

# Ethics in NLP

CSE 5525: Foundations of Speech and Natural Language  
Processing

<https://shocheen.github.io/courses/cse-5525-spring-2025>



THE OHIO STATE UNIVERSITY

---

# Logistics

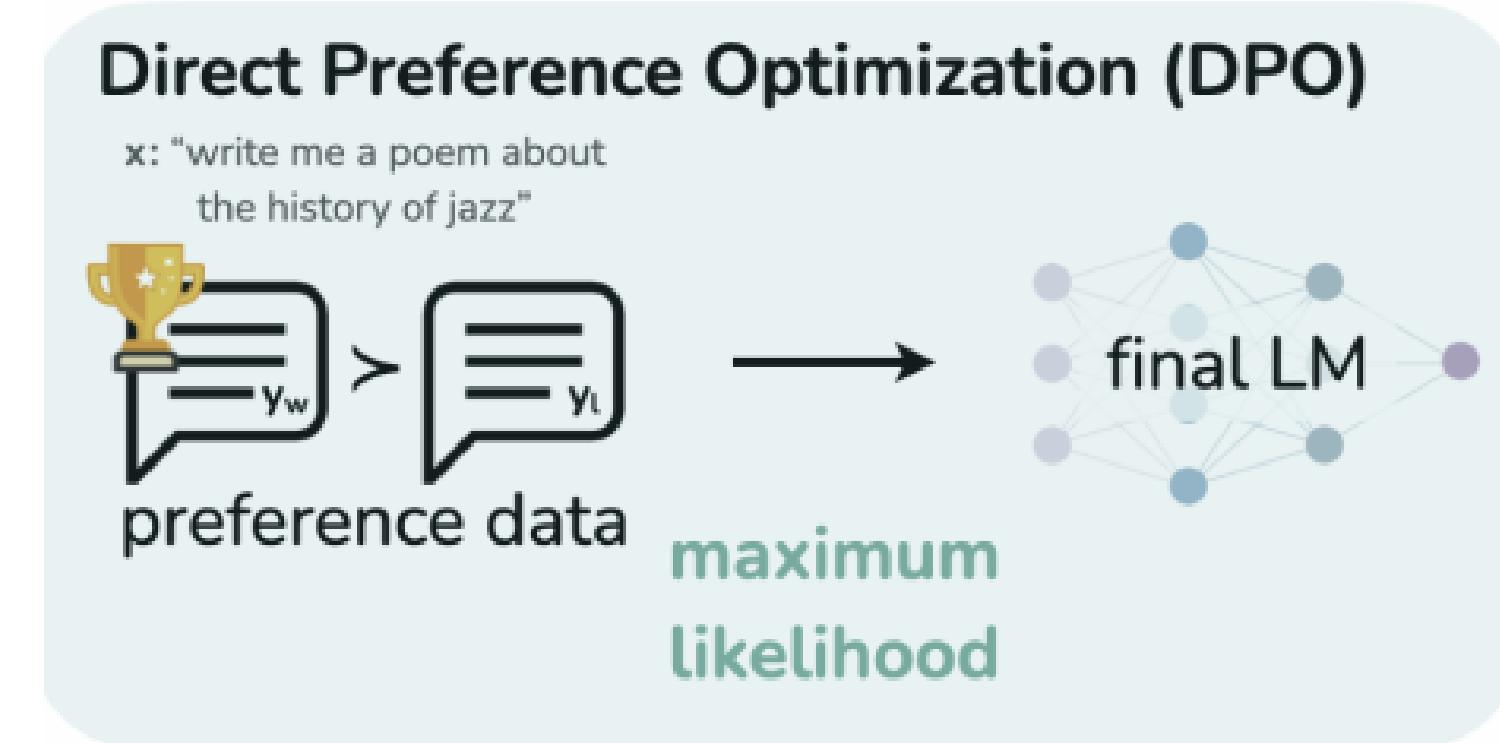
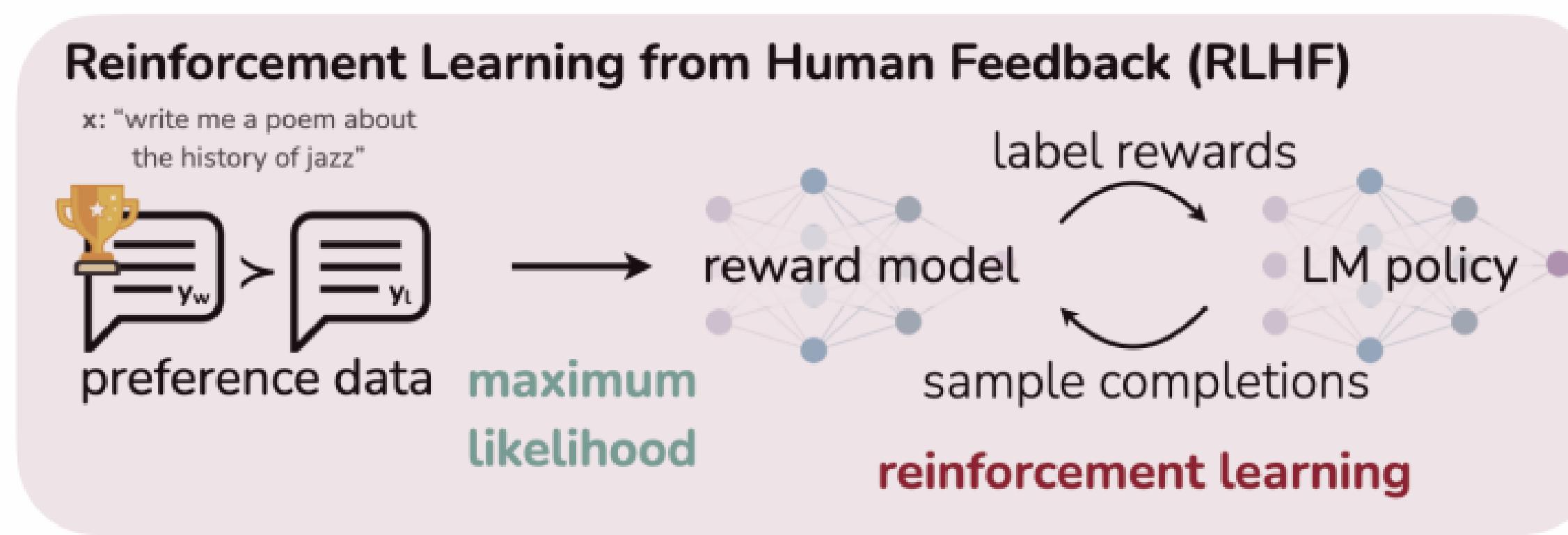
- Final project:
  - Project presentations: April 16, 18 (check your presentation day on teams)
  - Time to present: 5 minutes (all or one can present), include the research question/motivation and main findings / what's left.
  - I create a blank slide deck and share, please all your slides there in order.
- Final project report due date: April 25

# Last Class Recap

- Why build multilingual NLP models?
- Linguistic resources (High resource to no resources)
- Efforts on creating pretraining corpora
- Efforts on instruction tuning
  - Translation instructions, ask humans to write instructions, use templates

