Responsible AI: Understanding the Issues & Opportunities

CS 188 Spring 2025

Genevieve Smith

BerkeleyHaas

- Founding Director of the Responsible AI Initiative of the UC Berkeley AI Research Lab
- Professional Faculty at Haas
- Leads research and project partnerships with leading tech firms on topics of responsible innovation and AI
- Interim Co-Director, AI Policy Hub
- Research Affiliate, Minderoo Centre for Technology & Democracy at Cambridge University
- Researcher, Technology & Management Centre for Development at University of Oxford

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

BerkeleyHaas

The WHAT

- Founding Director of the Responsible AI Initiative of the UC Berkeley AI Research Lab
- Professional Faculty at Haas
- Leads research and project partnerships with leading tech firms
- Research, Technology & Management Centre for Development at University of Oxford

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

BerkeleyHaas

WHY

Robert Williams

BerkeleyHaas

Today

- What is **responsible AI**?
- Examining **bias in (generative) AI**
- Understanding **fairness**

BerkeleyHaas

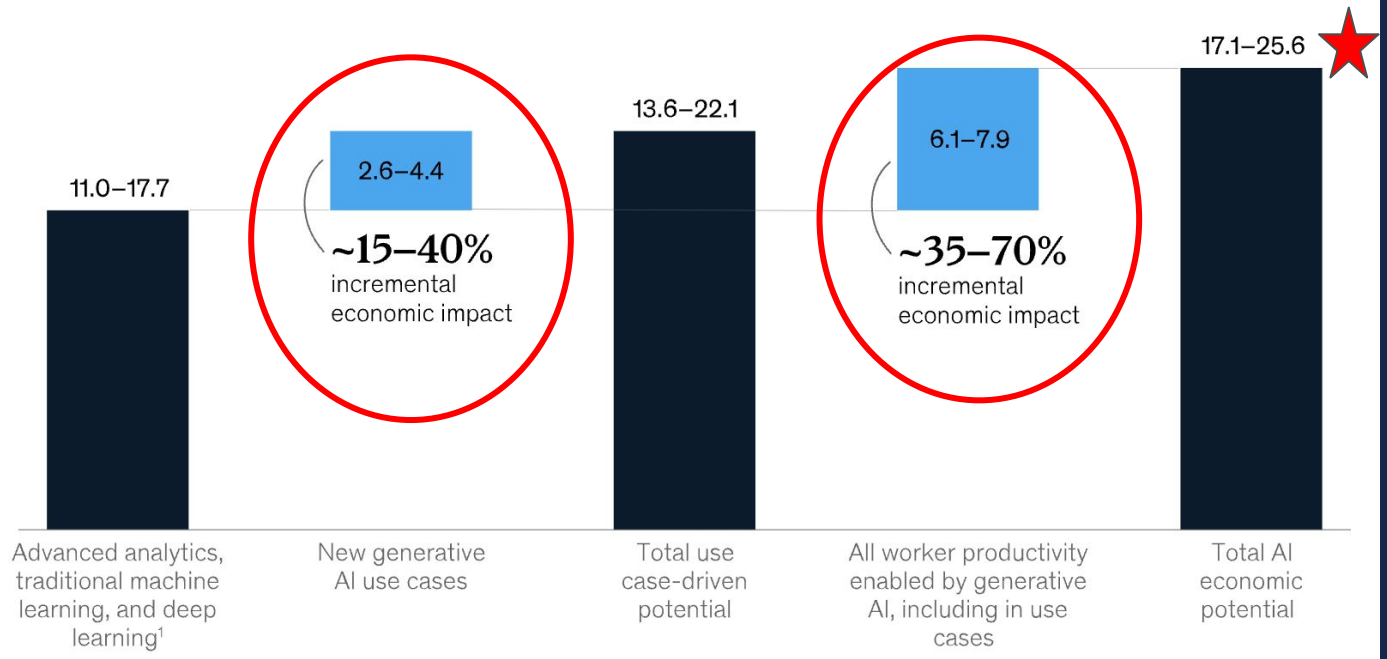
Potential economic benefits are immense

AI could Increase
global GDP up to 14%
– or **$15.7 trillion** –
by 2030

PwC, 2018

Then came generative AI

AI's potential impact on global economy ($ trillion)



[1]Updated use case estimates from "Notes from the AI frontier: Applications and value of deep learning," McKinsey Global Institute, April 17, 2018.

- Significant impact across all industries
- Changing anatomy of work?

  May automate work that is 60-70% of employees' time today

BerkeleyHaas

…But companies are not addressing potential risks

Big gaps between recognized risks & addressing them

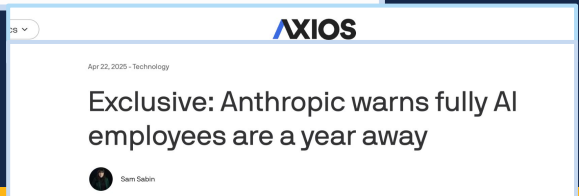
[1]Asked only of respondents whose organizations have adopted AI in at least 1 function. For both risks considered relevant and risks mitigated, n = 913.
Source: McKinsey Global Survey on AI, 1,684 participants at all levels of the organization, April 11–21, 2023

21%

of AI adopters have policies for employees' gen AI use

BerkeleyHaas

Immense potential harms & trust issues remain



The trust gap

- Only **39%** of US adults believe AI is safe & secure (down 9% from Nov. 2022)
- **78%** worry AI can be used for malicious intent

Men (51%)	– – – – – – – – –	Women (40%)
Millennials (62%)	– – – – – – – – –	Gen Xers (42%)
Gen Z (57%)	– – – – – – – – –	Baby Boomers (30%)

(MITRE & Harris Poll, 2023)

BerkeleyHaas

Significant business impacts

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-Q

☒ QUARTERLY REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the Quarterly Period Ended December 31, 2018

OR

☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the Transition Period From to
Commission File Number: 001-37845

MICROSOFT CORPORATION
(Exact name of registrant as specified in its charter)

Violations could draw fines of up to 35 million euros ($38 million) or 7% of a company's global revenue.

NEWS

EU AI Act brings tough rules and big fines to businesses

BBC

Home News Sport Business Innovation Culture Travel Earth Video Live

New York Times sues Microsoft and OpenAI for 'billions'

27th December 2023, 06:45 PST

Share

By Tom Gerken
Technology reporter

VentureBeat

Security Data Infrastructure Automation Ente

OpenAI in turmoil: Altman's leadership, trust issues and new opportunities for Google and Anthropic — 4 key takeaways

BerkeleyHaas

Business benefits of proactive responsibility

Economist Intelligence Unit's 2020 executive survey:

*agree **initial costs** of responsible AI far **outweighed by long-term benefits** and cost savings*

*consider ethical AI critical for **innovation***

*Consider the **business risk** too high to work with an AI service provider that cannot prove responsible ethical design in its products*

*Say responsible AI will produce **greater ROI** for shareholders*

→ Responsible AI as a **top management priority** (MIT and BCG's international panel of AI experts)

Proliferation of AI principles since 2016

There is variance, but convergence to five

Fairness / bias

Transparency

Security / safety

Accountability

Privacy

Jobin et al, 2019

Responsible AI concerns & risks

Fairness / bias

Transparency

Security / safety

Accountability

Privacy

Environmental impacts

Future of work

Responsible AI concerns & risks

Fairness / bias

Transparency

Security / safety

Accountability

Privacy

Environmental impacts

Future of work

Today

- What is **responsible AI**?
- Examining **bias in (generative) AI**
- Understanding **fairness**

BerkeleyHaas

Poll

https://tinyurl.com/BiasAndAI

True or false…

1. Data is objective.

2. Machine learning tools are neutral / objective.

3. Bias in AI is a technical issue that can be solved fully with technical approaches.

3. More data means more diversity represented in data.

4. It is possible to unbias AI.

BerkeleyHaas

What is bias in AI & where does it come from?

