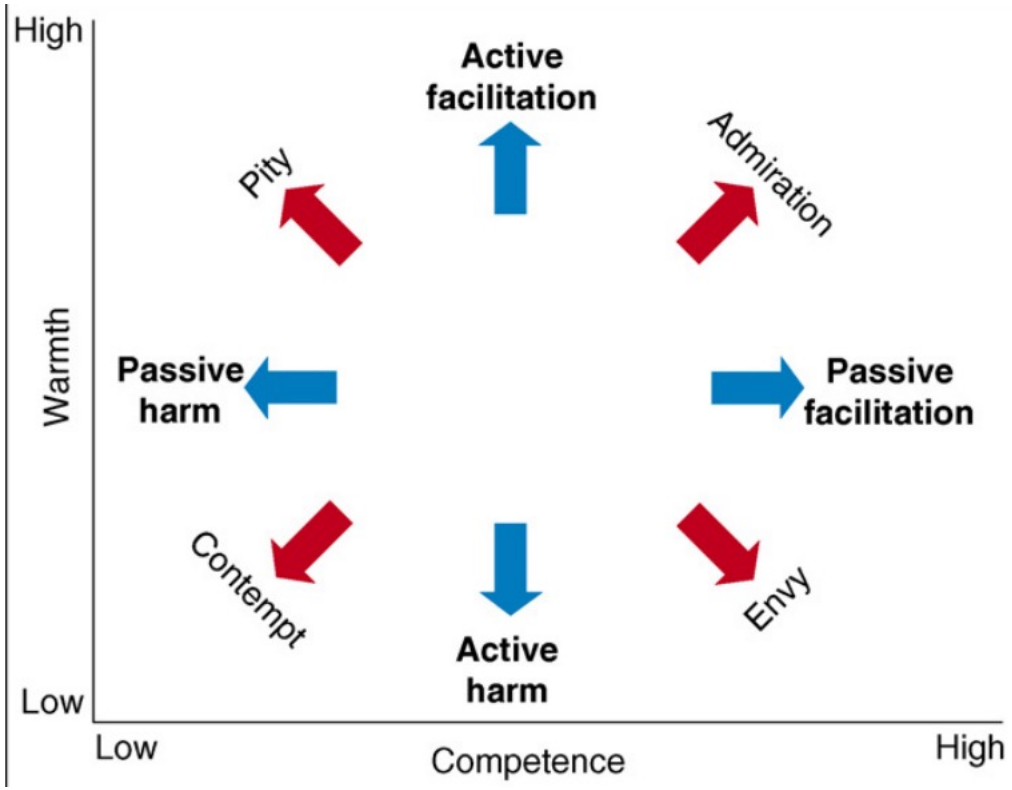


Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, *11*(2), 77-83.

The warmth-competence framework

- The 2×2 Stereotype Content Model: Combining warmth & competence yields four quadrants

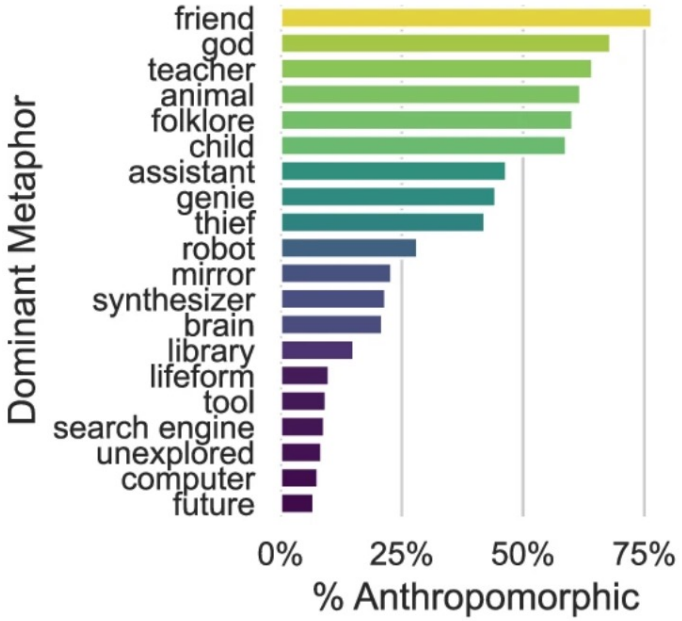
- **High Warmth / High Competence**
  - Admiration
- **Low Warmth / High Competence**
  - Envy
- **High Warmth / Low Competence**
  - Pity
- **Low Warmth / Low Competence**
  - Contempt

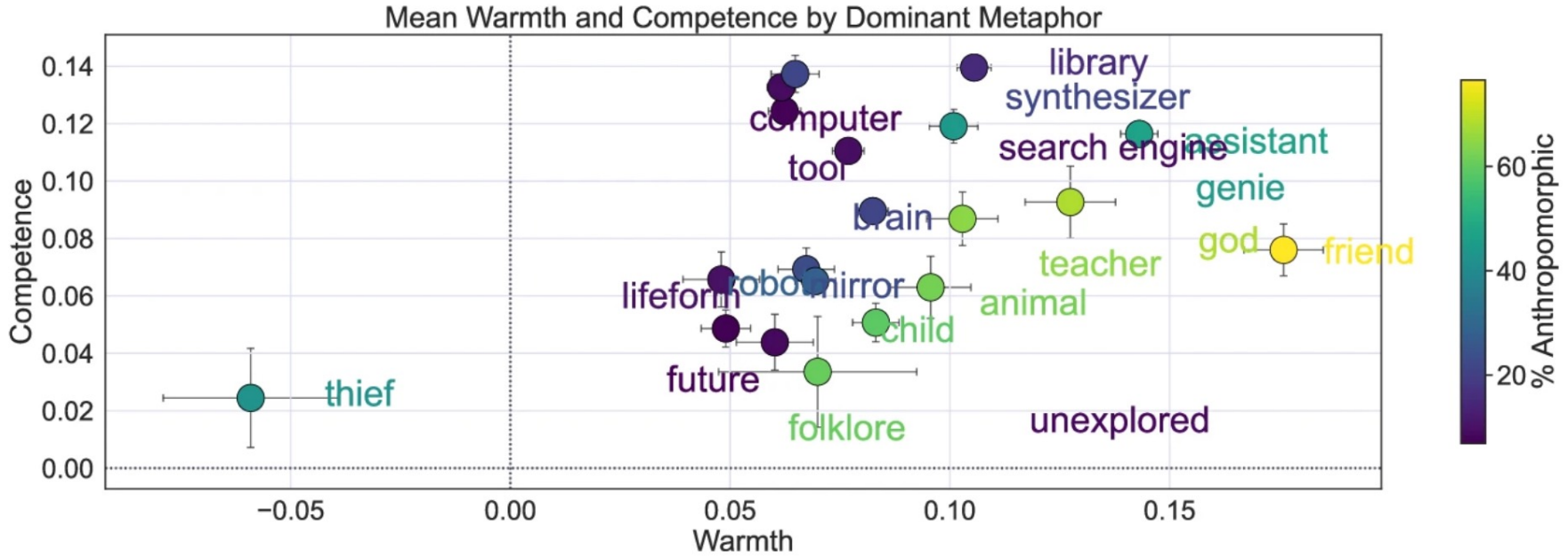


1) People use this framework to evaluate AI

- Humans apply the same social cognition dimensions to AI systems
- Warmth judgments:
  - Is the AI aligned with me?
  - Does it “care”?
  - Is it fair? Is id bad for society?
- Competence judgments:
  - Is it accurate?
  - Is it intelligent?
  - Does it make good decisions?
- **Warmth** often shapes trust and adoption
- **Competence** shapes perceived intelligence and reliability