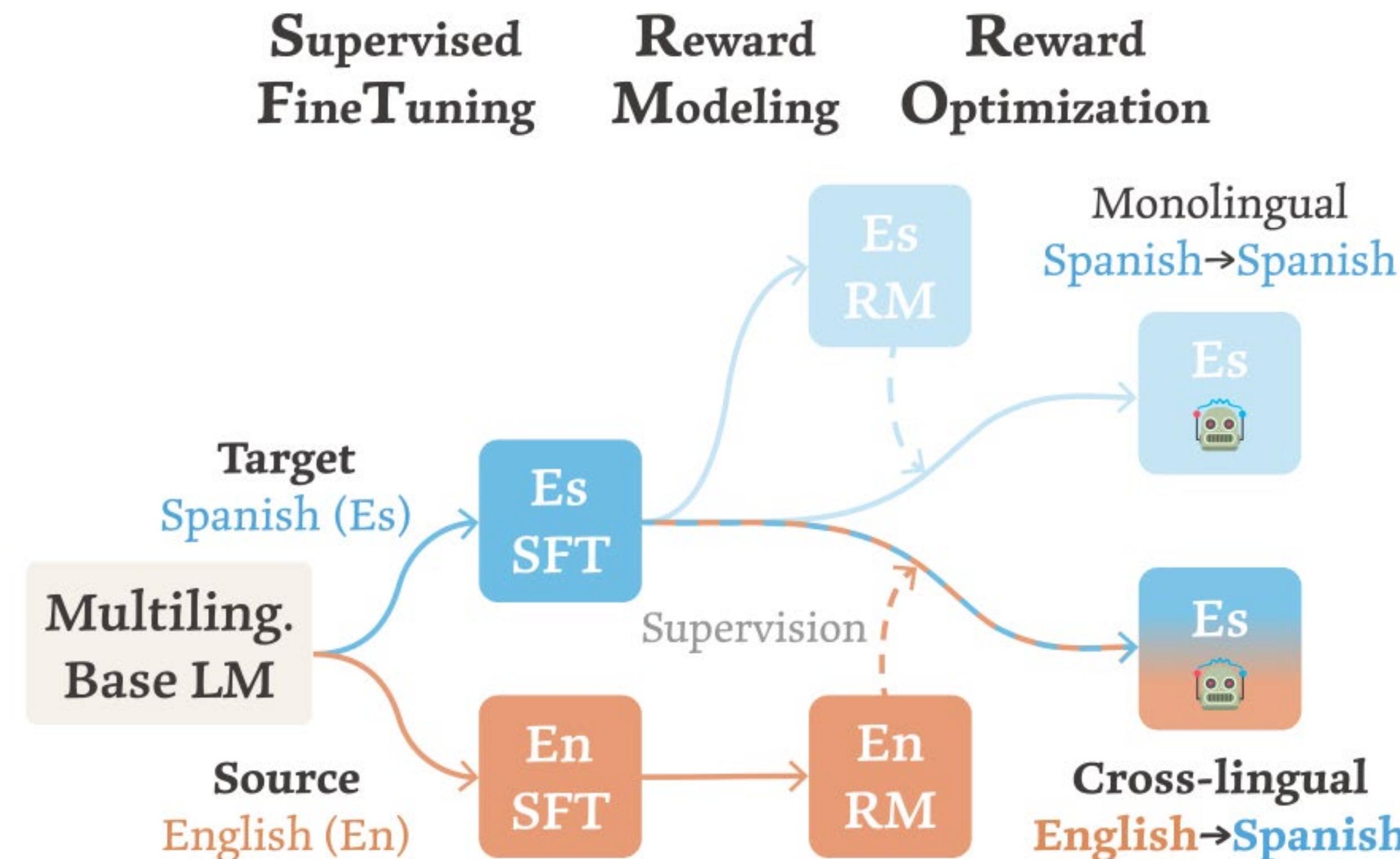
# Multilingual Alignment

# Preference Optimization

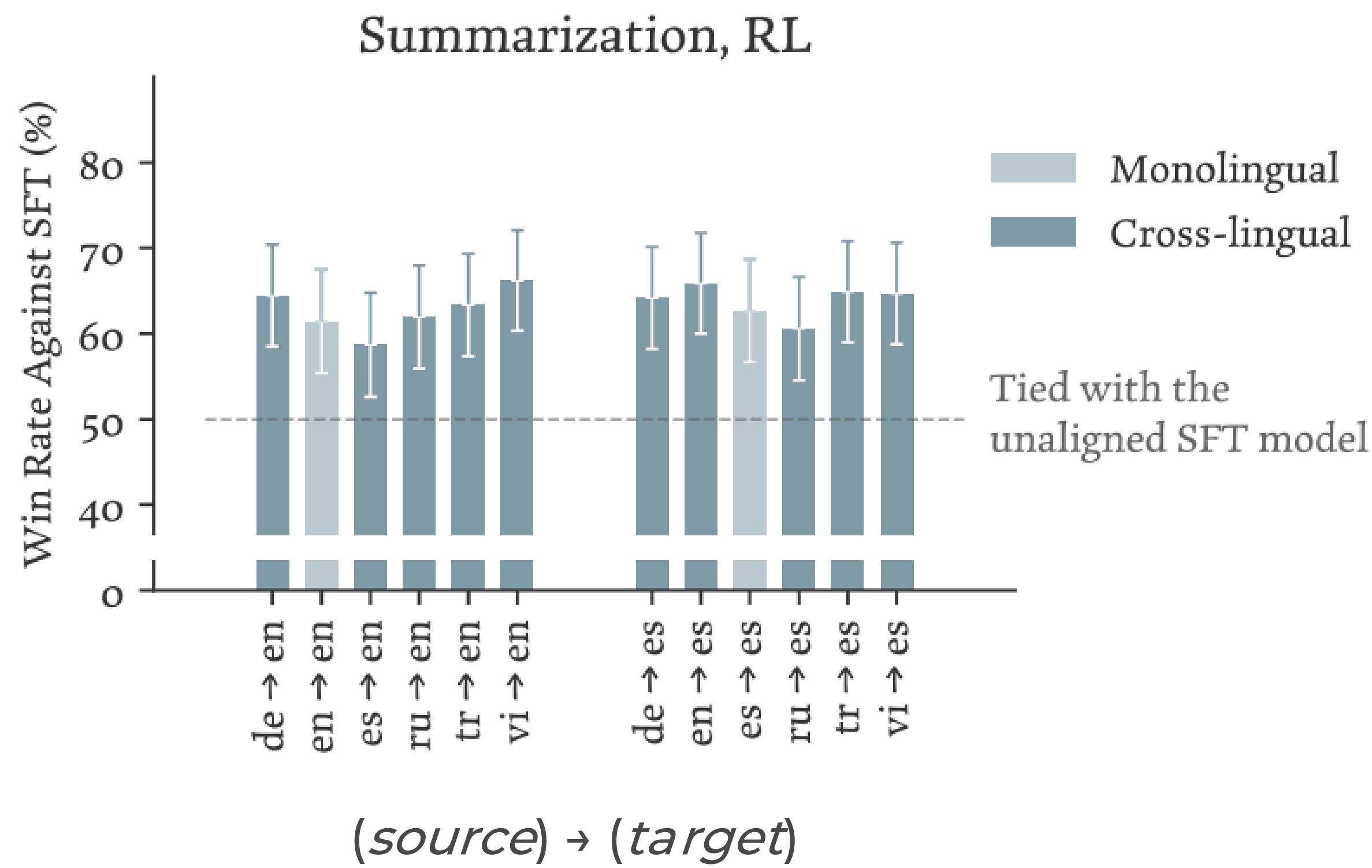


# Cross-lingual (X-Lingual) Alignment

- Reward model trained on preference data of language **X** (source)
- Applied to preference tune for language **Y** (target)



# Cross-lingual Alignment: Does it Work?



- **Evaluation:** Head-to-head win-rates as judged by humans
- **Base SFT model:** mT5-XL
- **Optimization:** Online (PPO)

Cross-lingual alignment sometimes **outperforms** in-language alignment

# Can't I Just Translate Source Preference Data

## Cross-lingual

Src \ Tgt	De	En	Es	Ru	Tr	Vi
De	52.3	50.8	63.0	66.7	63.0	60.4
En	56.4	55.5	66.1	70.7	67.2	63.1
Es	51.9	51.2	62.4	66.0	64.4	57.5
Ru	48.1	46.5	59.2	63.6	59.0	56.3
Tr	53.3	52.9	62.6	66.6	60.4	59.0
Vi	46.5	48.2	60.0	65.6	62.1	58.0

**Table 6:** Cross-lingual alignment results using **best-of- $n$**  with  $n = 64$ , for the **summarization** task, measured in win rate (%) against the target-language SFT model as judged by **PaLM-2-L** (Figure 4).

Translation > Cross-lingual

## Translation

Src \ Tgt	De	En	Es	Ru	Tr	Vi
De	—	50.0	61.9	66.1	66.1	54.6
En	47.9	—	63.3	64.9	64.5	53.1
Es	50.6	52.9	—	64.1	64.5	59.0
Ru	47.4	51.2	60.3	—	63.3	57.7
Tr	50.6	52.5	61.8	65.6	—	50.8
Vi	42.0	50.8	59.1	64.4	63.6	—

**Table 17:** Alignment quality using RM trained by translating the source language data into the target language using **best-of- $n$**  with  $n = 64$ , for the summarization task, measured in win rate (%) against the target-language SFT model as judged by PaLM-2-L (§5.1).

Can't say much!!

- English benefits from translation
- Russian (different script) doesn't transfer well

# Challenges

## Curse of multilinguality<sup>[1,2]</sup>

Packing more languages into a model decreases per language performance

## Cost of Technology<sup>[3]</sup>

- GPT\* models are behind paid APIs; cost  $\propto$  input & generation tokens
- Poor tokenization in non-English languages → more tokens
- More tokens → more latency & money
- Efforts made but far from parity<sup>[4,5]</sup>

## Dialectal Biases<sup>[6]</sup>

- Whose dialect matters the most?<sup>[7,8]</sup>
- Whose English?<sup>[9,10]</sup>

**and many more .....**

- [1] [Unsupervised Cross-lingual Representation Learning at Scale](#) (Conneau et al., ACL 2020)
- [2] [When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages](#) (Chang et al., 2023))
- [3] [Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models](#) (Ahia et al., EMNLP 2023)
- [4] <https://cohere.com/blog/command-r-plus-microsoft-azure>
- [5] <https://openai.com/index/hello-qpt-4o/>
- [6] [A Survey of Corpora for Germanic Low-Resource Languages and Dialects](#) (Blaschke et al., NoDaLiDa 2023)
- [7] [Decolonizing NLP for “Low-resource Languages”](#) (Ògúnřémí et al., AI Frameworks Discussion of Abeba Birhane's "Algorithmic Injustice" and Social Impact Articles 2023)
- [8] [Which Humans?](#) (Atari et al., 2023)
- [9] [What to do about non-standard \(or non-canonical\) language in NLP](#) (Plank, KONVENS 2016)
- [10] [AI makes racist decisions based on dialect](#) (Science, 24 August 2024)

# Other Interesting Directions

## Multilingual Architectures

- Efficient solutions for the curse of multilinguality
- Adding some language-specific parameters
- E.g.: Adapters<sup>[1]</sup>, Cross-lingual expert models<sup>[2]</sup>

## Tokenization and Vocabulary

- Efficient tokenization methods to reduce costs and latency
- E.g.: Vocab budgeting<sup>[6]</sup>, allocation<sup>[7]</sup>

## Adapting to a New Language

- Increasing support of an **N** language multilingual model to **N+K** languages
- E.g.: Continued pretraining<sup>[3]</sup>, Adapters<sup>[4]</sup>, Efficient Initializations<sup>[5]</sup>

## Data Creation and Verification

- Methods for synthetic data generation<sup>[8]</sup> and verification of labeled data<sup>[9]</sup>

- [1] [MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer](#) (Pfeiffer et al., EMNLP 2020)
- [2] [Breaking the Curse of Multilinguality with Cross-lingual Expert Language Models](#) (Blevins et al., 2024)
- [3] [How to Adapt Your Pretrained Multilingual Model to 1600 Languages](#) (Ebrahimi & Kann, ACL-IJCNLP 2021)
- [4] [BLOOM+1: Adding Language Support to BLOOM for Zero-Shot Prompting](#) (Yong et al., ACL 2023)
- [5] [OFA: A Framework of Initializing Unseen Subword Embeddings for Efficient Large-scale Multilingual Continued Pretraining](#) (Liu et al., Findings 2024)
- [6] [XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models](#) (Liang et al., EMNLP 2023)
- [7] [Tokenization Impacts Multilingual Language Modeling: Assessing Vocabulary Allocation and Overlap Across Languages](#) (Lemisiewicz et al., Findings 2023)
- [8] [Multilingual Arbitrage: Optimizing Data Pools to Accelerate Multilingual Progress](#) (Odumakinde et al., 2024)
- [9] [Verifying Annotation Agreement without Multiple Experts: A Case Study with Gujarati SNACS](#) (Mehta & Srikumar, Findings 2023)

# AI Ethics

# What is Ethics?

Ethics is a study of what are **good and bad** ends to pursue in life and what it is **right and wrong** to do in the conduct of life.

It is therefore, above all, a practical discipline.

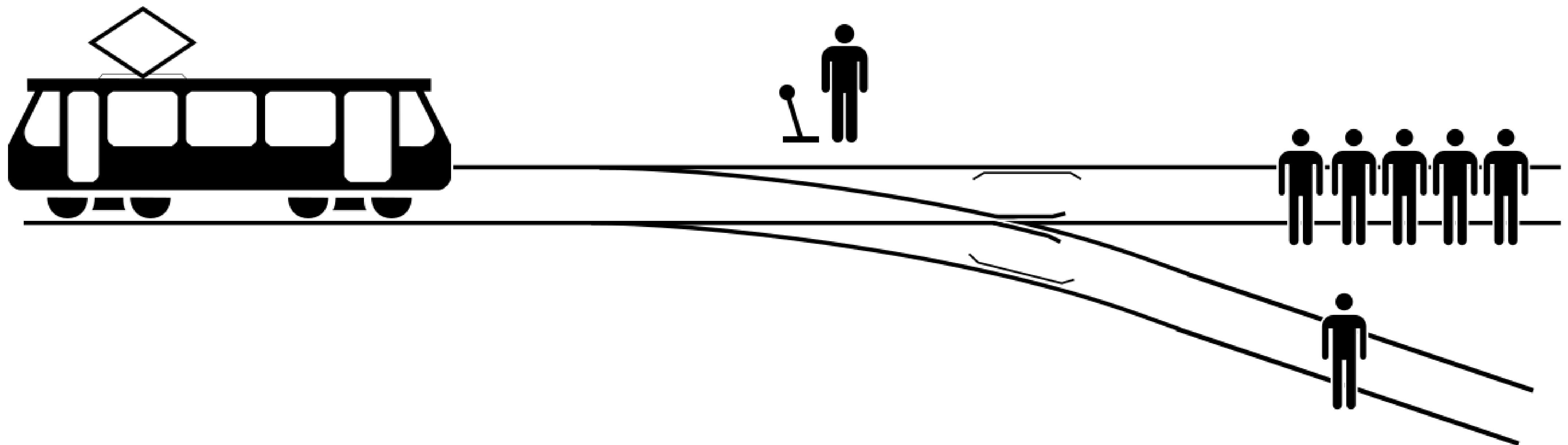
Its primary aim is to determine how one ought to live and what actions one ought to do in the conduct of one's life.”