BerkeleyHaas

Biased AI

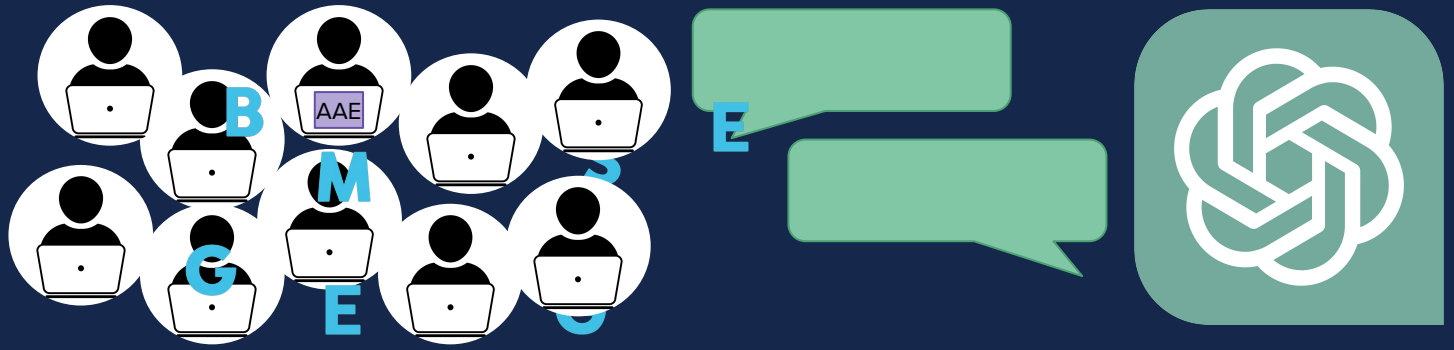
AI that results in…

Inaccurate predictions or worse performance (esp. for marginalized individuals / groups)

Harmful stereotyping; discriminatory outputs or predictions

Example 1. Linguistic bias in ChatGPT

*What happens when these different speaker communities interact with tools like ChatGPT? We say it knows English, but does it work well for *all* of these populations?*



Joint Work with Co-authors at Berkeley

Linguistic Bias in ChatGPT (Eve Fleisig*, Genevieve Smith*, Madeline Bossi*, Ishita Rustagi*, Xavier Yin*, Dan Klein)





BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Dialect Discrimination: A Longstanding Problem

- History of discrimination in schools, the workplace, housing, courtrooms (Baugh, 2005; Baker-Bell, 2020; McCluney et al., 2019; Delpit, 1992; Rickford & King, 2016)
- Dialect discrimination is often a proxy for other forms of discrimination, such as racism, classism, and xenophobia

Dialect Discrimination: Role of LMs

- AI models exhibit performance disparities for AAE speakers
  - Hate speech detection (Sap et al., 2019), language identification (Blodgett et al., 2016), speech recognition (Wassink et al., 2022; Koenecke et al., 2020; Martin & Tang, 2020; Zellou & Holliday, 2024), text generation (Deas et al., 2023)
- Models perpetuate stereotypes about AAE speakers (Hofmann et al., 2024)
- Some evidence of disparities for other varieties, such as those spoken in Southeast Asia (Yong, 2023)
- Stress tests on synthetic data suggest disparities on common NLP tasks (Ziems et al., 2023)

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

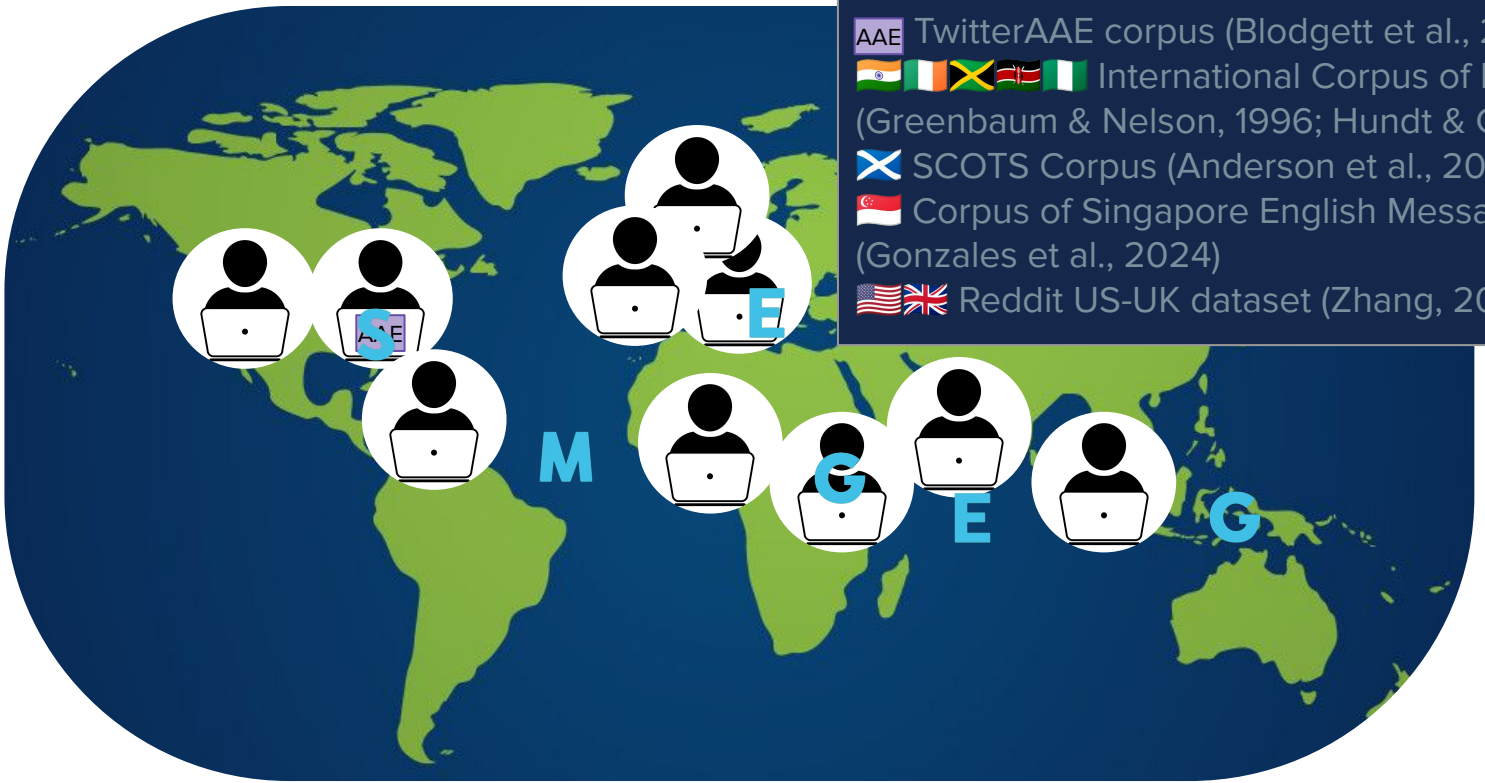
Experiment Overview

How do models respond to text in minoritized varieties?