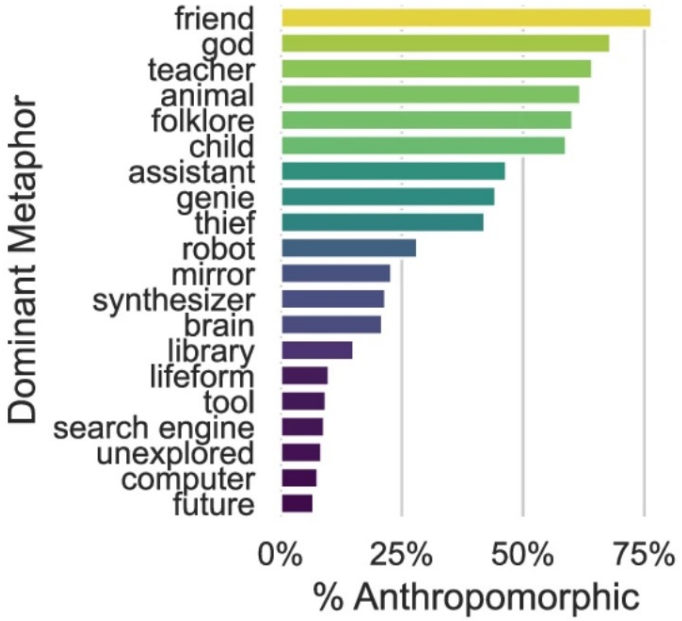
1) People use this framework to evaluate AI

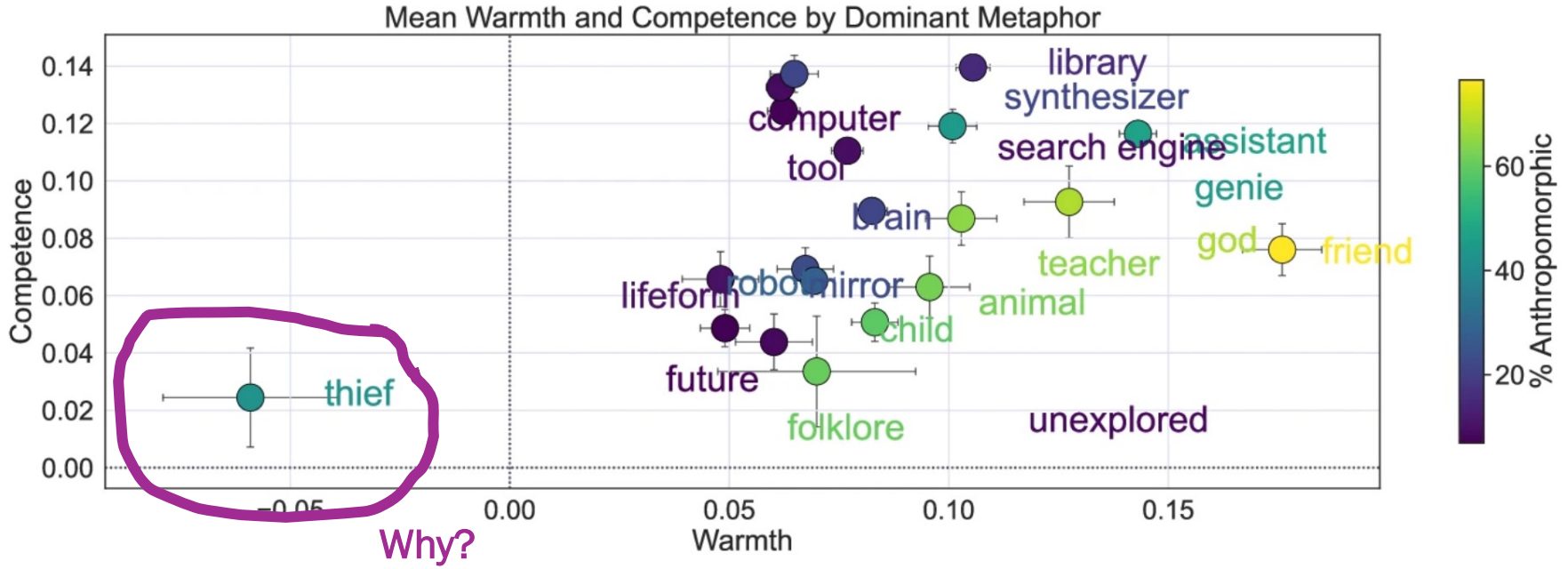




Cheng, M., Lee, A. Y., Rapuano, K., Niederhoffer, K., Liebscher, A., & Hancock, J. (2026). Metaphors of AI indicate that people increasingly perceive AI as warm and human-like. *Communications Psychology*.

1) People use this framework to evaluate AI



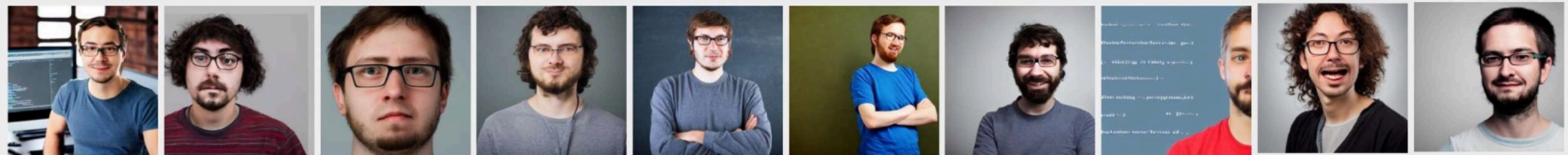


Cheng, M., Lee, A. Y., Rapuano, K., Niederhoffer, K., Liebscher, A., & Hancock, J. (2026). Metaphors of AI indicate that people increasingly perceive AI as warm and human-like. *Communications Psychology*.