# What is Ethics?

It's the **good** things

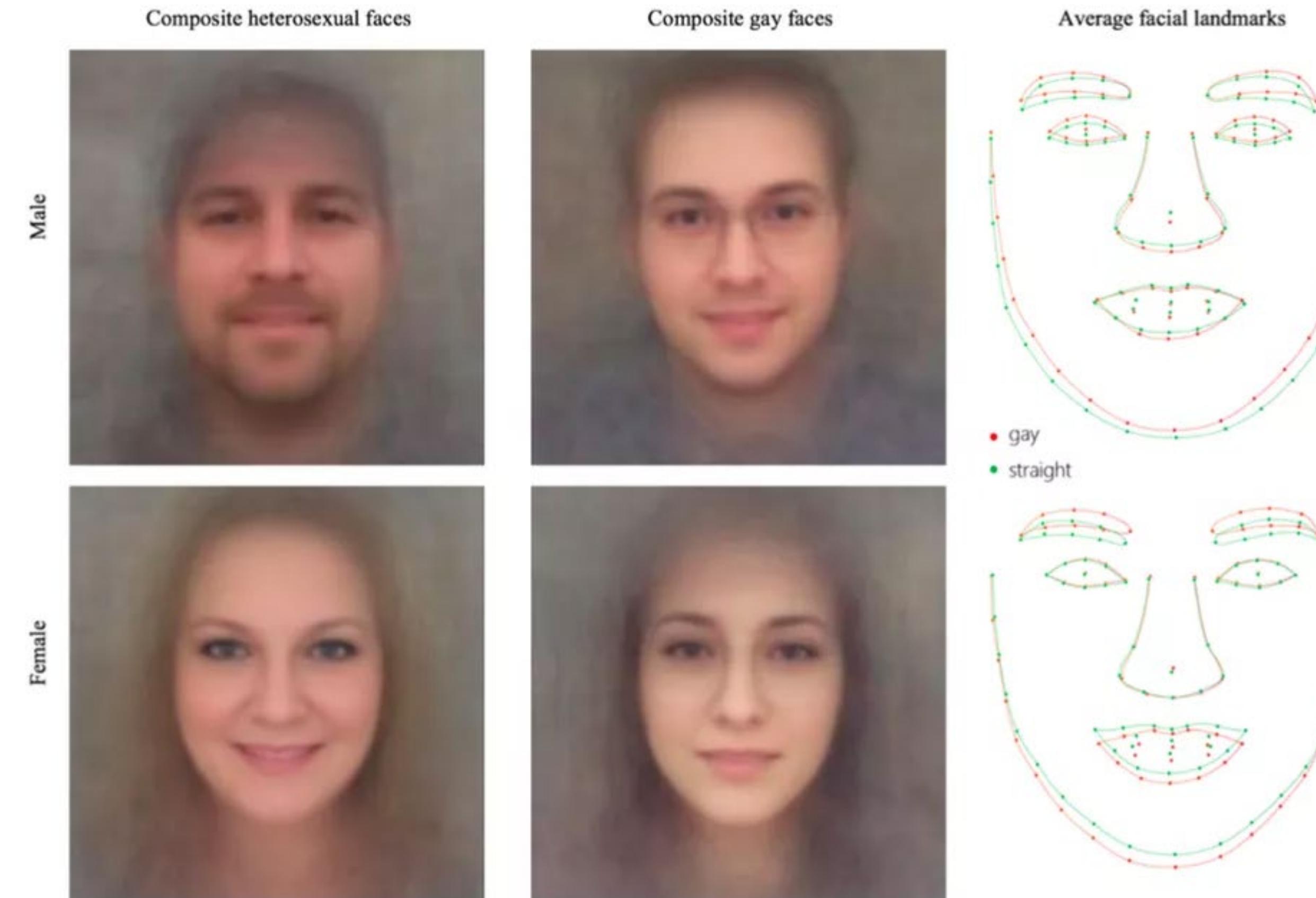
It's the **right** things

# The Trolley Dilemma



What are important ethical questions to ask in development and deployment of AI systems?

# A recent study: the “AI Gaydar”, 2017



# A recent study: the “AI Gaydar”

- Research question
  - Identification of sexual orientation from facial features
- Data collection
  - Photos downloaded from a popular American dating website
  - 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly
- Method
  - A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification
- Accuracy
  - 81% for men, 74% for women
- Motivation for the study: expose a threat to the privacy and safety of gay men and women

# Let's discuss...

- Research question
  - Identification of sexual orientation from facial features
- Data collection
  - Photos downloaded from a popular American dating website
  - 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly
- Method
  - A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification
- Accuracy
  - 81% for men, 74% for women

What went wrong?

# Questioning the ethics of the research question

- Identification of sexual orientation from facial features



# Sexual orientation classifier - who can be harmed?

- . In many countries being gay person is prosecutable (by law or by society) and in some places there is even death penalty for it
- . It might affect people's employment; family relationships; health care opportunities;
- . Gender, race, sexual orientation, religion are personal attributes and/or social constructs. Many of them can change over time. They can be non-binary. They are private, intimate, often not visible publicly.
- . Importantly, these are properties for which people are often discriminated against.

# DALLE-2 result in 2022

- Image of a teacher:



- *Let's discuss*: what do you notice about this image generation result?

- Only women
- Only white people
- Certain age range

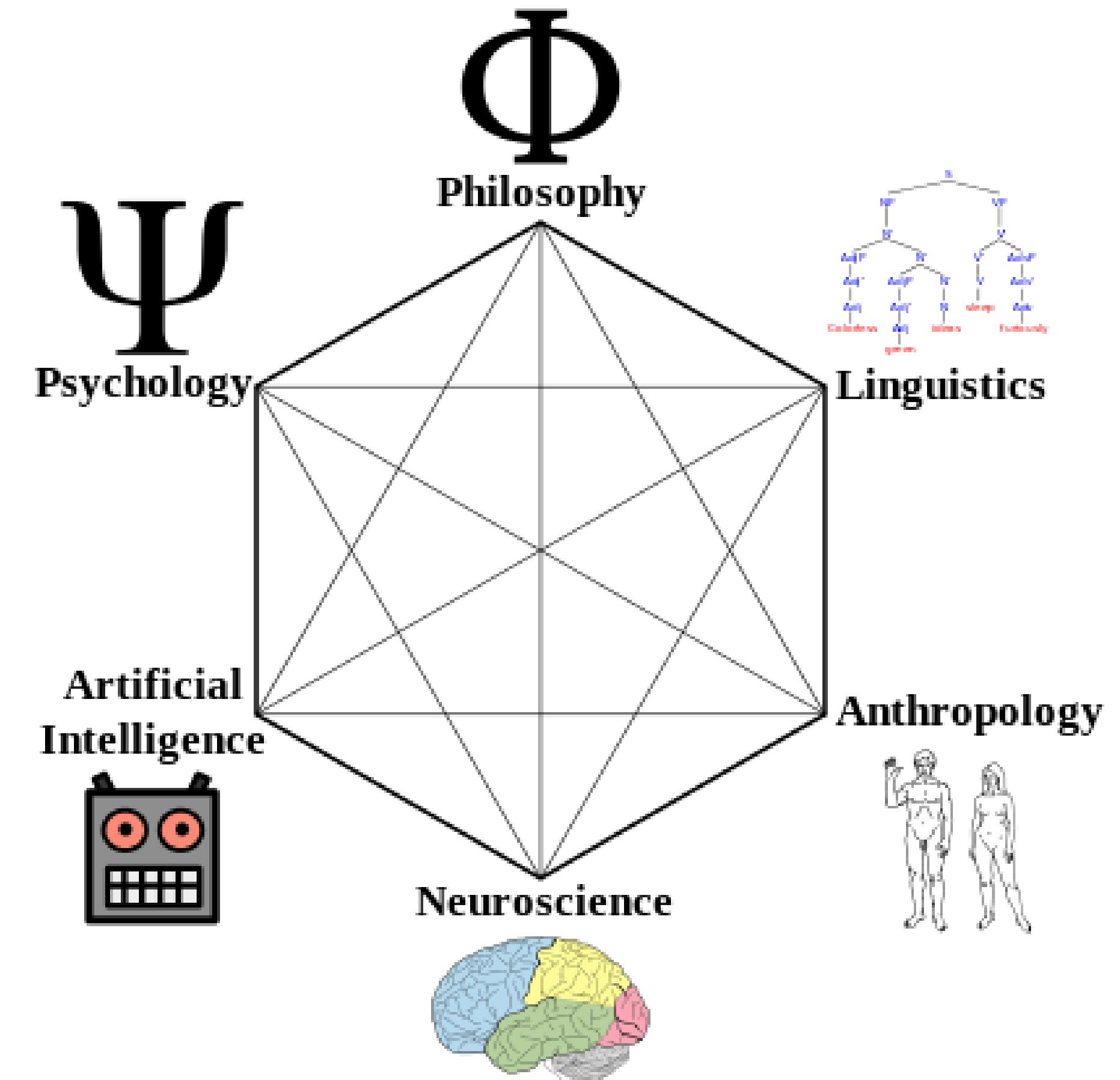
# NLP tools have biases, and pose ethical risks

The collage consists of five news snippets arranged in a grid-like layout:

- VICE MOTHERBOARD**: Headline: "'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says'. Subtext: 'The incident raises concerns about guardrails around quickly proliferating conversational AI models.' By Chloe Xiang, March 30, 2023.
- BUSINESS FORTUNE Well.**: Headline: 'They fell in love with AI bots. A software update broke their hearts.' Subtext: 'Loneliness is widespread. Artificial intelligence can help combat it, but it comes with risks.' By Pranshu Verma, March 30, 2023 at 6:00 a.m. EDT.
- The Washington Post**: Headline: 'They fell in love with AI bots. A software update broke their hearts.' Subtext: 'Loneliness is widespread. Artificial intelligence can help combat it, but it comes with risks.' By Pranshu Verma, March 30, 2023 at 6:00 a.m. EDT.
- GIZMODO**: Headline: 'Move Aside, Crypto. AI Could Be The Next Climate Disaster.' Subtext: 'A new Stanford report highlights the staggering carbon emissions required to train and maintain large language models like OpenAI's ChatGPT.' By Mack DeGeurin, Published April 3, 2023 | Comments (6).

# Ethics in AI

- Data origin and ownership
- Social bias in AI
- Algorithmic fairness
- Privacy risks and protection
- Mis-use of AI:
  - disinformation, opinion manipulation
- AI for good
  - content moderation
  - assistive technologies, disaster response
- Societal impacts of AI
  - environmental
  - economic, educational, policy impacts
- ....