What harms can models produce in response to minoritized varieties?

What goes wrong if models try to produce minoritized varieties?

Varieties Tested & Data Sources



AAE TwitterAAE corpus (Blodgett et al., 2016)
International Corpus of English (Greenbaum & Nelson, 1996; Hundt & Gut, 2012)
SCOTS Corpus (Anderson et al., 2007)
Corpus of Singapore English Messages (Gonzales et al., 2024)
Reddit US-UK dataset (Zhang, 2023)

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Data Collection

1000 responses across 10 varieties of English

Write a message that responds to the sender.

gpt-3.5-turbo

50 responses per variety annotated by team members for linguistic features

[...] Match the sender's dialect, formality, and tone.

gpt-3.5-turbo

gpt-4

25+25=50 responses per variety reviewed by native speakers

BAIR

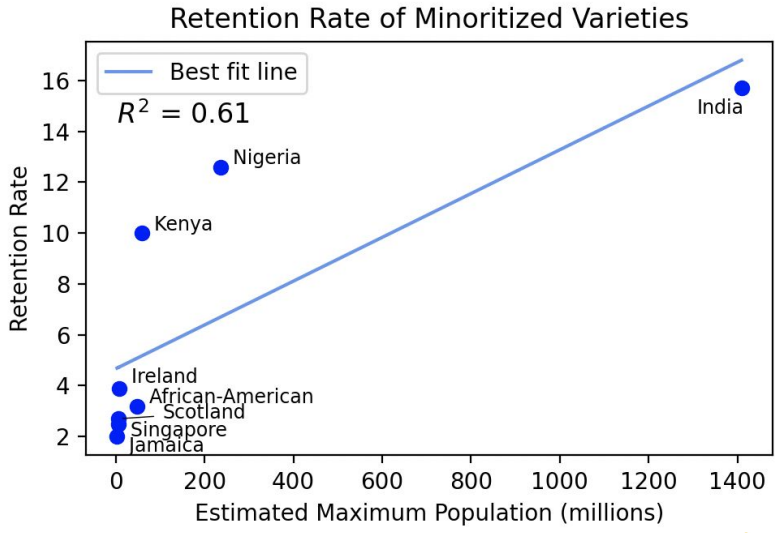
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Default Behavior in Responses

Inputs & responses coded for features of each variety

Retention rate correlates with estimated population

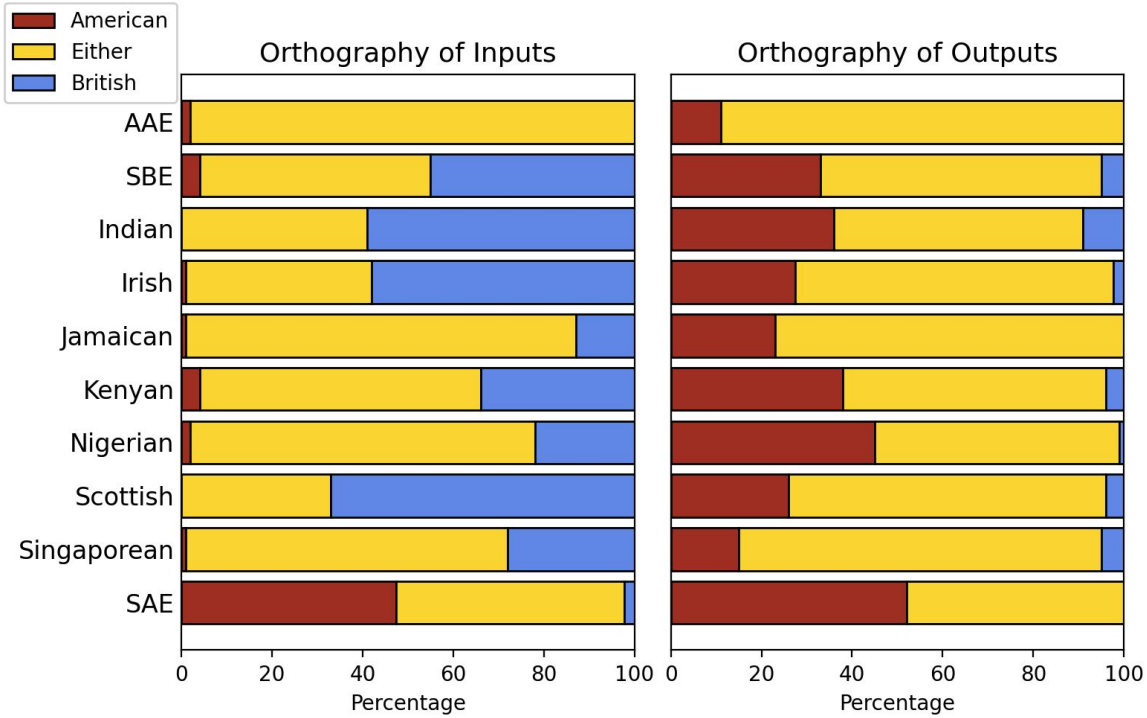
Variety of English	Percentage Retention ↑
Standard British	72.2%
Standard American	77.9%
Indian	15.7%
Nigerian	12.6%
Kenyan	10.0%
Irish	3.9%
African American	3.2%
Scottish	2.7%
Singaporean	2.5%
Jamaican	2.0%



Default Behavior in Responses

Most of these dialects typically use British orthography (spelling), but responses usually switch to American orthography

Responses retain borrowed words from minoritized varieties much more than other grammatical features



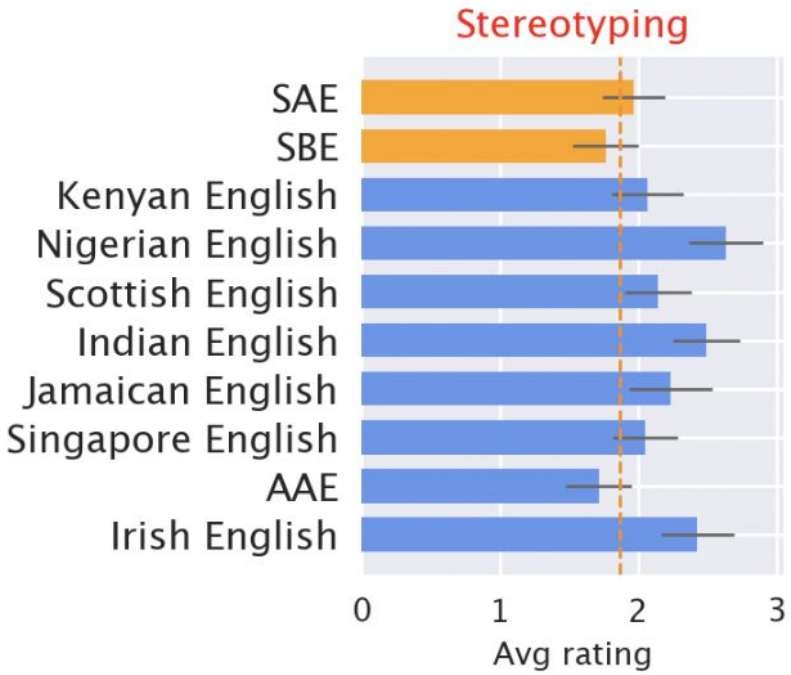
Berkeley

Harms in Responses

- Recruited native speakers of each variety through Prolific
- Speakers assessed a random sample of responses (5-point scale) for:
  - Naturalness
  - Warmth
  - Friendliness
  - Respect
  - Comprehension
  - Formality
  - Stereotyping
  - Demeaning content
  - Condescension