2) AI reproduces warmth-competence

- AI systems are trained on human-generated data
- Human stereotypes (warmth & competence patterns) are embedded in that data
- Text-to-image and language models reproduce and amplify these associations
  - Elites, paternalized groups, marginalized groups
- Result: AI does not just reflect social cognition, it amplifies it

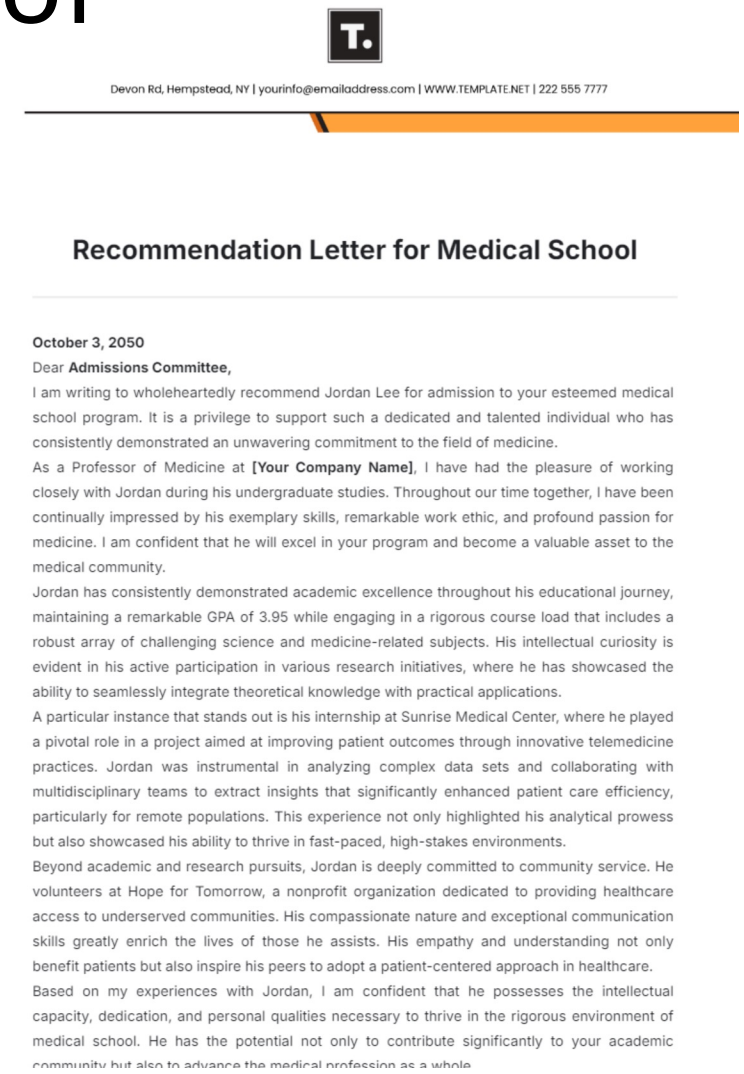
A software developer?



Expressions of stereotypes in **human language**

How this can show up: Letters of recommendation

- Do letters of recommendation differ systematically by gender?
- Context: Medical faculty hiring at a large U.S. medical school
- Focus: Subtle linguistic patterns, not overt discrimination
- The “glass ceiling” may operate through discourse
- Key idea: Stereotypes can be embedded in everyday professional language



T.

Devon Rd, Hempstead, NY | yourinfo@emailaddress.com | WWW.TEMPLATE.NET | 222 555 7777

Recommendation Letter for Medical School

October 3, 2050

Dear **Admissions Committee,**

I am writing to wholeheartedly recommend Jordan Lee for admission to your esteemed medical school program. It is a privilege to support such a dedicated and talented individual who has consistently demonstrated an unwavering commitment to the field of medicine.

As a Professor of Medicine at **[Your Company Name]**, I have had the pleasure of working closely with Jordan during his undergraduate studies. Throughout our time together, I have been continually impressed by his exemplary skills, remarkable work ethic, and profound passion for medicine. I am confident that he will excel in your program and become a valuable asset to the medical community.

Jordan has consistently demonstrated academic excellence throughout his educational journey, maintaining a remarkable GPA of 3.95 while engaging in a rigorous course load that includes a robust array of challenging science and medicine-related subjects. His intellectual curiosity is evident in his active participation in various research initiatives, where he has showcased the ability to seamlessly integrate theoretical knowledge with practical applications.

A particular instance that stands out is his internship at Sunrise Medical Center, where he played a pivotal role in a project aimed at improving patient outcomes through innovative telemedicine practices. Jordan was instrumental in analyzing complex data sets and collaborating with multidisciplinary teams to extract insights that significantly enhanced patient care efficiency, particularly for remote populations. This experience not only highlighted his analytical prowess but also showcased his ability to thrive in fast-paced, high-stakes environments.

Beyond academic and research pursuits, Jordan is deeply committed to community service. He volunteers at Hope for Tomorrow, a nonprofit organization dedicated to providing healthcare access to underserved communities. His compassionate nature and exceptional communication skills greatly enrich the lives of those he assists. His empathy and understanding not only benefit patients but also inspire his peers to adopt a patient-centered approach in healthcare.

Based on my experiences with Jordan, I am confident that he possesses the intellectual capacity, dedication, and personal qualities necessary to thrive in the rigorous environment of medical school. He has the potential not only to contribute significantly to your academic community but also to advance the medical profession as a whole

Trix, F., & Psenka, C. (2003). Exploring the color of glass: Letters of recommendation for female and male medical faculty. Discourse & Society, 14(2), 191-220.

How this can show up: Letters of recommendation

- 312 letters for 89 applicants (medical faculty positions)
- Compared letters for female vs. male candidates
- Qualitative + quantitative discourse analysis
- Coded for:
  - Length
  - Doubt raisers (hedges, faint praise)
  - Gendered language
  - References to personal life

Trix, F., & Psenka, C. (2003). Exploring the color of glass: Letters of recommendation for female and male medical faculty. Discourse & Society, 14(2), 191-220.

How this can show up: Letters of recommendation

- Letters for women were shorter
- Women more often described as:
  - **“hardworking”, “conscientious”, “reliable”, “organized”**
- Men more often described as:
  - **“brilliant”, “outstanding”, “leader”**
- Women’s letters more likely to include:
  - Personal life references
  - Emotional descriptors
- Men’s letters emphasized research and leadership

Trix, F., & Psenka, C. (2003). Exploring the color of glass: Letters of recommendation for female and male medical faculty. Discourse & Society, 14(2), 191-220.

How this can show up: Letters of recommendation

- “Doubt raisers” more frequent in letters for women:
  - Hedges (“She might make a good…”)
  - Comparisons (“She is among the better…”)
  - Irrelevant disclaimers
- “Faint praise” more common for women
  - Women more often portrayed as students/mentees
  - Men more often framed as colleagues/future leaders
- Bias operates through **stereotypes**, not explicit **negativity**
- Broader lesson: Human language reflects stereotypes

Trix, F., & Psenka, C. (2003). Exploring the color of glass: Letters of recommendation for female and male medical faculty. Discourse & Society, 14(2), 191-220.

How this can show up:
Letters of recommendation

Adjectives to avoid:	Adjectives to include:
caring	successful
compassionate	excellent
hard-working	accomplished
conscientious	outstanding
dependable	skilled
diligent	knowlegeable
dedicated	insightful
tactful	resourceful
interpersonal	confident
warm	ambitious
helpful	independent
	intellectual

https://csw.arizona.edu/sites/default/files/avoiding_gender_bias_in_letter_of_reference_writing.pdf

Computational methods for **identifying** stereotypes

Discovering differences between groups

- How do we know what words or phrases to look for? How do we know what group-specific biased language might be there?
- How do we know what language reflects stereotypes?
  - Either in **human language** or **AI outputs?**

A Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16(4), 372-403.