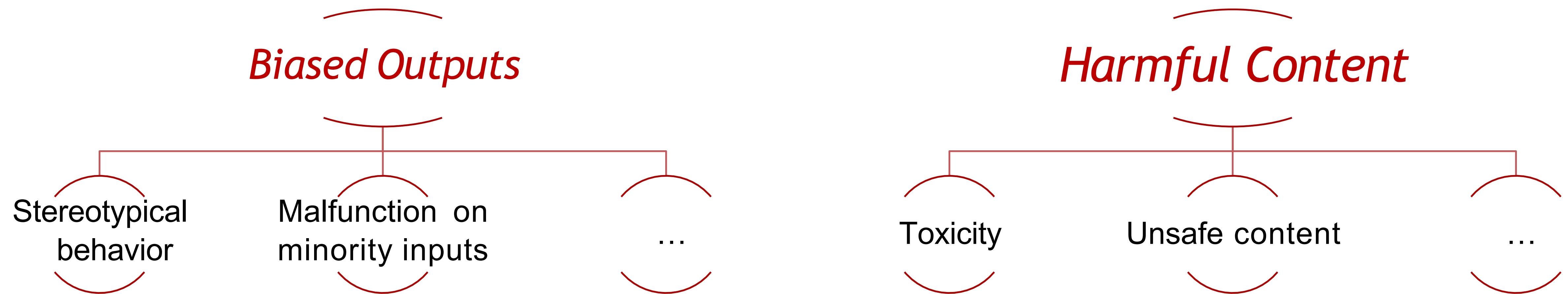
# Risks of Harms of NLP Models

Theme	Subcategory
Representational Harms	Stereotyping Demeaning Social Groups
Classification	Harm
Discrimination, Exclusion and Toxicity	<p>Social stereotypes and unfair discrimination Exclusionary norms Toxic language Lower performance for some languages and social groups</p>
Information Hazards	<p>Compromising privacy by leaking private information Compromising privacy by correctly inferring private information Risks from leaking or correctly inferring sensitive information</p>
Misinformation Harms	<p>Disseminating false or misleading information Causing material harm by disseminating false or poor information e.g. in medicine or law Leading users to perform unethical or illegal actions</p>
Malicious Uses	<p>Making disinformation cheaper and more effective Facilitating fraud, scams and more targeted manipulation Assisting code generation for cyber attacks, weapons, or malicious use Illegitimate surveillance and censorship</p>
Human-Computer Interaction Harms	<p>Anthropomorphising systems can lead to overreliance or unsafe use Creating avenues for exploiting user trust, nudging or manipulation Promoting harmful stereotypes by implying gender or ethnic identity</p>
Automation, access, and environmental harms	<p>Environmental harms from operating LMs Increasing inequality and negative effects on job quality Undermining creative economies Disparate access to benefits due to hardware, software, skill constraints</p>

# What we will **not** discuss

- . We will not discuss on notions of existential AI risk, and other related theories.
- . We will not discuss policy related solutions for AI safety.

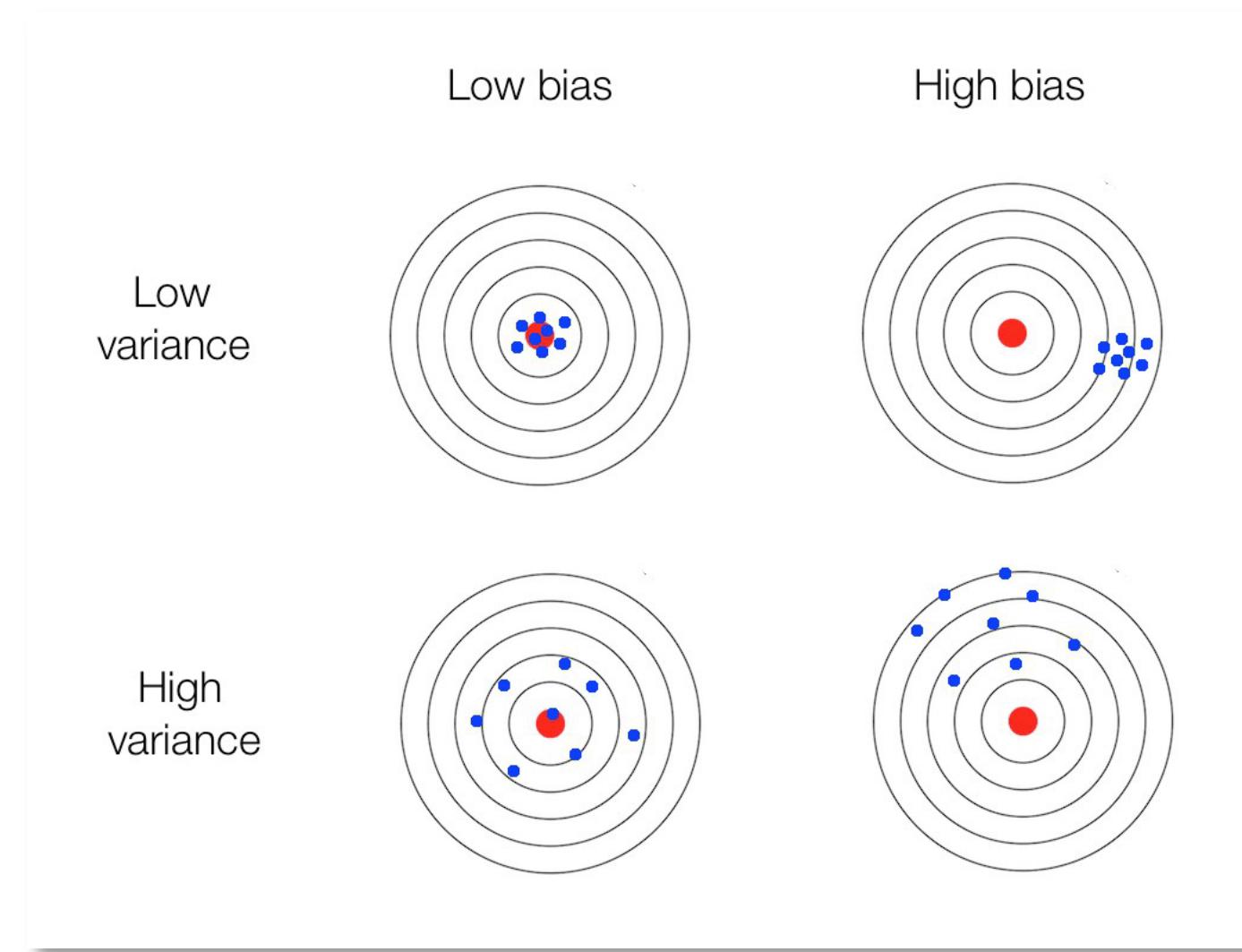
# Today: a story in two parts



# Part 1 – Bias

# Some definitions of bias

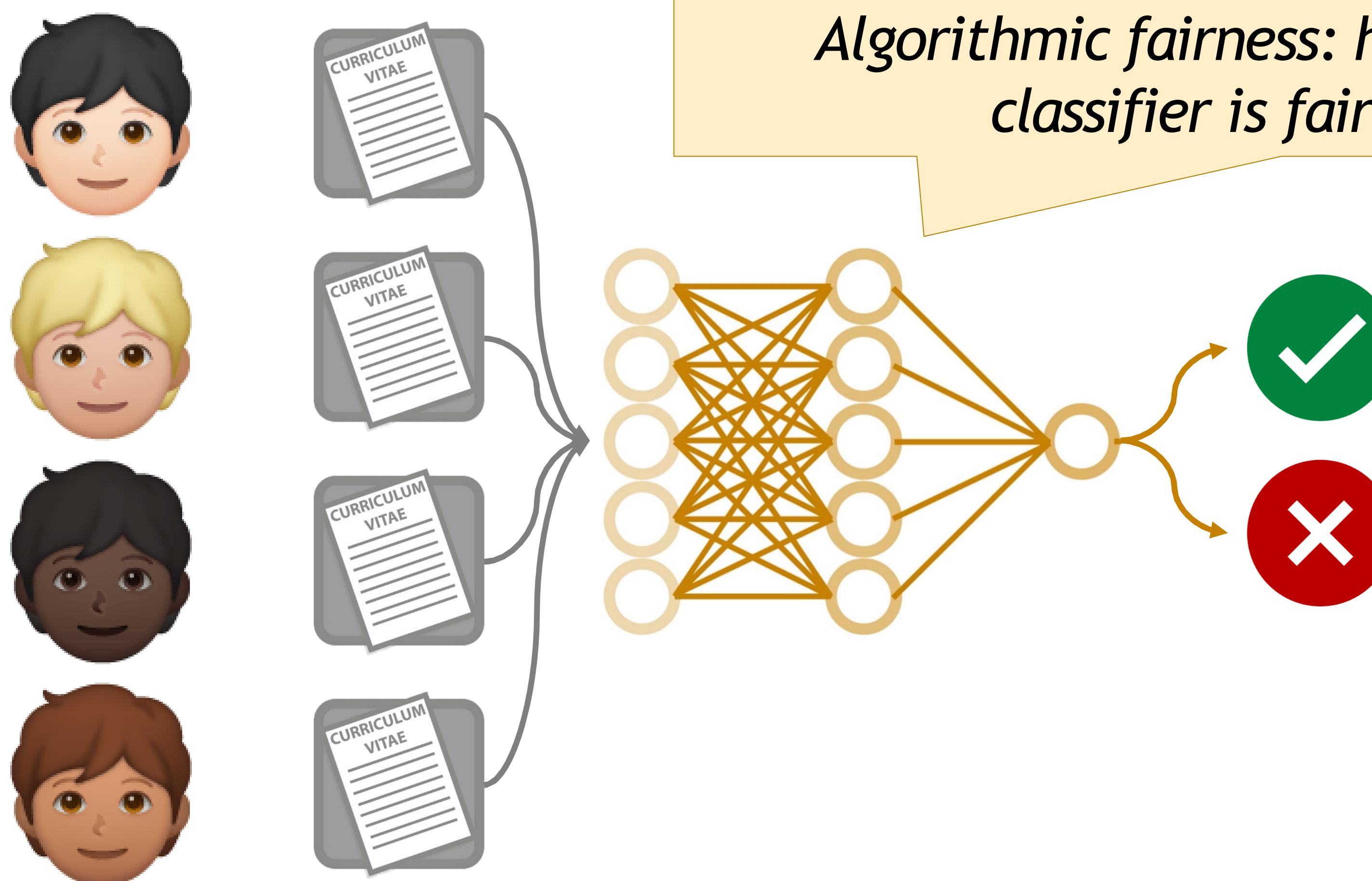
- Bias [statistics]: systematic tendency causing differences between model estimates / predictions
- Bias [general]: “*disproportionate weight in favor of or against an idea or thing, usually in a way that is closed-minded, prejudicial, or unfair*” – Wikipedia



Presence of bias  $\simeq$  absence of fairness  
 Algorithmic fairness: attempts to correct biases in ML systems  
 But... how is fairness defined?

# Algorithmic fairness

Let's assume a toy task: given a resumé, predict whether a candidate is qualified



*Algorithmic fairness: how do you know if your classifier is fair (e.g., w.r.t. ethnicity)?*

Evaluation:  
Accuracy  
Precision  
Recall  
F-1 score

# Fairness metrics

- Accuracy quality: a classifier is fair if the people from different groups have the same accuracy
- Statistical parity: groups should have the same probability of being assigned positive class



*Accuracy*



*Accuracy*



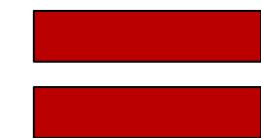
*Accuracy*



*Accuracy*



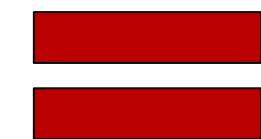
$p(\checkmark | \text{Person A})$



$p(\checkmark | \text{Person B})$



$p(\checkmark | \text{Person C})$



$p(\checkmark | \text{Person D})$

# Equalized odds criterion [Hardt et al '16]

A classifier  $c$  is fair if the *false positive (FP)* and *true positive (TP)* rates are the same for different groups

- *False positives*

$$p(c(\text{🟡}) = \checkmark \mid l(\text{🟡}) = \times) =$$

$$p(c(\text{🟤}) = \checkmark \mid l(\text{🟤}) = \times)$$

- *True positives*

$$p(c(\text{🟡}) = \checkmark \mid l(\text{🟡}) = \checkmark) =$$

$$p(c(\text{🟤}) = \checkmark \mid l(\text{🟤}) = \checkmark)$$

		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

# Other fairness metrics

- *Treatment equality*
  - Ratio of false negatives and false positives should be the same for groups
- *Fairness through unawareness*
  - Models should not employ sensitive attributes when making decisions
- *Causality-based*
  - *Counterfactual fairness*: outcome of the classifier would not changed if the sensitive attribute (e.g., race) were the only thing changed
- Many more...
  - [https://en.wikipedia.org/wiki/Fairness\\_\(machine\\_learning\)](https://en.wikipedia.org/wiki/Fairness_(machine_learning))
  - <https://fairmlbook.org/>

## FAIRNESS AND MACHINE LEARNING

### Limitations and Opportunities

*Solon Barocas, Moritz Hardt, Arvind Narayanan*

### CONTENTS

[PREFACE](#)

[ACKNOWLEDGMENTS](#)

1 [INTRODUCTION](#) [PDF](#)

2 [WHEN IS AUTOMATED DECISION MAKING LEGITIMATE?](#) [PDF](#)

We explore what makes automated decision making a matter of normative concern, situated in bureaucratic decision making and its mechanical application of formalized rules.