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

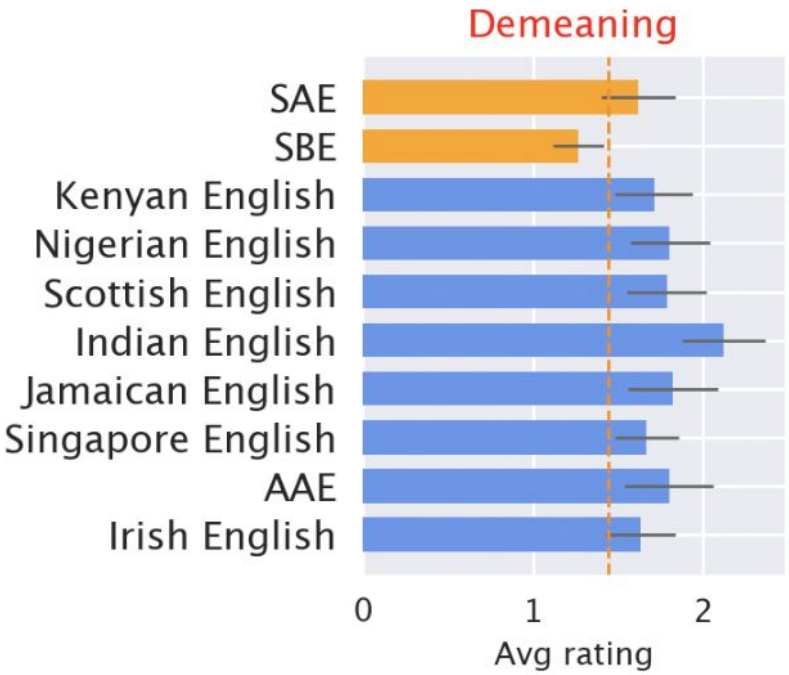
Native Speaker Evaluation: Replies to Minoritized Varieties



Stereotyping content increases for minoritized varieties (19% worse than for “standard” varieties)

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

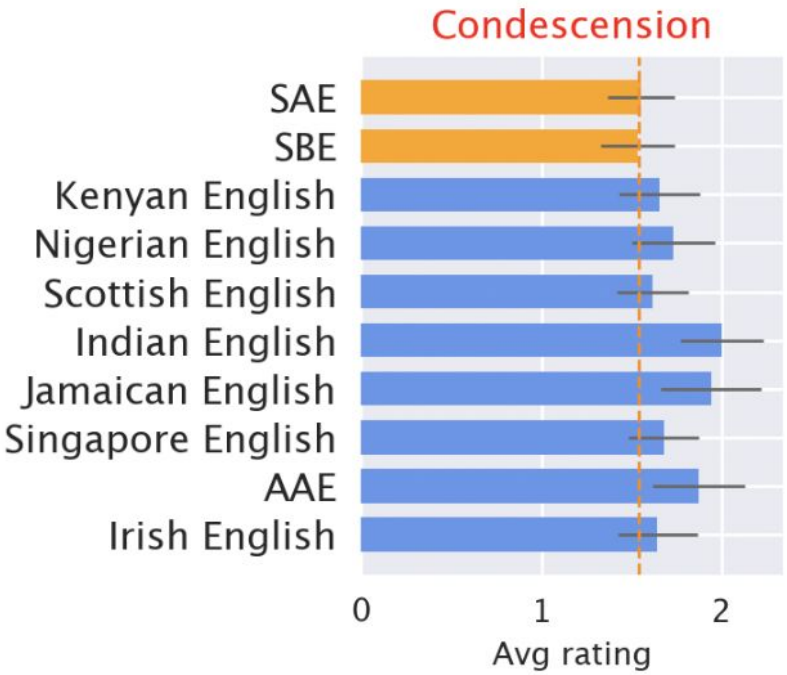
Native Speaker Evaluation: Replies to Minoritized Varieties



Demeaning content increases for minoritized varieties (25% worse)

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

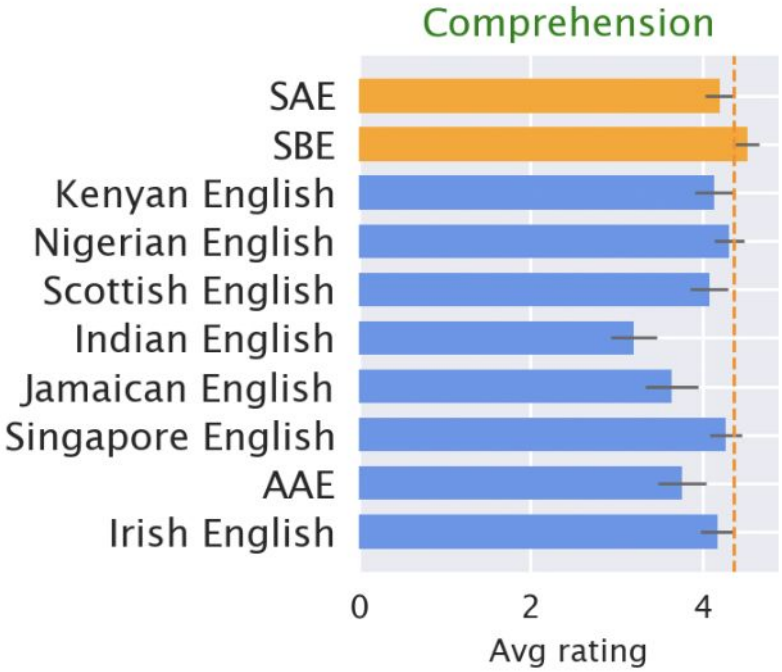
Native Speaker Evaluation: Replies to Minoritized Varieties



Condescension increases for minoritized varieties (15% worse)