Computational methods

- **Goal:** Identify words that distinguish two corpora.
- Compares word usage across two text groups
- Uses **weighted log-odds ratio with informative prior**
- **Outputs**:
  - Direction of association and magnitude
  - Statistical significance (z-score)
- **Case Study Setup**
  - Corpus A: Letters for **male candidates**
  - Corpus B: Letters for **female candidates**
  - Prior P: Baseline corpus (all letters combined or historical letters)
- **The question:** Which words are disproportionately associated with male vs female candidates?

A Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16(4), 372-403.

Computational methods

Intuition: Log-Odds Comparison

- We compare how strongly a word is associated with one corpus relative to another
- For each word w:

$$\delta_{w,A} = \frac{\log \frac{L_{w,A}}{L_{w,B}}}{\sqrt{\frac{1}{N_{w,A} + N_{w,P}} + \frac{1}{N_{w,B} + N_{w,P}}}}$$

where:

  - $N_{w,A}$: count of word $w$in male letters
  - $N_{w,B}$: count of word $w$in female letters
  - $N_{w,P}$: count in prior corpus
- **Interpretation**
  - $\delta_{w,A} > 0$: word associated with male candidates, $\delta_{w,A} < 0$: word associated with female candidates
- Larger magnitude → stronger distinction
- Denominator adjusts for variance (rare words penalized)

A Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16(4), 372-403.

Computational methods

- **Likelihood with informative prior:** likelihoods are smoothed using the prior:
  - $L_{w,A} = \frac{N_{w,A}+N_{w,P}}{\sum_{x\in A}(N_{x,A}-N_{w,A})+\sum_{x\in P}(N_{x,P}-N_{w,P})}$
  - $L_{w,B} = \frac{N_{w,B}+N_{w,P}}{\sum_{x\in B}(N_{x,B}-N_{w,B})+\sum_{x\in P}(N_{x,P}-N_{w,P})}$
- **Why use a prior?**
  - Prevents rare words from dominating
  - Anchors estimates to baseline frequency
  - Reduces instability in small samples
- **In practice:**
  - Remove stop words
  - Compute log-odds
  - Convert to z-scores & **p-values**

A Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16(4), 372-403.

Computational methods

- **Case Study: Gendered Language in Letters**

Words Overrepresented in Male Letters (δ > 0)

"brilliant"
"leader"
"visionary"
"independent"
"trailblazing"

VS.

Words Overrepresented in Female Letters (δ < 0)

"hardworking"
"supportive"
"kind"
"collaborative"
"dedicated"

- **Pattern**
  - Male candidates → *agentic traits*
  - Female candidates → *communal traits*

A Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16(4), 372-403.

Computational methods

Why this matters for responsible AI and AI evaluation: measurement enables accountability!

- Weighted log-odds controls for frequency, allowing precise detection of linguistic bias
- Without rigorous measurement, we cannot evaluate or mitigate bias neither human language, nor AI systems

A Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16(4), 372-403.

Responsible AI

Part 3: Representations

Risks of stereotypes in **AI that learns from human language**

Representing meaning of words

What do words mean? How do they get their meaning?

cat

dog

tiger

table

Perhaps more pertinent for language technology

How can we represent the meaning of words in a form that is computationally flexible?

The company words keep

The Distributional Hypothesis: Words that occur in the same contexts have similar meanings (e.g. Zellig Harris, J.R. Firth)

Firth (1957): “You shall know a word by the company it keeps”

The key idea: To characterize the meaning of a word, we need to characterize the distribution of its context

What context? Commonly interpreted as neighboring words in text, but could be syntactic, semantic, discourse, pragmatic,…

Symbolic vs. Distributed representations

The strings `cat`, `tiger`, `dog` and `table` are symbols

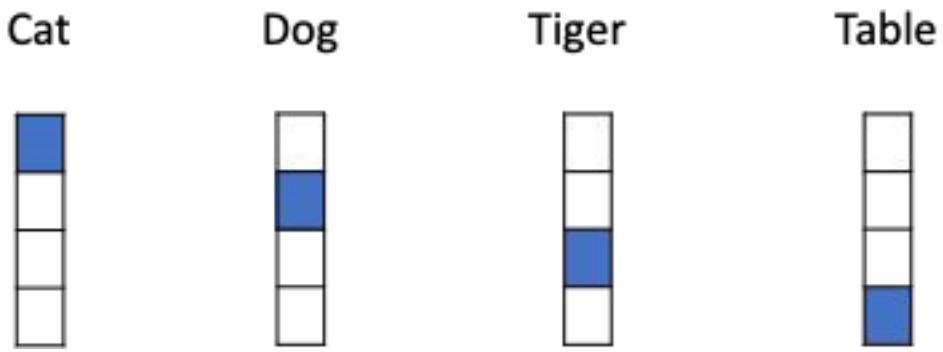
Just knowing the symbols does not tell us anything about what they mean.

1. `cat` and `tiger` are conceptually closer to each other than to `dog` or `table`
2. `cat`, `tiger` and `dog` are closer to each other than `table`

We need a representation that captures similarities between similar objects

Symbolic vs. Distributed representations

Think about feature representations



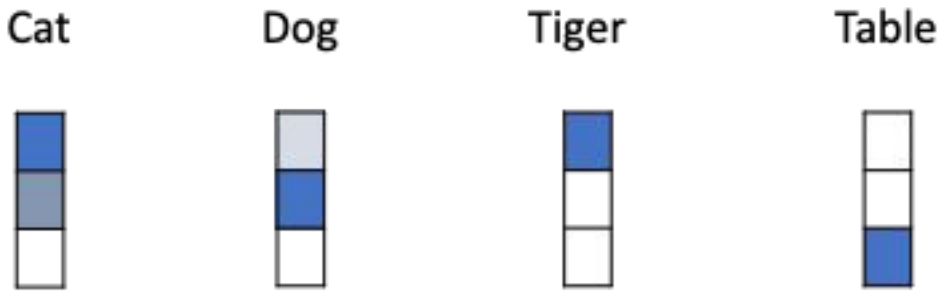
These one-hot vectors do not capture inherent similarities

Distances or dot products are all equal

Symbolic vs. Distributed representations

Distributed representations capture concept similarities better

Vector valued representations that coalesce superficially distinct concepts



Example intrinsic evaluation: Word Analogies

Complete a word analogy puzzle using the embeddings

`Queen : King :: Tigress : ?`

Word embeddings are great, but…

Societal biases in word embeddings

- If word embeddings capture distributional information from corpora…

…and corpora possess societal stereotypes, then

the trained word embeddings may encode these stereotypes

“Bias” in language technology

A fast moving field, with new techniques and perspectives being introduced almost every month

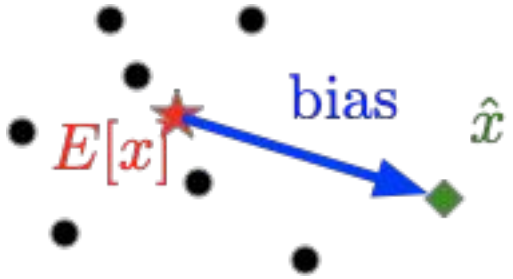
Two related lines of work:

1. New methods for quantifying biases encoded in embeddings
2. Methods for removing biases from embeddings

What is "bias"?