3 [CLASSIFICATION](#) [PDF](#)

We introduce formal non-discrimination criteria in a decision-theoretic setting, establish their relationships, and illustrate their limitations.

4 [RELATIVE NOTIONS OF FAIRNESS](#) [PDF](#)

W W W . F A I R M L B O O K . O R G

# Other fairness metrics

- *Treatment equality*
  - Ratio of false negatives and false positives should be the same for groups
- *Fairness through unawareness*
  - Models should not employ sensitive attributes when making decisions

FAIRNESS AND MACHINE LEARNING

Limitations and Opportunities

*Solon Barocas, Moritz Hardt, Arvind Narayanan*

- But, do these definitions really matter if no harms are caused?  
Many argue that unfairness/bias should be measured in terms of the harms that it causes

- Many more...
  - [https://en.wikipedia.org/wiki/Fairness\\_\(machine\\_learning\)](https://en.wikipedia.org/wiki/Fairness_(machine_learning))
  - <https://fairmlbook.org/>

We explore what makes automated decision making a matter of normative concern, situated in bureaucratic decision making and its mechanical application of formalized rules.

## 3 CLASSIFICATION

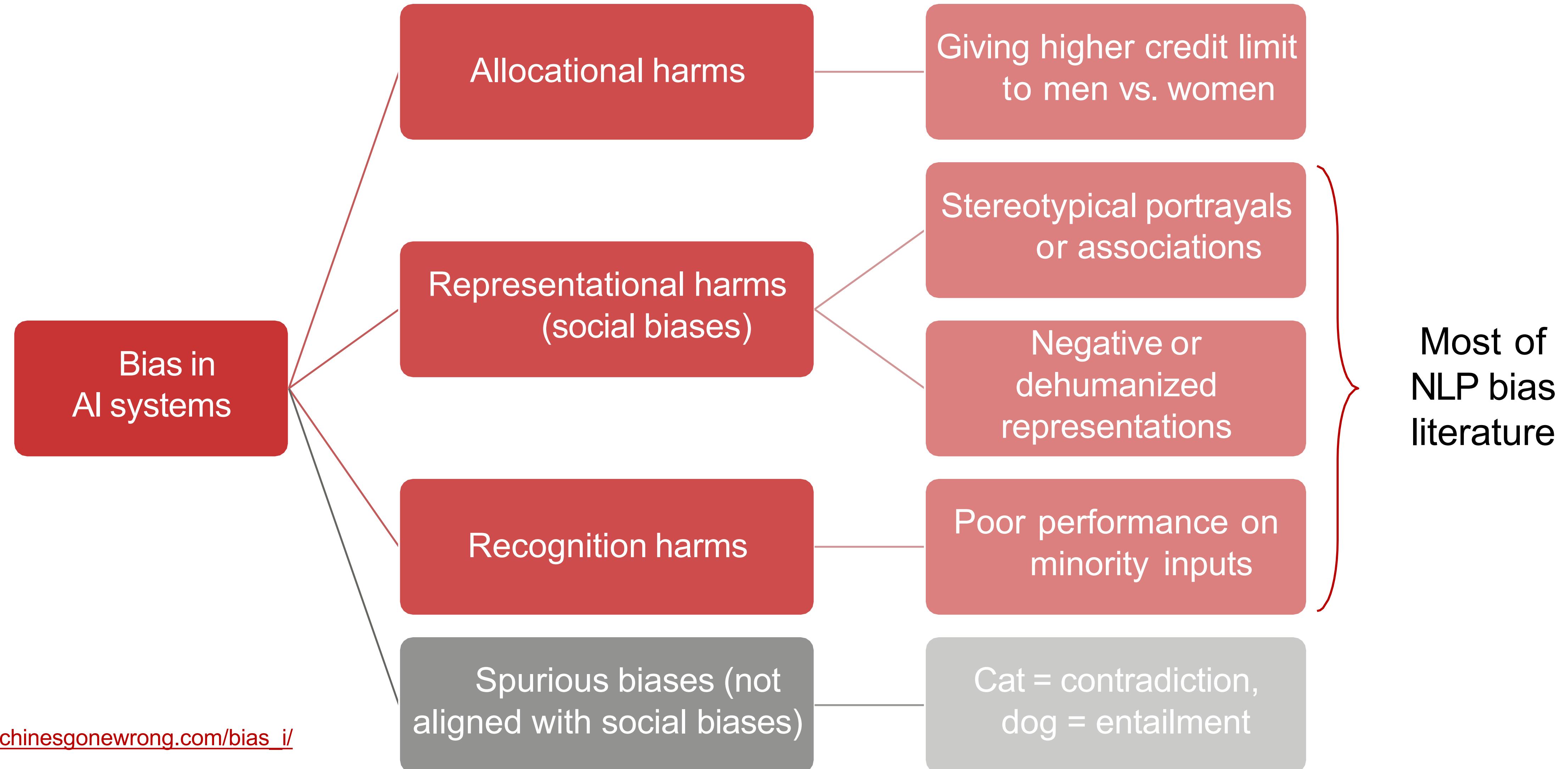
[PDF](#)

We introduce formal non-discrimination criteria in a decision-theoretic setting, establish their relationships, and illustrate their limitations.

## 4 RELATIVE NOTIONS OF FAIRNESS

[PDF](#)

# Bias in terms of the harms it causes



# “Bias” is an overloaded term

- Blodgett et al 2020 examined ~150 NLP papers with “bias” in the title, found that many papers use term “bias” in ill-defined or vague ways

Some recommendations

## Biased behavior

- What kinds of system behaviors are described as “bias”? What are their potential sources (e.g., general assumptions, task definition, data)?

## Harms from biases

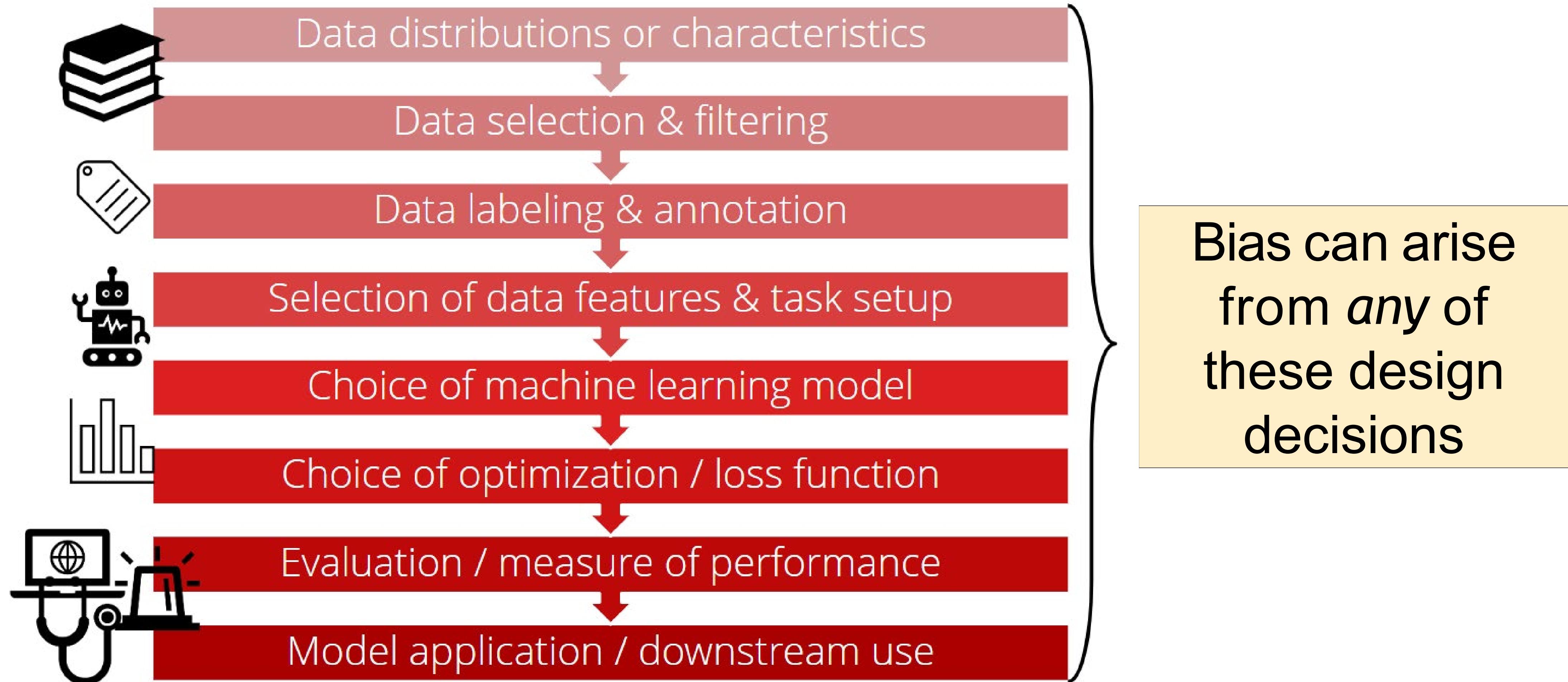
- In what ways are these system behaviors harmful, to whom are they harmful, and why?

## Social values

- What are the social values (obvious or not) that underpin this conceptualization of “bias?”