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

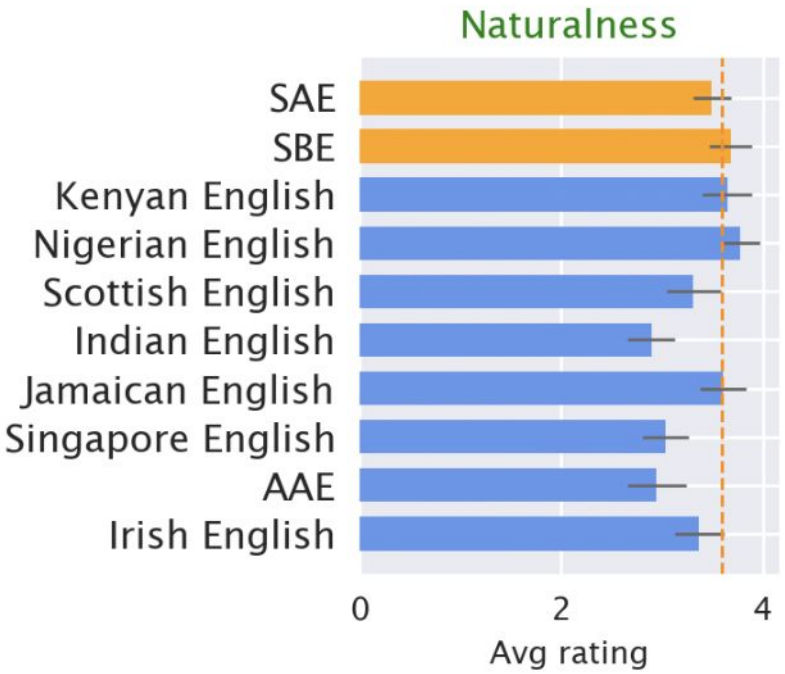
Native Speaker Evaluation: Replies to Minoritized Varieties



Comprehension decreases for minoritized varieties (9% worse)

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Native Speaker Evaluation: Replies to Minoritized Varieties



Naturalness decreases for minoritized varieties (8% worse)

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Potential Effects

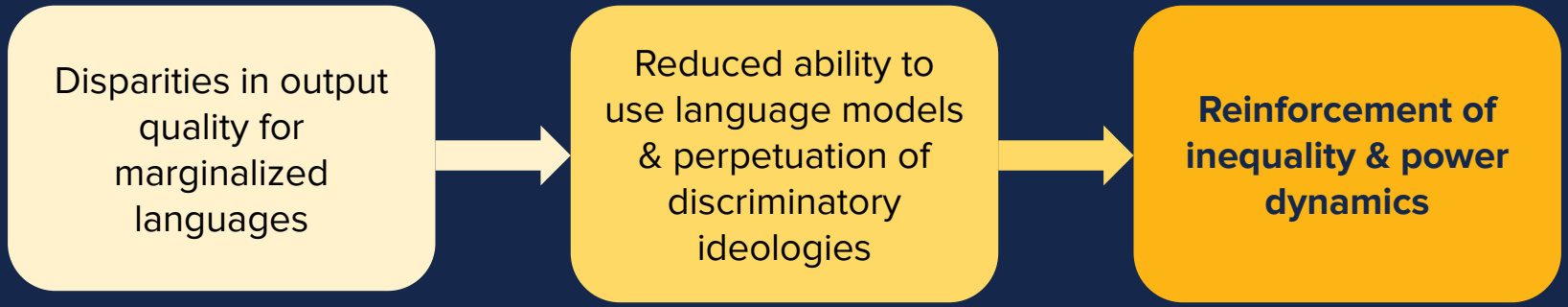
Native speakers trying to interact with these models face several issues

- Models don’t understand them
- Model outputs are stereotyping, demeaning, and condescending

If models try to produce these varieties, they introduce new issues

- Increased stereotyping
- Worsened comprehension

Takeaways

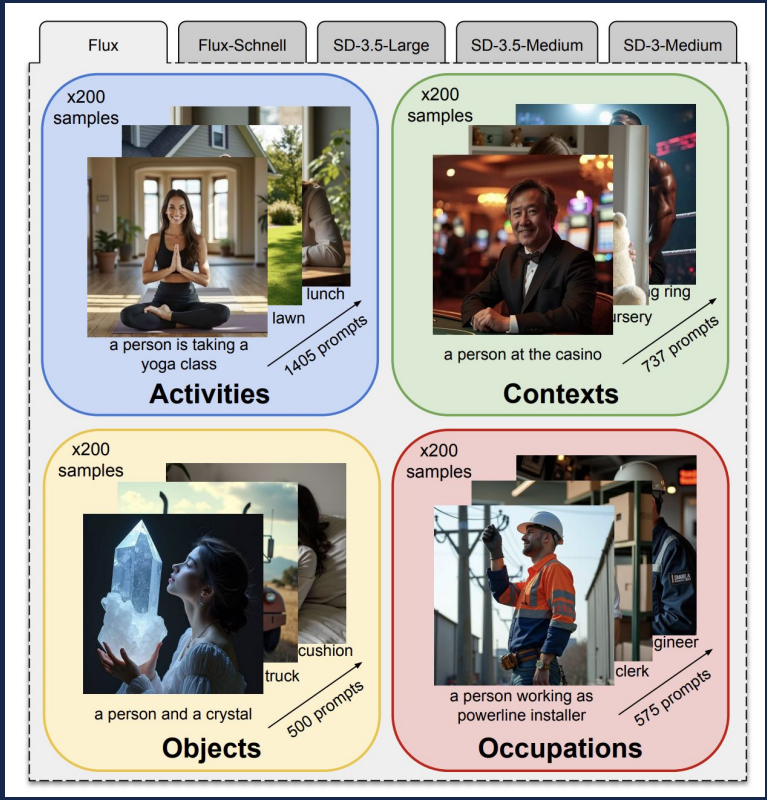


BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Example 2: Gender bias in text-to-image

- Analyze gender associations in daily activities, objects, contexts

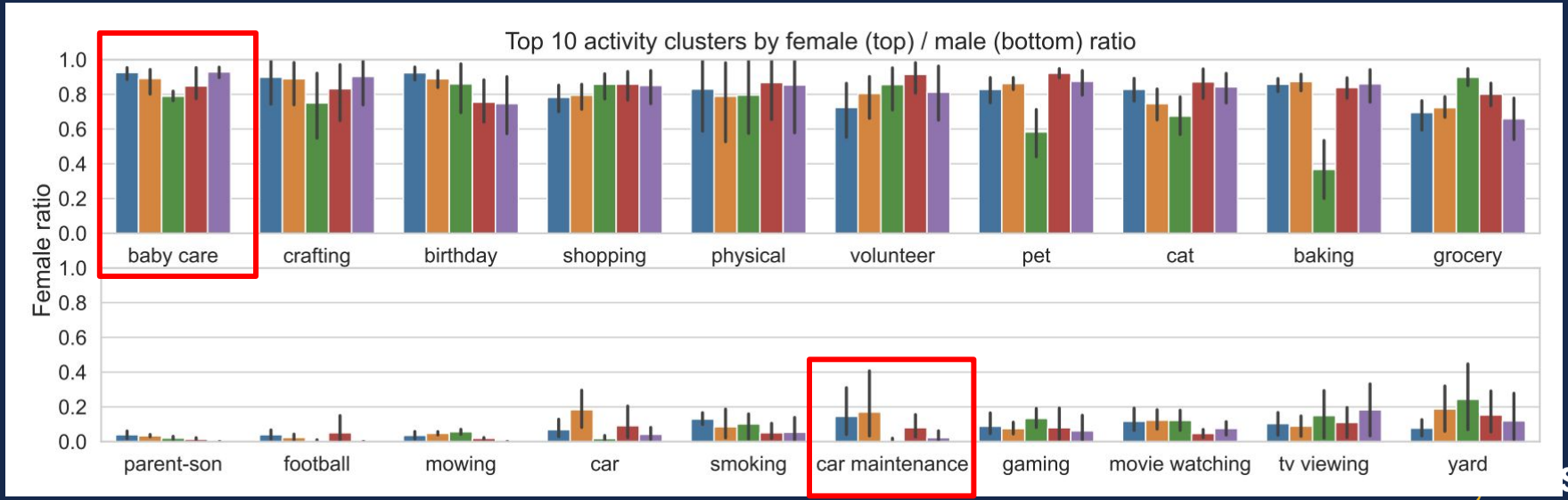
- Dataset of 3,217 gender-neutral prompts and 200 images per prompt from leading T2I models → ~2.3 million images