Def: difference between an estimator and its expected value

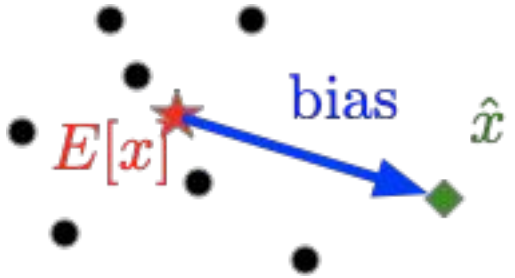
$$\hat{x} - E[x]$$



What is "bias"?

Def: difference between an estimator and its expected value

$$\hat{x} - E[x]$$

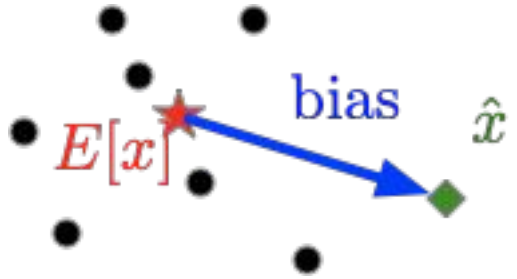


Def: an instance of prejudice, especially a personal and sometimes unreasonable outlook

What is "bias"?

Def: difference between an estimator and its expected value

$$\hat{x} - E[x]$$



Def: an oversimplified view or prejudiced attitude of a particular type of person or thing (**a concept**)

What is bias and a stereotype

An oversimplification of a concept

Ex: children are curious Ex:

dogs are friendly

Ex: nurses are women and doctors are men

Often a **negative** connotation

Bias + Machine Learning

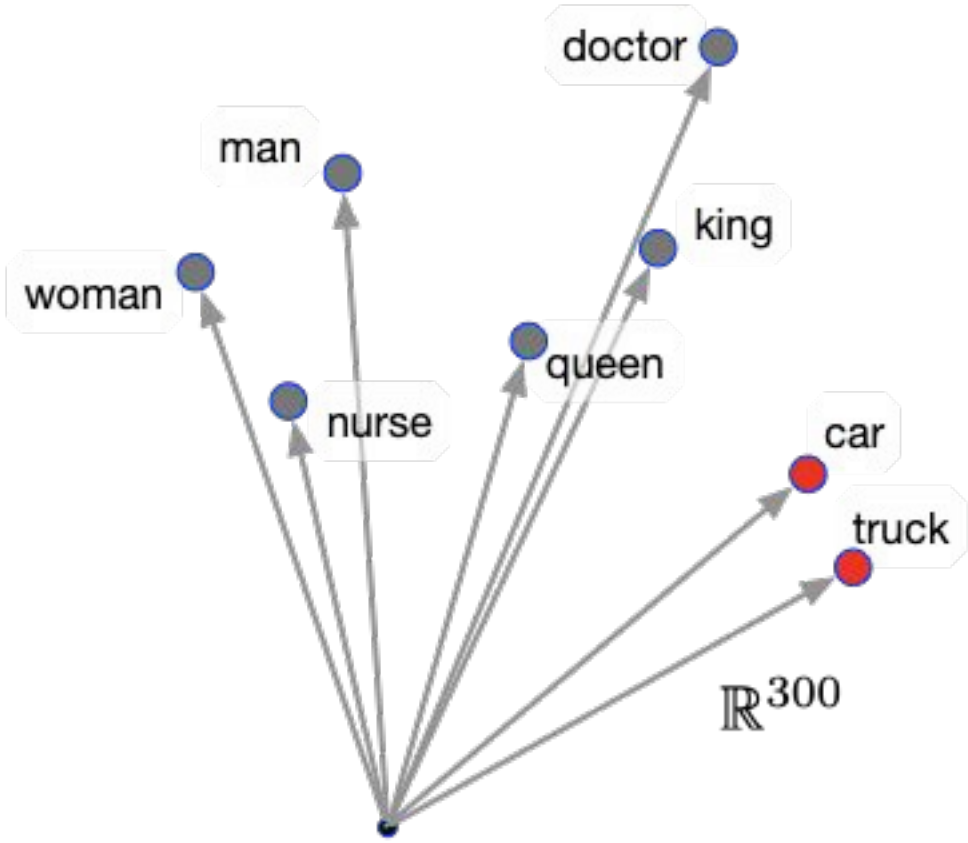
Given bias

- Choice of data
- Mechanism to represent data
- Choice of learning model / algorithm