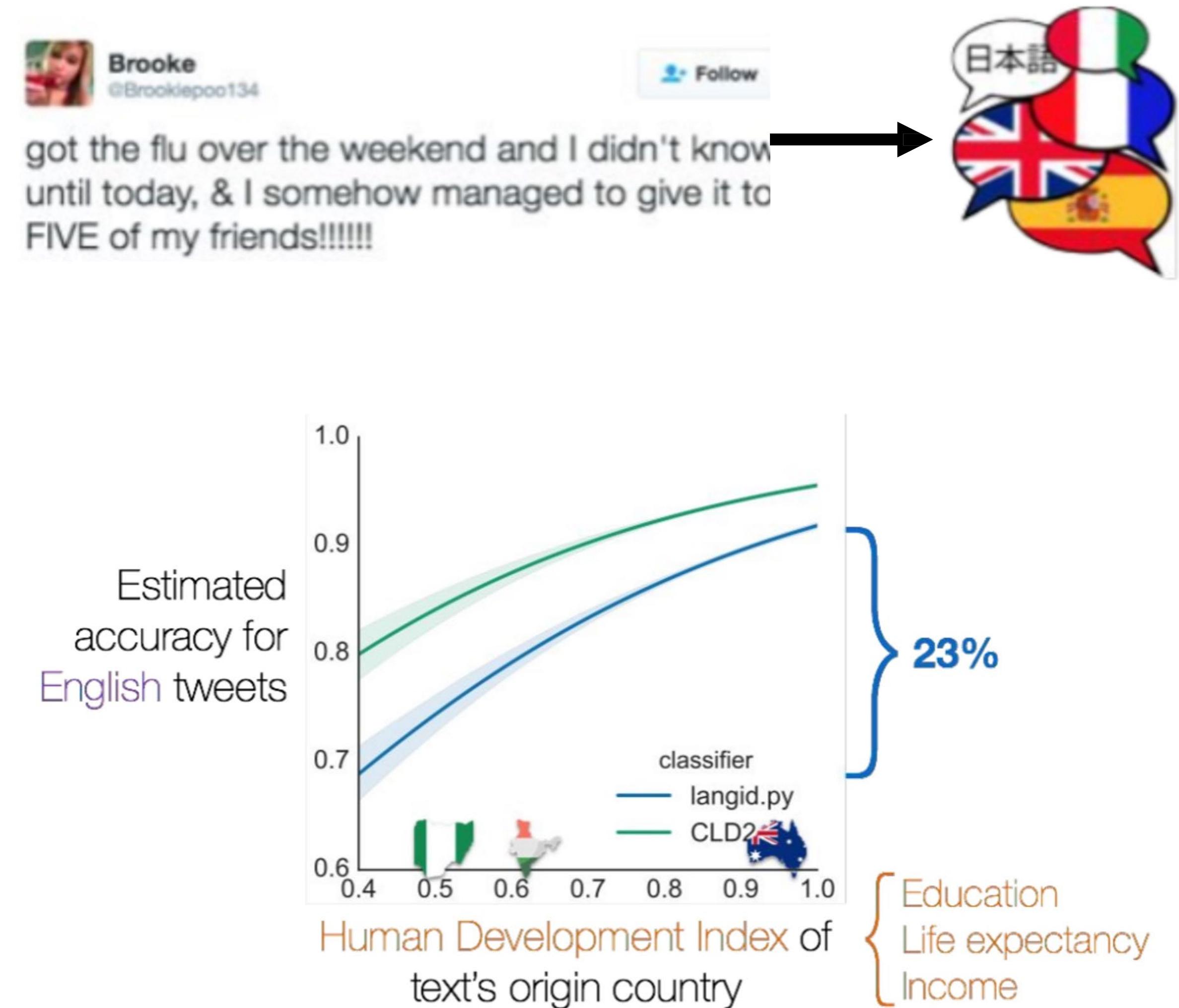
# Where does bias come from?

# Machine learning pipeline



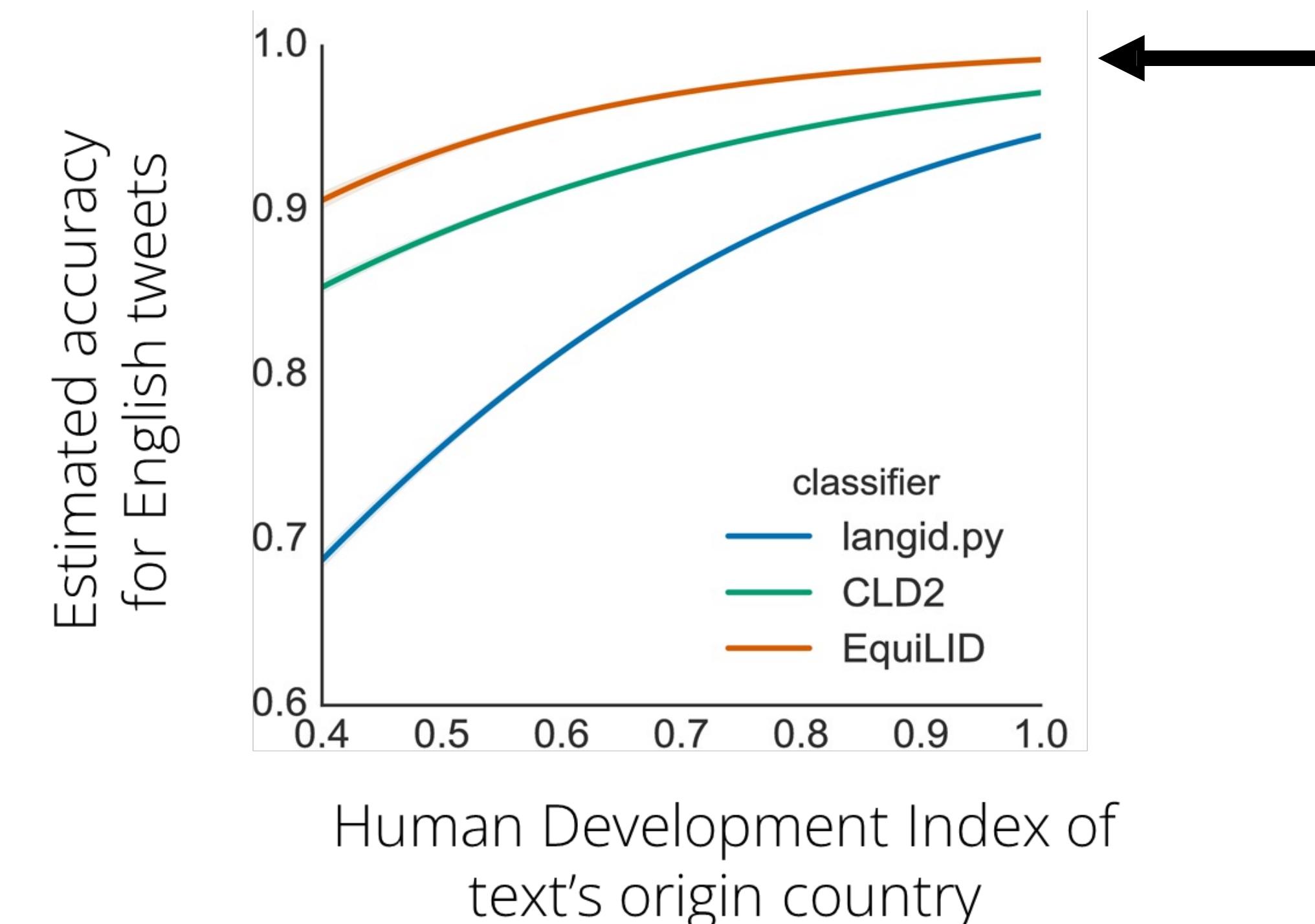
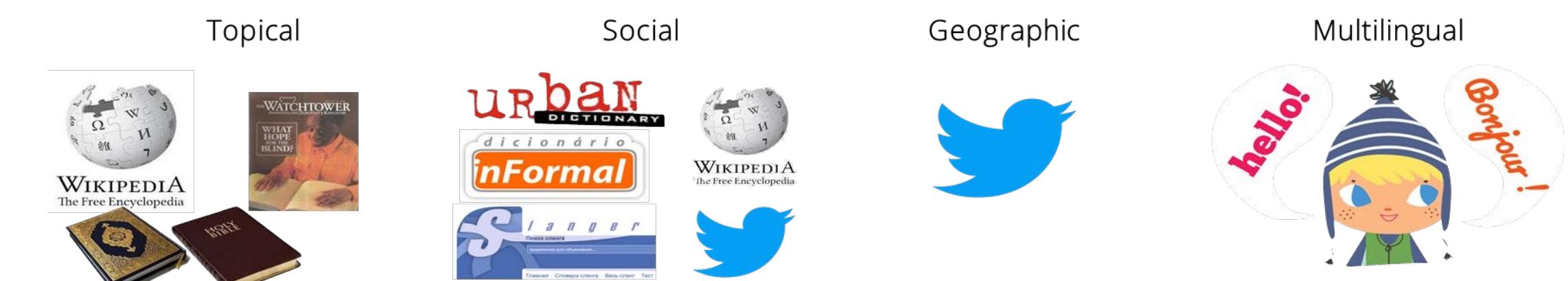
# Example of bias from data: LangID tool

- LangID task: determine which language an input text is in
  - Considered a “*a solved problem* suitable for undergraduate instruction” (McNamee, 2005)
- Often a first step in most NLP and CSS preprocessing pipelines
  - e.g., filtering LLM pretraining data
- *But*, many variations of English in the world
  - *Int'l*: Nigerian English, Indian English, etc.
  - *Within US*: African American English, etc.
- Jurgens et al. (2017) found that accuracy of LID tool correlated with wealth/development level of country; works worse for low HDI countries

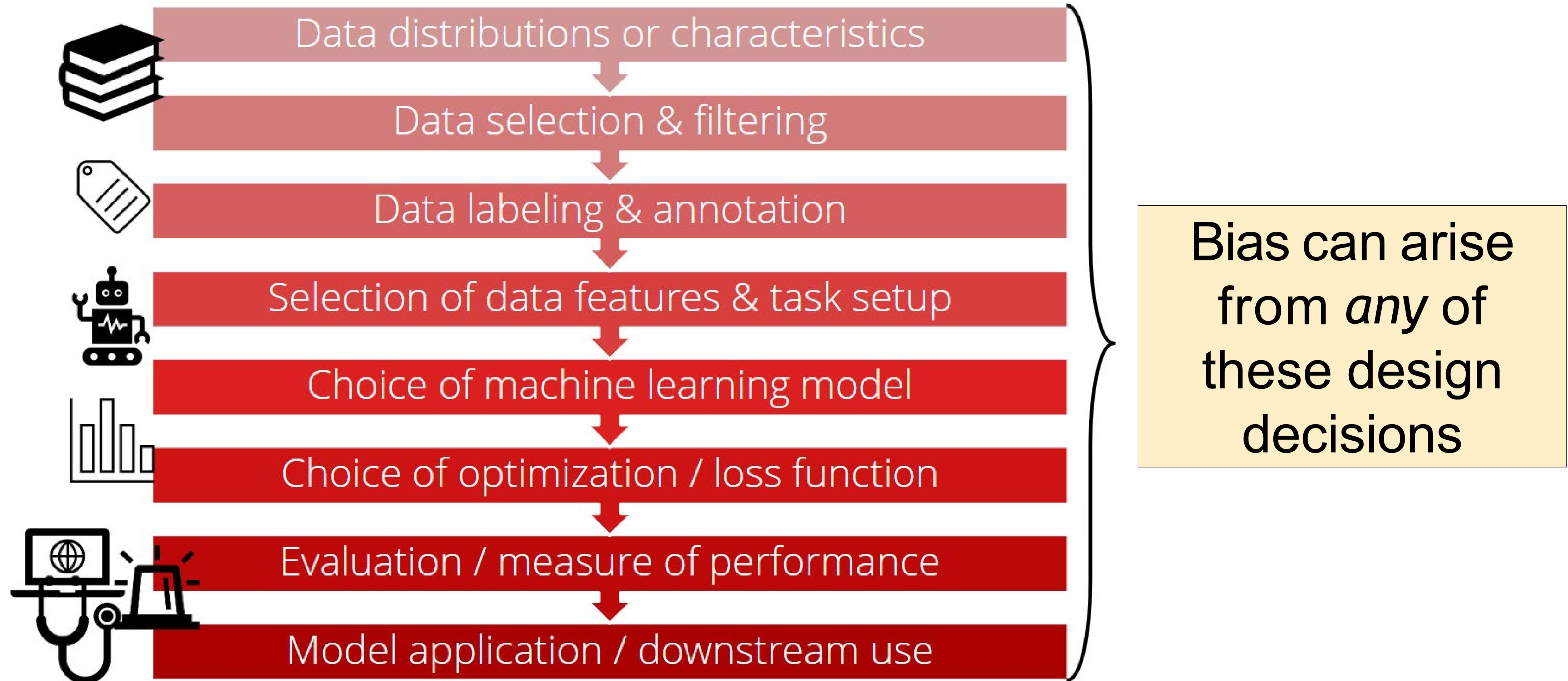


# Example of bias from data: LangID tool (2)

- Jurgens et al (2017) introduce EquiLID
  - Trained by sampling more variety of data, topically, socially, geographically diverse, and even multilingual data
- Find that tool works much better than original LID systems
  - Bonus: even improved accuracy on highly developed countries!
- Takeaway: bias can be mitigated by making better data choices
  - But that's not the only source of bias...