Joint work with co-authors: Leander Girrbach, Stephan Alaniz, and Zeynep Akata (and ongoing Vongani Maluluke, Trevor Darrell)

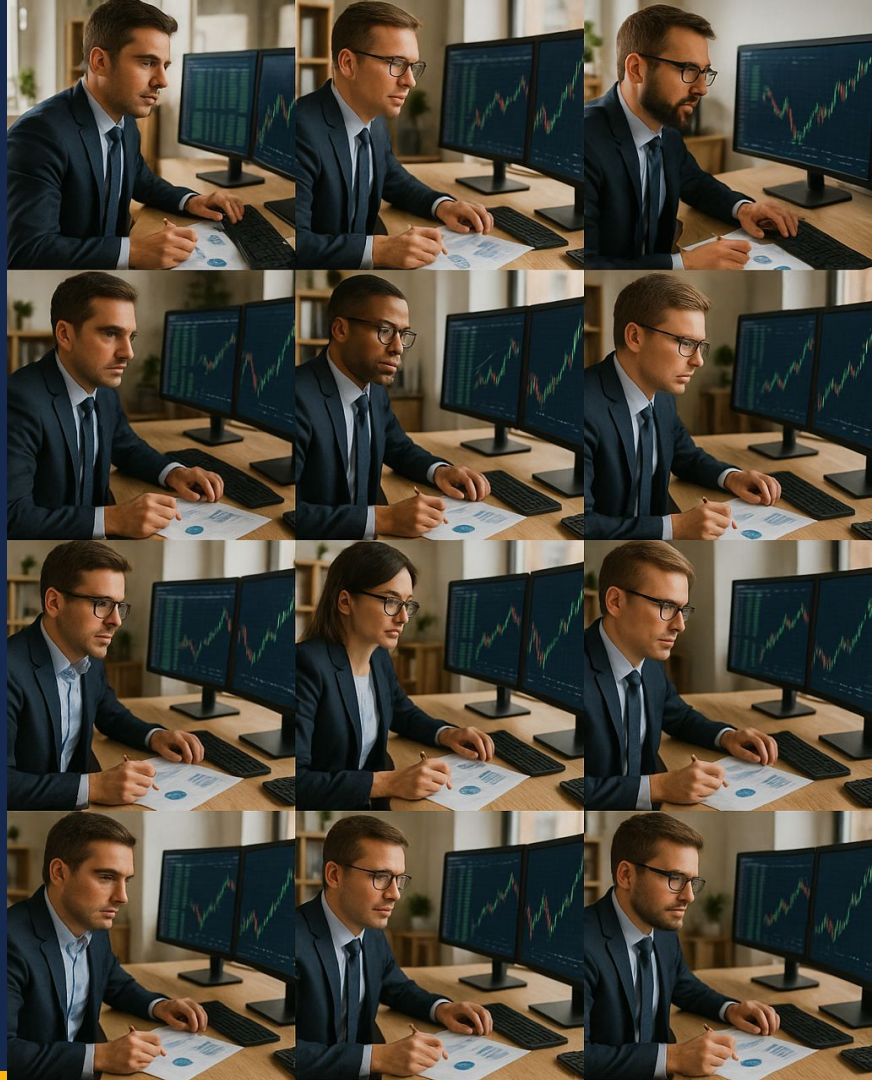
Gender bias in text-to-image – Results

- Reinforcement of traditional gender roles
- Reflection of common gender stereotypes in household roles (e.g., caretaking vs physical labor)



Not only reflecting bias, but amplifying

- Images of financial analysts → 16% of image outputs included women
- Women are 43.9% of financial analysts in the US



Not only reflecting bias, but amplifying

- Images of financial analysts → 16% of image outputs included women
- Women are 43.9% of financial analysts in the US

	Male majority		Female majority	
	reduced	amplified	reduced	amplified
Flux	7.12%	44.11%	19.18%	29.59%
Flux-Schnell	6.03%	45.21%	17.81%	30.96%
SD-3.5-Large	7.67%	43.56%	23.01%	25.75%
SD-3.5-Medium	10.68%	40.55%	28.22%	20.55%
SD-3-Medium	9.32%	41.92%	28.77%	20.00%

…By reinforcing stereotypes, these tools can negatively shape public perceptions and impact one's behavior.

Where is bias coming from?

BerkeleyHaas

Why are AI systems biased?