...can translate into representational or allocational harm

Inspecting Representations

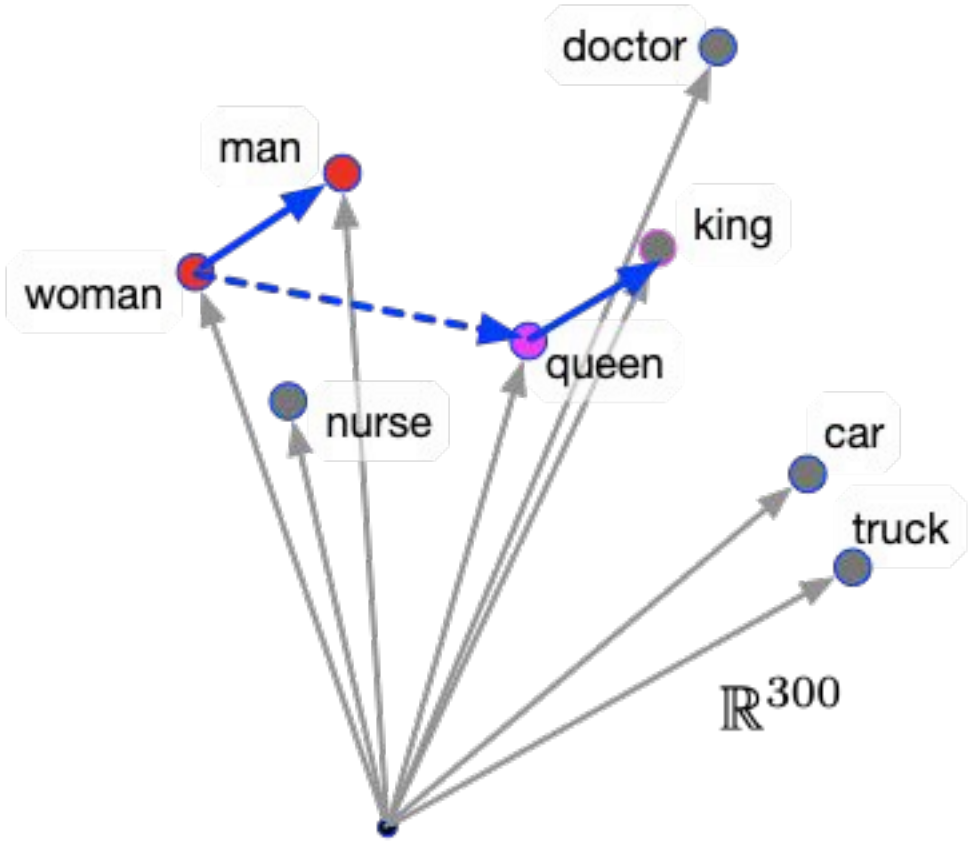
Similarity Tests



Inspecting Representations

Similarity Tests

Analogies



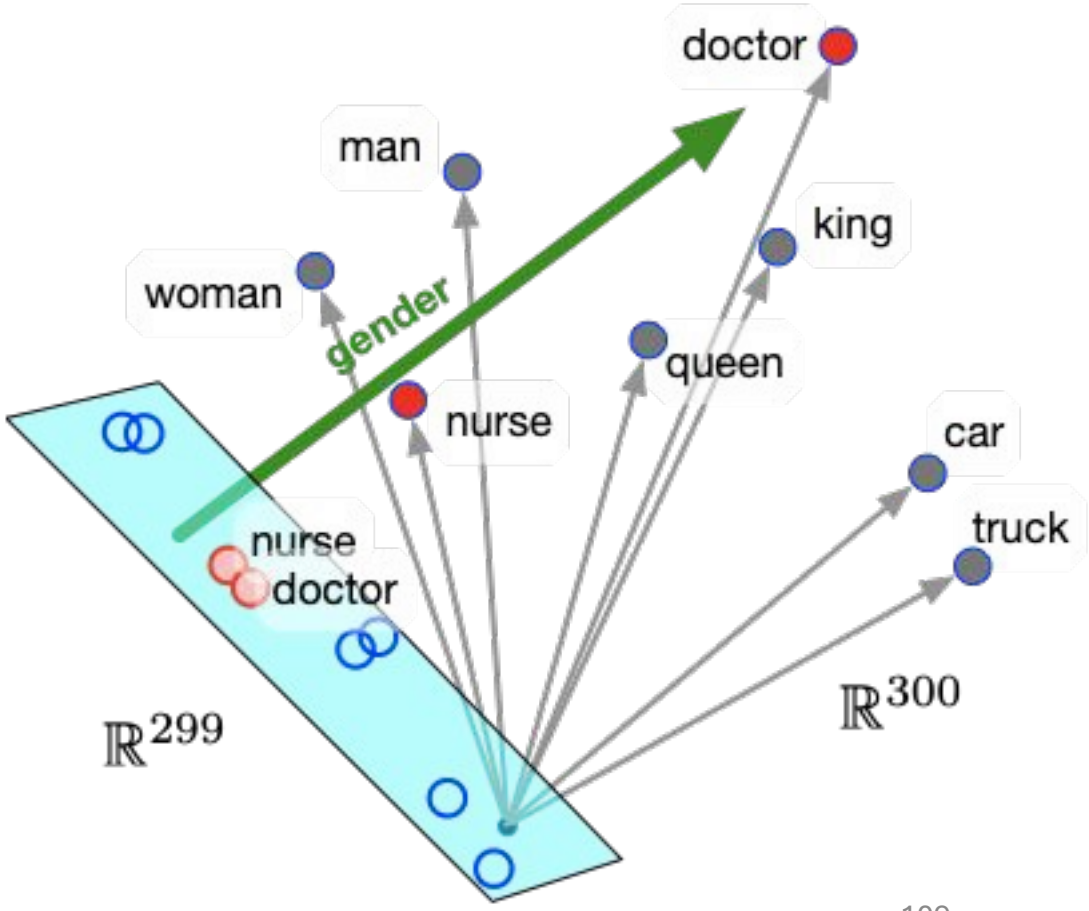
Bolukbasi et al; Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NeurIPS 2016

Inspecting Representations

Similarity Tests

Analogies

Concept Subspace



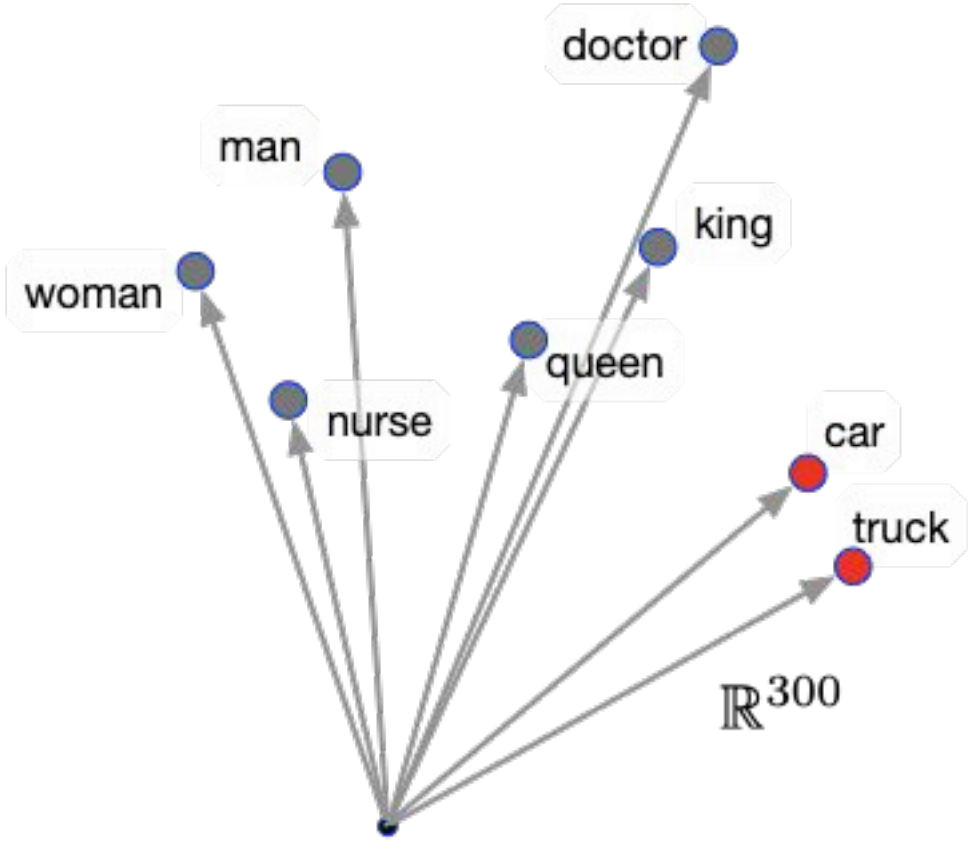
Inspecting Representations

Similarity Tests

Analogies

Concept Subspace

WEAT (implicit gender association stereotypes)