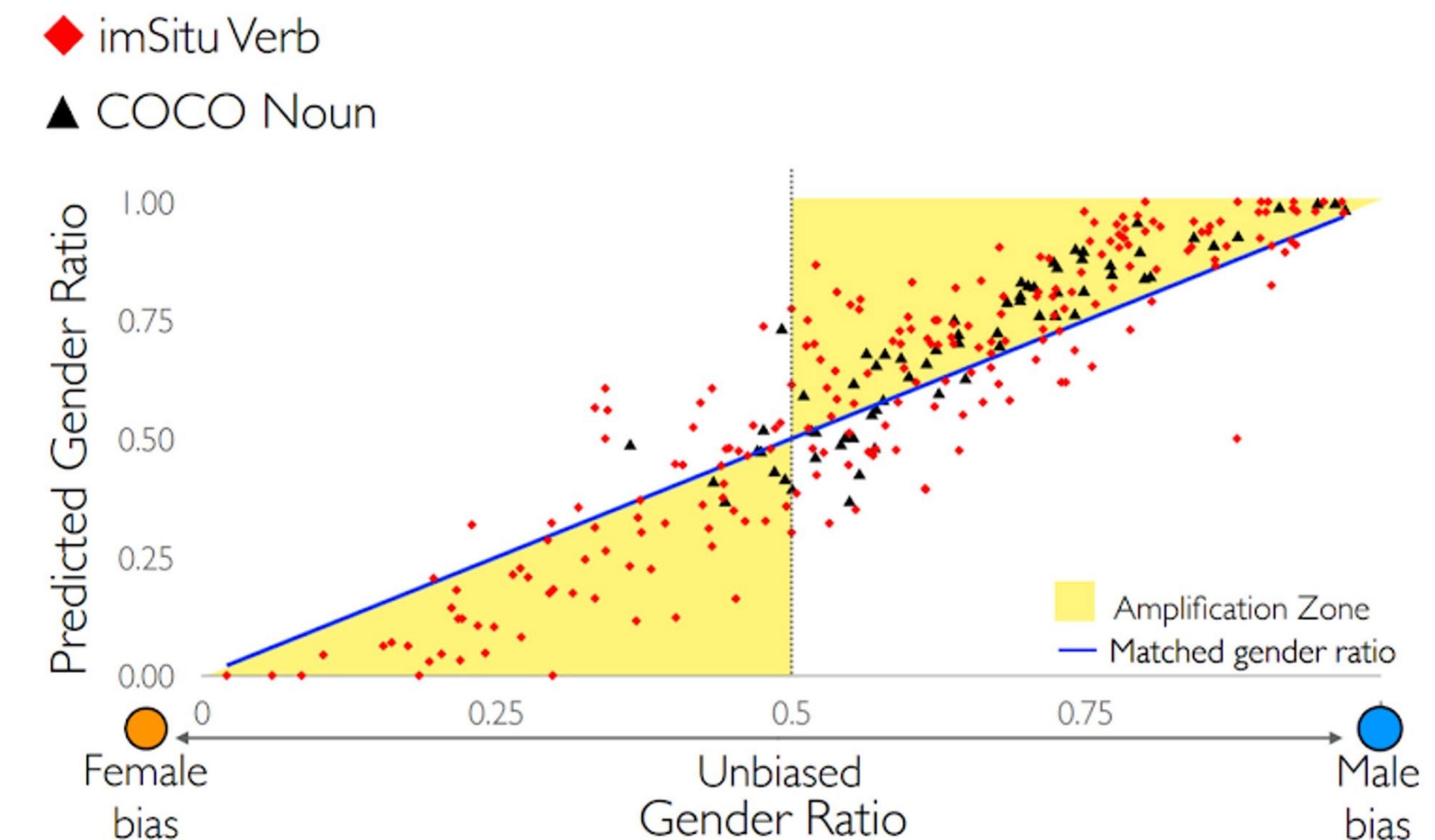
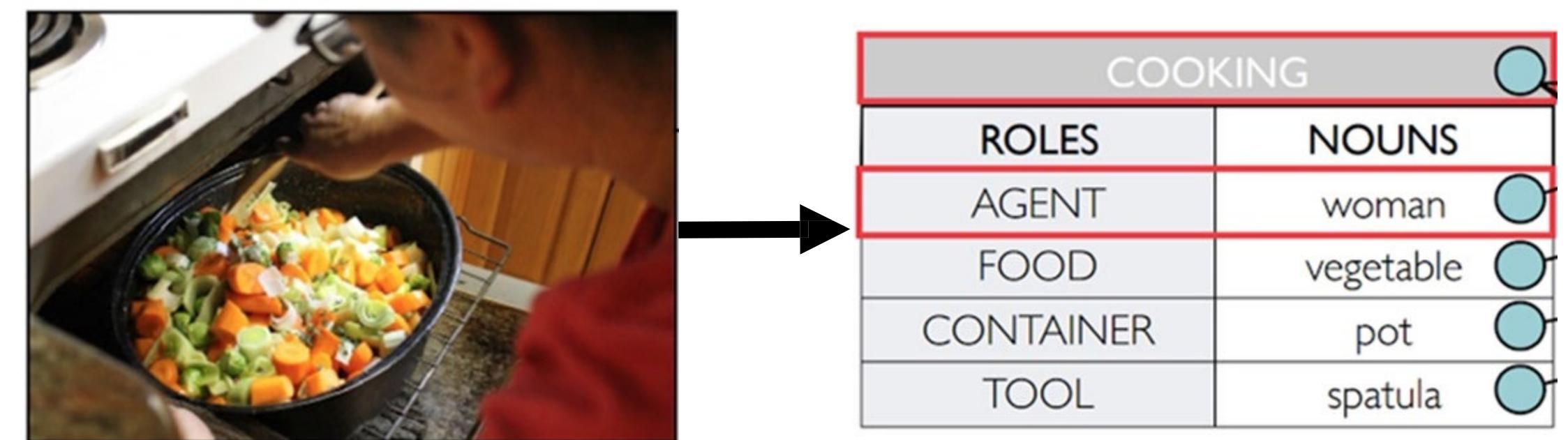


# Machine learning pipeline



# Bias amplification from models

- Zhao et al (2017) examined visual semantic role labeling task
  - Given an image, predict various semantic roles, including agent (person doing the action)
- Found skews in training dataset
  - E.g., 66% of training cooking images had agent=woman
- Found that models *amplified* biases
  - E.g., 84% of test cooking images predicted as agent=woman (~18% men mis-labeled)



# Model biases: mathematical links

- *Competing losses*: objective functions aim to minimize loss globally → learns to predict most frequent class
  - Often at the expense of less frequent classes (e.g. minority groups)
- *Simplicity bias*: neural networks biased towards learning simpler functions [[Valle Pérez et al. 2019](#)]
- Intuitively, if a model has limited learning capacity, makes sense that it learns shortcuts first
  - Shortcuts are often stereotypes or majority biases; e.g., CEOs are men
- Takeaway: ML/optimization choices also affect biases