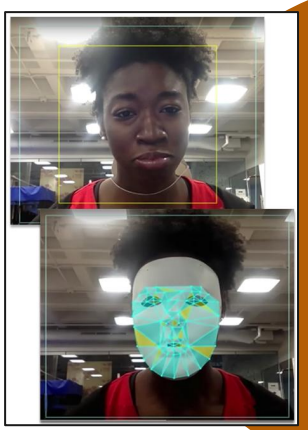
BerkeleyHaas

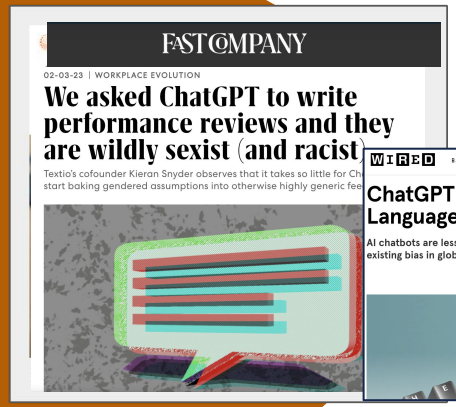
DATA GENERATION & COLLECTION

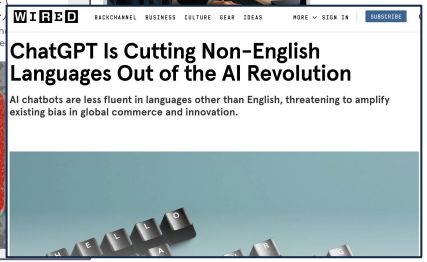
DATASET DEVELOPMENT

DATA LABELING

BIASED DATASET

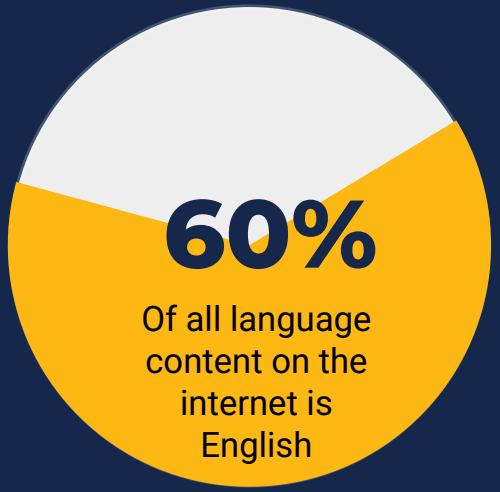






Where *are ML* datasets *coming from?*

Large language models rely on data from the internet, but Internet use varies



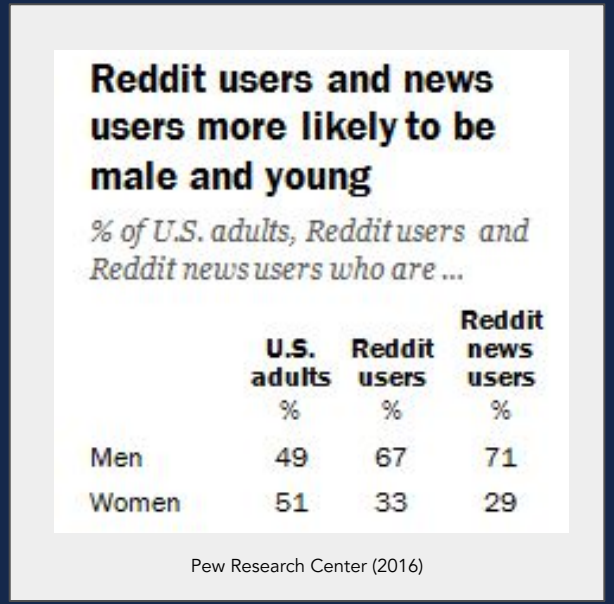
Only 7

Out of ~7000 languages in use in the world have large digital data typically called for in machine learning

Differing performance & opportunities

Even among well-represented languages…

- Some perspectives are overrepresented
- Others are actively marginalized
  - Ex: Harassment of women on Twitter can lead to self-censorship



Reddit users and news users more likely to be male and young

% of U.S. adults, Reddit users and Reddit news users who are …

	U.S. adults	Reddit users	Reddit news users
	%	%	%
Men	49	67	71
Women	51	33	29

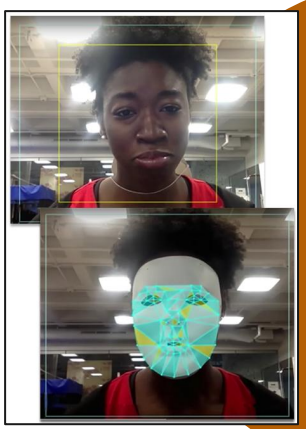
Pew Research Center (2016)

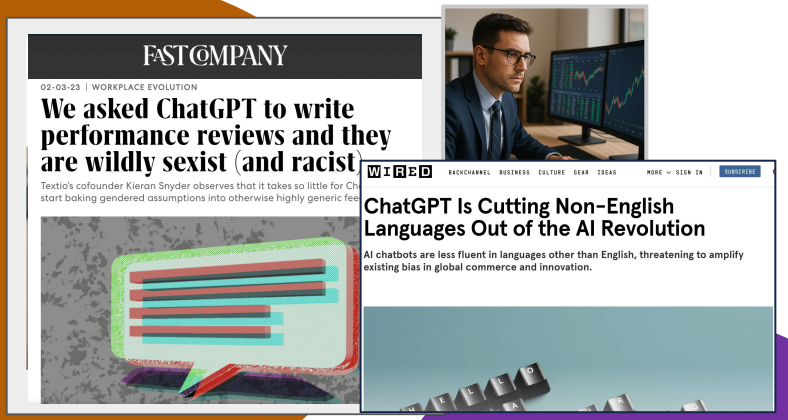
DATA GENERATION & COLLECTION | DATASET DEVELOPMENT | DATA LABELING

BIASED DATASET

PURPOSE | INPUTS | EVALUATION

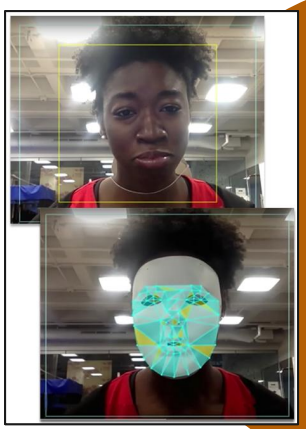
BIASED ALGORITHM





DATA GENERATION & COLLECTION | DATASET DEVELOPMENT | DATA LABELING

BIASED DATASET



PURPOSE | INPUTS | EVALUATION

BIASED ALGORITHM

INACCURATE PREDICTIONS

×

DISCRIMINATORY PREDICTIONS & OUTPUTS

CONTEXT | ALTERATION | INTERPRETATION

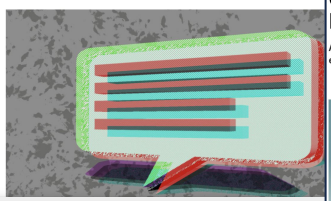
USE OF AN AI SYSTEM

FAST COMPANY

02-03-23 | WORKPLACE EVOLUTION

We asked ChatGPT to write performance reviews and they are wildly sexist (and racist)

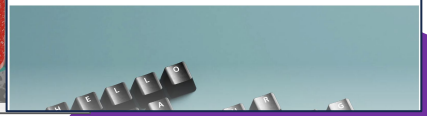
Textio's cofounder Kieran Snyder observes that it takes so little for Ch... start baking gendered assumptions into otherwise highly generic fee...



WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS MORE SIGN IN SUBSCRIBE

ChatGPT Is Cutting Non-English Languages Out of the AI Revolution

AI chatbots are less fluent in languages other than English, threatening to amplify existing bias in global commerce and innovation.



ICE rigged its algorithms to keep immigrants in jail, claims lawsuit / A 'secret no-release policy'

By ADI ROBERTSON / @thedextriarchy
Mar 3, 2020, 5:59 PM GMT | 0 Comments

Screenshot: The Verge

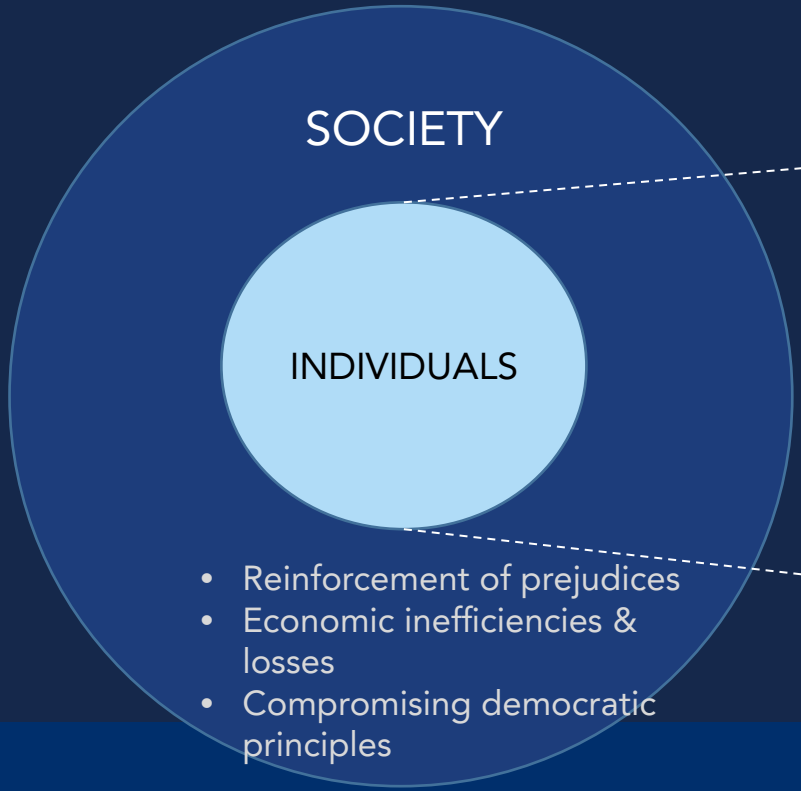
Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

By Starre Vartan on October 24, 2019

Health care algorithms can reinforce existing inequality. Credit: Getty Images

What happens if AI making important decisions is biased?