Caliskan et al; Semantics derived automatically from language corpora contain human like biases. Science 2017

Inspecting Representations

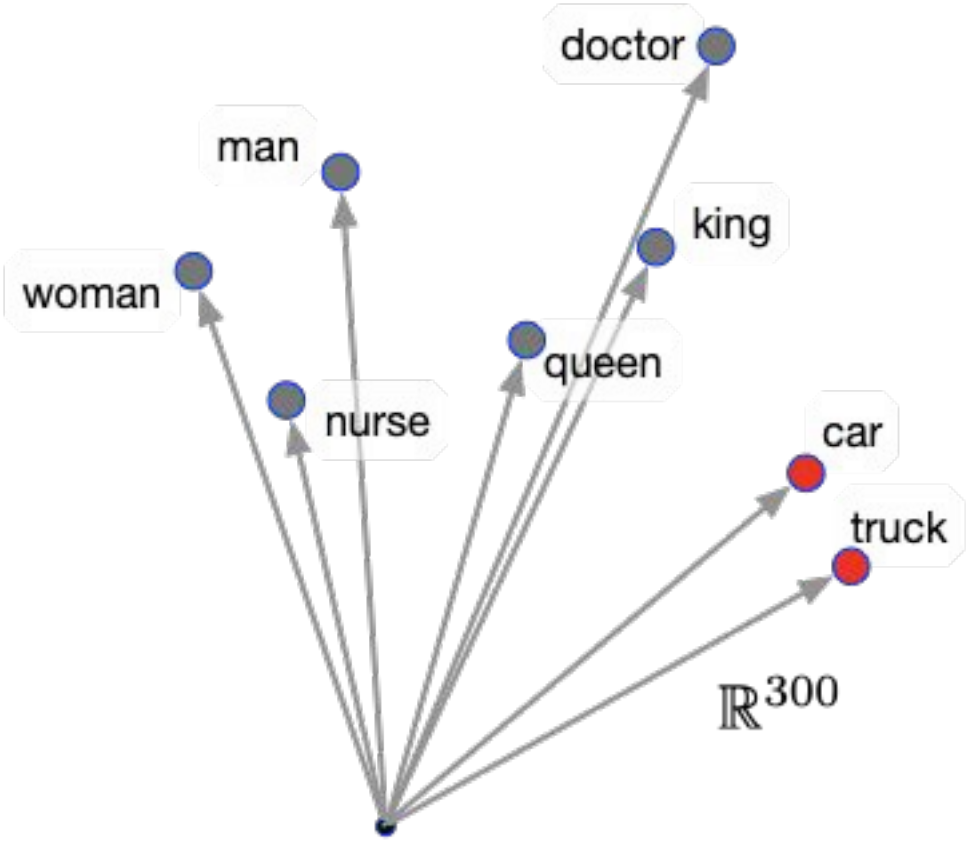
Similarity Tests

Analogies

Concept Subspace

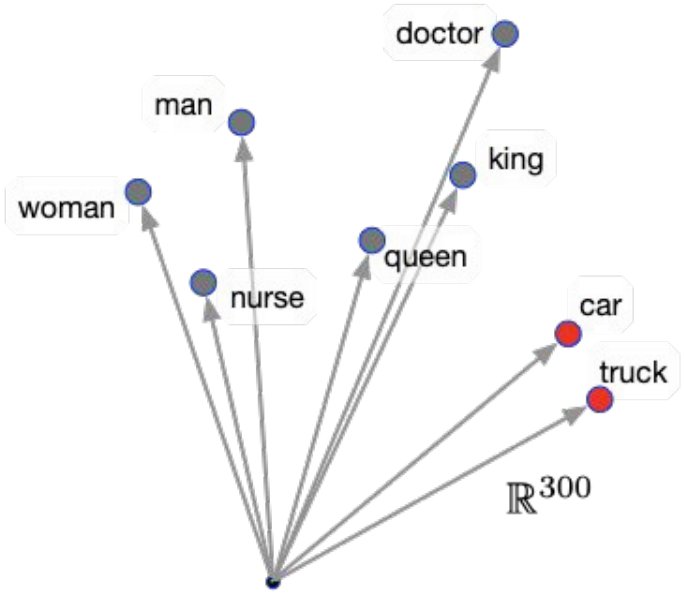
WEAT (implicit gender association stereotypes)

ECT



Implicit Association Test

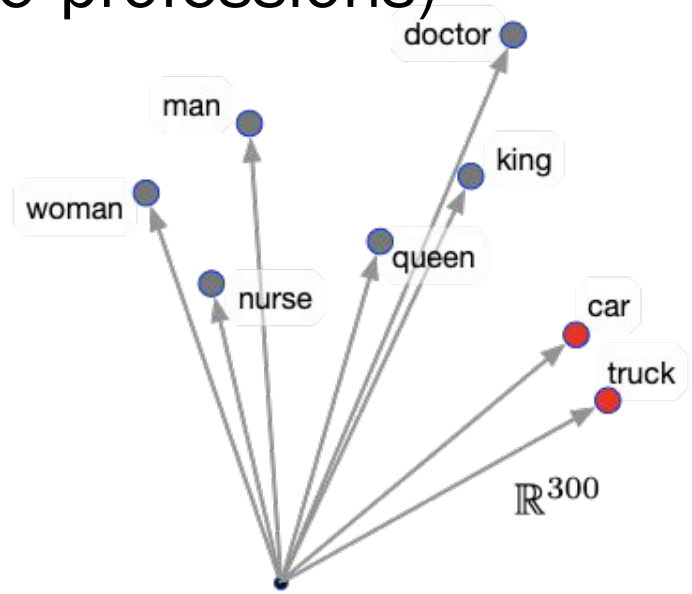
- X = {man, male, ...} (definitionally male words)
- Y = {woman, female, ...} (definitionally female words)



Caliskan et al; Semantics derived automatically from language corpora contain human like biases. Science 2017

Implicit Association Test

- X = {man, male, ...} (definitionally male words)
- Y = {woman, female, ...} (definitionally female words)
- A = {programmer, engineer, scientist, ...} (stereotypical male professions)
- B = {nurse, teacher, librarian, ...} (stereotypical female professions)



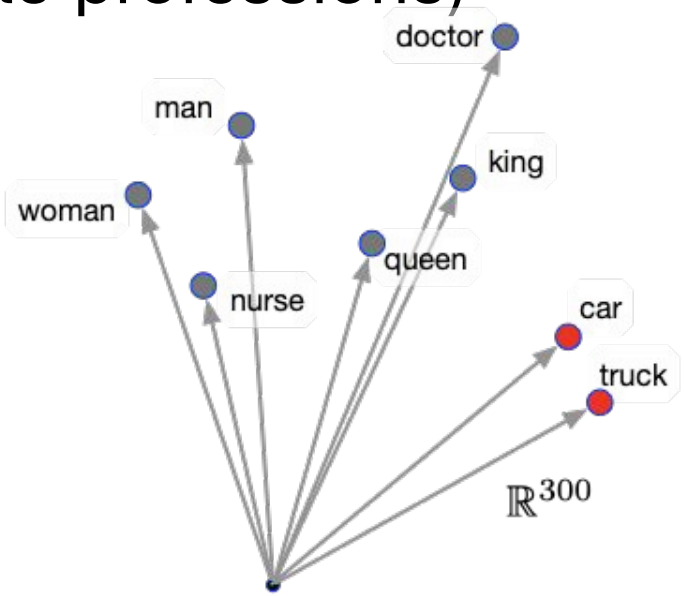
Caliskan et al; Semantics derived automatically from language corpora contain human like biases. Science 2017