- Unfair resource allocation
- Sub-par service
- Health hazard
- Safety detriment
- Infringement of civil liberties
- Derogatory treatment

TRUST?

The role of “fairness”

BerkeleyHaas

Bias tied to notions of fairness

Quality or state of being fair, especially fair or impartial treatment

Law

Social science

Quant fields

Philosophy

Different things in different contexts to different people…

BerkeleyHaas

Bias tied to notions of fairness

Quality or state of being fair, especially fair or impartial treatment

Law

Social science

Quant fields

Philosophy

ML researchers tend to focus on quant perspectives…

BerkeleyHaas

Role of ‘Fairness’

- Often around sensitive, legally protected attributes → model to perform as optimally as possible while treating people “fairly” with respect to these sensitive attributes
- Simplest approach: demographic parity across subgroups; e.g., each subgroup receives positive outcome at equal rates/ same proportion

The definition of fairness used and the fairness approach taken can inform how bias both manifests and is interpreted

Role of ‘Fairness’: COMPAS

Correctly predicted recidivism for Black & White defendants at roughly same rate, but wrong in different ways

- Black arrestees who wouldn’t be rearrested in a 20-year horizon scored as high risk at 2x rate of White arrestees not subsequently arrested
- White people more likely than Black people to later commit a crime scored as lower

Equivant: fair

Model reflected same likelihood of recidivism across all groups

Treating all citizens according to same rules

ProPublica: not fair

Didn’t treat likes alike

Wrong in different ways, repeating (unjust) status quo

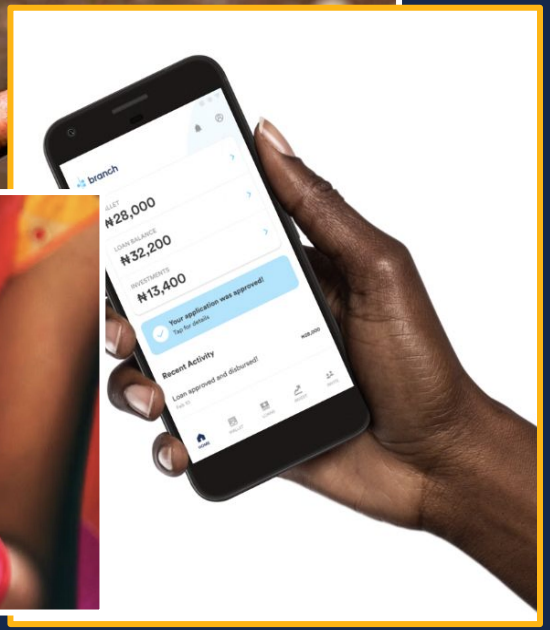
BerkeleyHaas

Role of ‘Fairness’: Fintechs & lending

- Gender differences in credit scores & lending
- Fintechs see fairness as being “accurate” – i.e. creditworthy people repay

Issues…

- Learning from / projecting economic inequities
- Creditworthiness as self-fulfilling prophecy



The New Jim Code

“The employment of new technologies that reflect and reproduce existing inequities but that are promoted and perceived as more objective or progressive than the discriminatory systems of a previous era.”

+ Default discrimination

BerkeleyHaas

TLDR on fairness

How might different tools perform differently for different people? And be used differently by different people?

Important to think beyond only technical definitions of fairness and bias.

BerkeleyHaas

Defining biased AI

AI that results in…

Inaccurate predictions (esp. for marginalized individuals / groups)

Harmful stereotyping; discriminatory outputs or predictions

Mitigating Bias in AI

BerkeleyHaas

Myth

- AI / algorithms / data are objective
- Fairness or bias in AI can be solved with technical solutions alone
- It is possible to completely debias or unbias AI

Reality

Data reflects historical and current social inequities which AI tools can then learn from

There are different ways to consider what is fair

Choices and tradeoffs are part of building and managing AI tools; Mitigating bias is the goal

BerkeleyHaas

Strategies to take forward

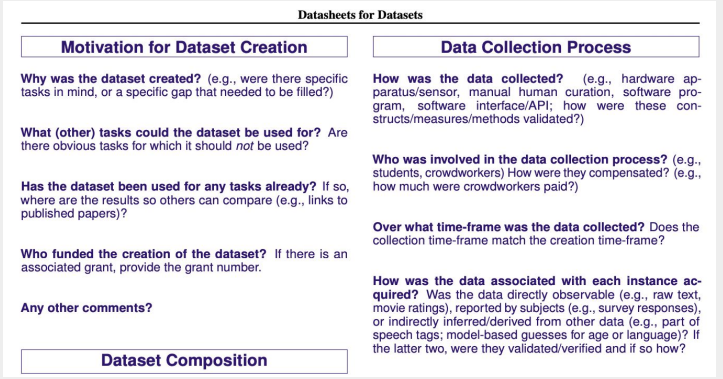
1. *Recognize limitations and keep a curious mindset.*
    a. Make explainability and transparency around shortcomings and pitfalls of AI systems the norm

BerkeleyHaas

Strategies to take forward

1. *Recognize limitations and keep a curious mindset.*
2. *Practice responsible dataset development*
    a. Assess existing datasets to check for over-/ under-representation of certain identities, underlying inequities
    b. Track & document how datasets were created, their content & limitations

Datasheets for Datasets



Datasheets for Datasets

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should *not* be used?

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

Who funded the creation of the dataset? If there is an associated grant, provide the grant number.

Any other comments?

Dataset Composition

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?

Strategies to take forward

1. *Recognize limitations and keep a curious mindset.*
2. *Practice responsible dataset development*
3. *Pursue responsible algorithm development*

    a. When establishing the algorithm's purpose & objective, consider fairness tradeoffs / ethics

    b. Ensure datasets and proxies chosen do not disadvantage certain identities; conduct audits

TL;DR

- ★ *At a high level*: Individuals & institutions that design, develop, manage AI systems matter
- ★ *More granular*: bias can enter in data, algorithm, how the tool is used
- ★ We have the power to mitigate bias and consider tradeoffs in the technology we create

Thank you.

BAIR Responsible AI Initiative:
https://re-ai.berkeley.edu/home

Fall course on Responsible AI Innovation & Management