Implicit Association Test

- X = {man, male, ...} (definitionally male words)
- Y = {woman, female, ...} (definitionally female words)
- A = {programmer, engineer, scientist, ...} (stereotypical male professions)
- B = {nurse, teacher, librarian, ...} (stereotypical female professions)

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(a, w) - \frac{1}{|B|} \sum_{b \in B} \cos(b, w)$$

association of gendered word w with sets A,B



Caliskan et al; Semantics derived automatically from language corpora contain human like biases. Science 2017

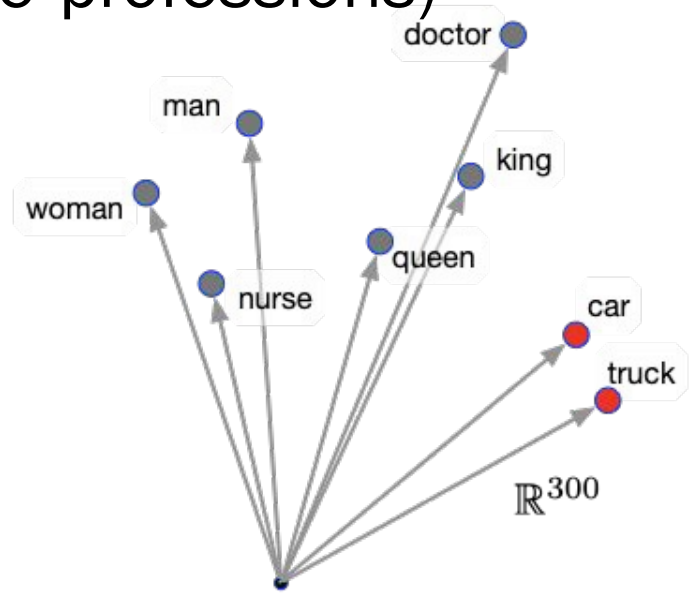
Implicit Association Test

- X = {man, male, ...} (definitionally male words)
- Y = {woman, female, ...} (definitionally female words)
- A = {programmer, engineer, scientist, ...} (stereotypical male professions)
- B = {nurse, teacher, librarian, ...} (stereotypical female professions)

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(a, w) - \frac{1}{|B|} \sum_{b \in B} \cos(b, w)$$

association of gendered word w with sets A,B

$$S(X, Y, A, B) = \frac{1}{|X|} \sum_{x \in X} s(x, A, B) - \frac{1}{|Y|} \sum_{y \in Y} s(y, A, B)$$



Caliskan et al; Semantics derived automatically from language corpora contain human like biases. Science 2017

Implicit Association Test

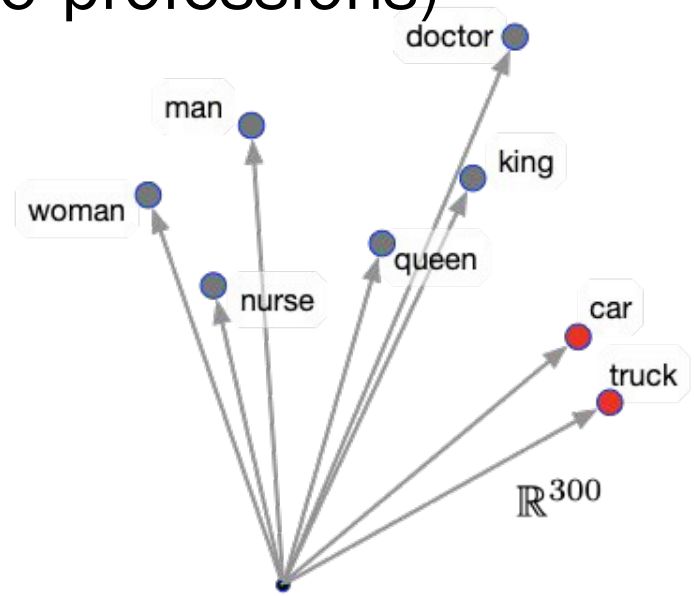
- X = {man, male, ...} (definitionally male words)
- Y = {woman, female, ...} (definitionally female words)
- A = {programmer, engineer, scientist, ...} (stereotypical male professions)
- B = {nurse, teacher, librarian, ...} (stereotypical female professions)

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(a, w) - \frac{1}{|B|} \sum_{b \in B} \cos(b, w)$$

association of gendered word w with sets A,B

$$S(X, Y, A, B) = \frac{1}{|X|} \sum_{x \in X} s(x, A, B) - \frac{1}{|Y|} \sum_{y \in Y} s(y, A, B)$$

S in [-2,2]. Neutral should be 0. Word2Vec = 1.89; GloVe 1.81



Word embeddings quantify 100 years of gender and ethnic stereotypes

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou Authors Info & Affiliations

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 12, 2018 (received for review November 22, 2017)

April 3, 2018 | 115 (16) E3635-E3644 | https://doi.org/10.1073/pnas.1720347115

What’s next?

- **Beyond representational: tangible harms for users**
- **Studying AI use**
- **Benchmarking & evaluation**
- **Over-reliance**
- **Erosion of skills**
- **Regulating and mitigating effects on:**
  - **Environment, job markets, creative work, relationships, etc.**

What’s next?

- **Lab tutorial putting these things into practice!**
- **Now that we see why we need to be responsible:**
  - **Part 2: LLM Evaluation (Delip Rao)!**

References

- Thanks to Yulia Tsvetkov and Dan Jurafsky for sharing their AI Ethics class materials!
- Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Journal of personality and social psychology, 114(2), 246. (Introduction only)
- Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., ... & Caliskan, A. (2023, June). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In Proceedings of the 2023 ACM conference on fairness, accountability, and transparency (pp. 1493-1504).
- Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36-45.
- A Visual Tour of Bias Mitigation Techniques for Word Representations (Part of KDD 2021 Tutorials)
- A Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16(4), 372-403.
- Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. Trends in cognitive sciences, 11(2), 77-83.
- Trix, F., & Psenka, C. (2003). Exploring the color of glass: Letters of recommendation for female and male medical faculty. Discourse & Society, 14(2), 191-220.
- Some of the materials adapted from slides by Irene Y. Chen (MIT) and Berkeley's CS 294: Fairness in Machine Learning and N[eur]IPS2017 by Solon Barocas (Cornell) and Moritz Hardt (Berkeley)
- Angwin, Julia, et al. "Machine bias." 2022. 254-264.
  - Method
  - Code