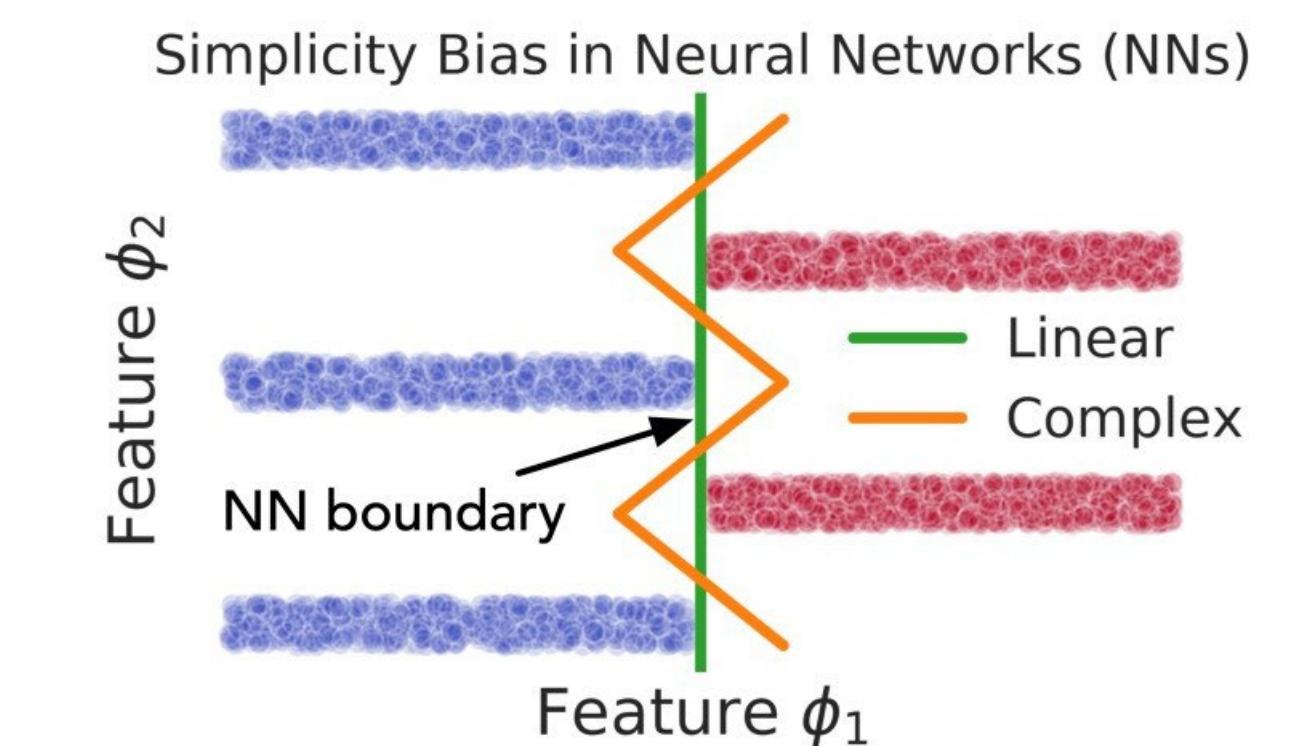
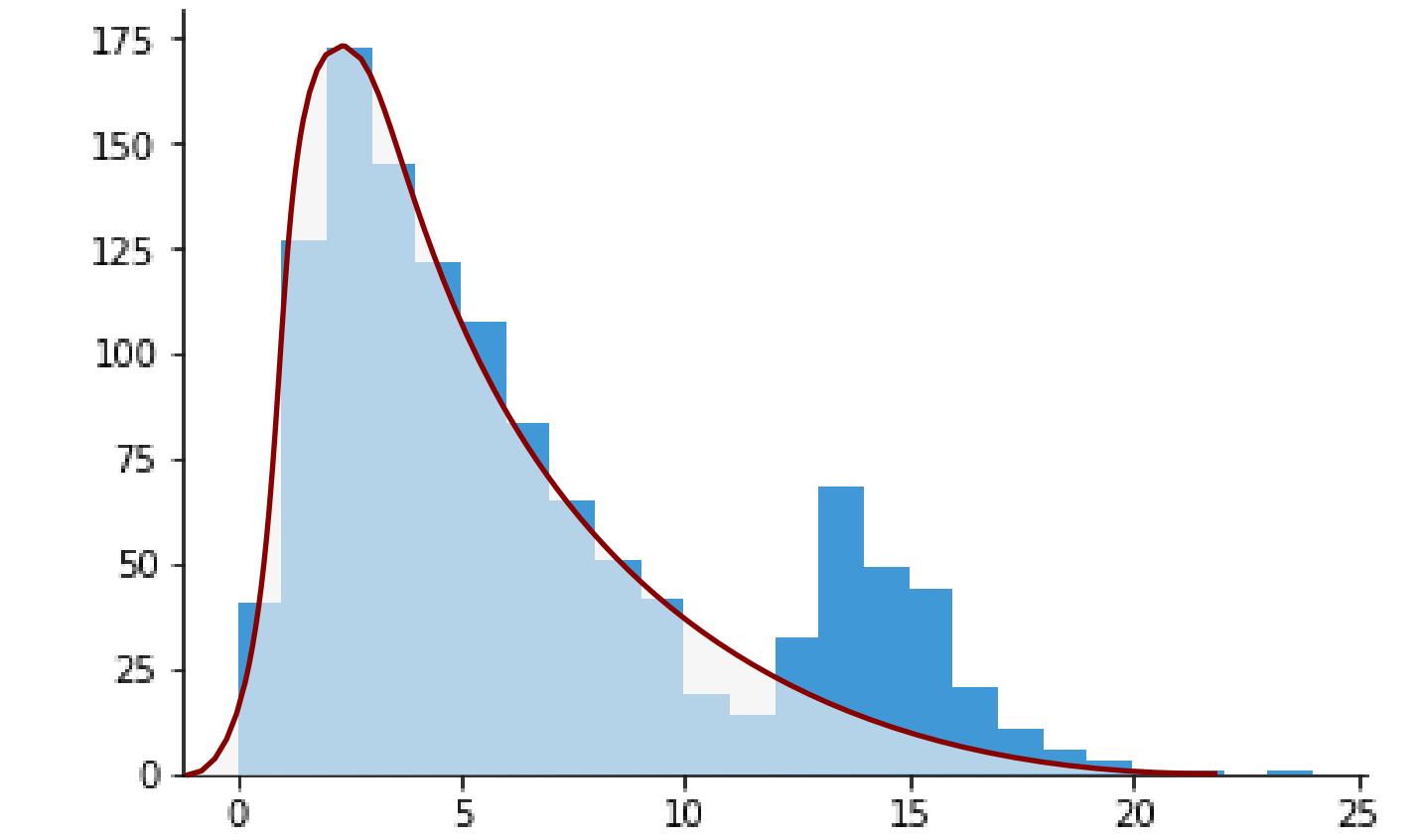
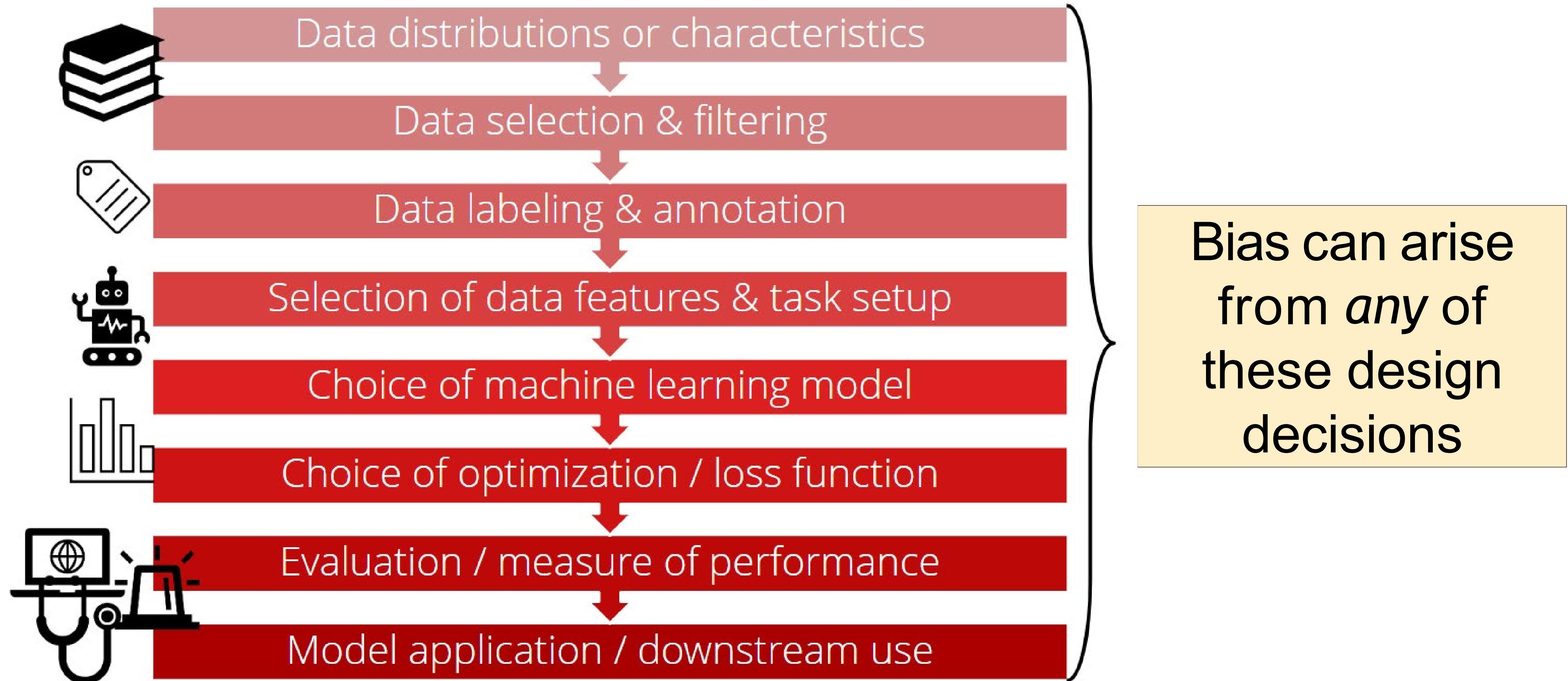


Figure 1: Simple vs. complex features

*Figure from [Shah et al 2020](#)*

# Machine learning pipeline



# Google Translate issue

The image shows two screenshots of the Google Translate mobile app interface, labeled 'Before' and 'After', illustrating a gender-specific translation issue.

**Before:** The user has inputted the Turkish sentence "o bir doktor" (he is a doctor). The English translation is displayed as a single result: "he is a doctor".

**After:** The same input is shown, but the English translation now appears twice: "she is a doctor (feminine)" and "he is a doctor (masculine)". A blue banner at the bottom of the translation box states "Translations are gender-specific. LEARN MORE".

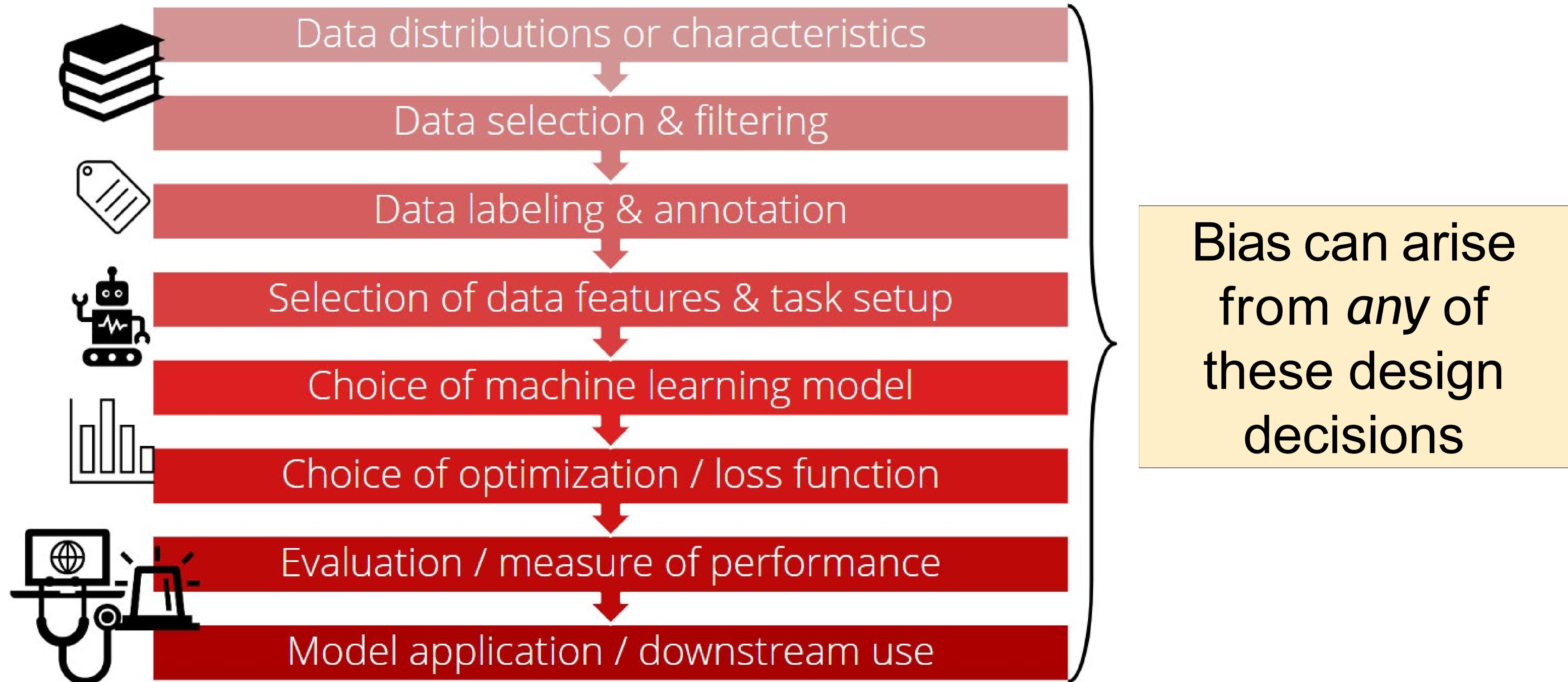
Annotations with red arrows point to specific elements:

- An annotation box on the left points to the input text with the text: "Input in non-gendered language (e.g., Turkish)".
- An annotation box below the input points to the single translation result with the text: "Stereotypical gender assigned in translation".
- An annotation box on the right points to the double result with the text: "Google's fix: translate twice, with male & female pronouns".

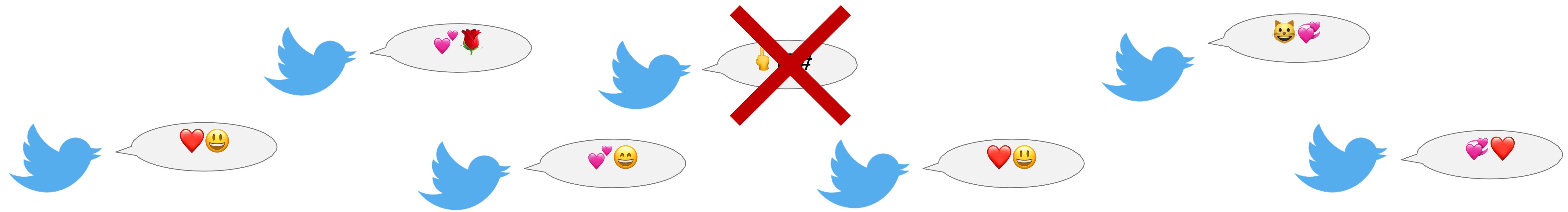
- Takeaways: mitigating bias may involve *system-level changes* to UI, input processing, output formatting, etc. while underlying AI model is similar
- ***Let's discuss:*** what do you think of this approach? What are some possible issues?

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

# Machine learning pipeline

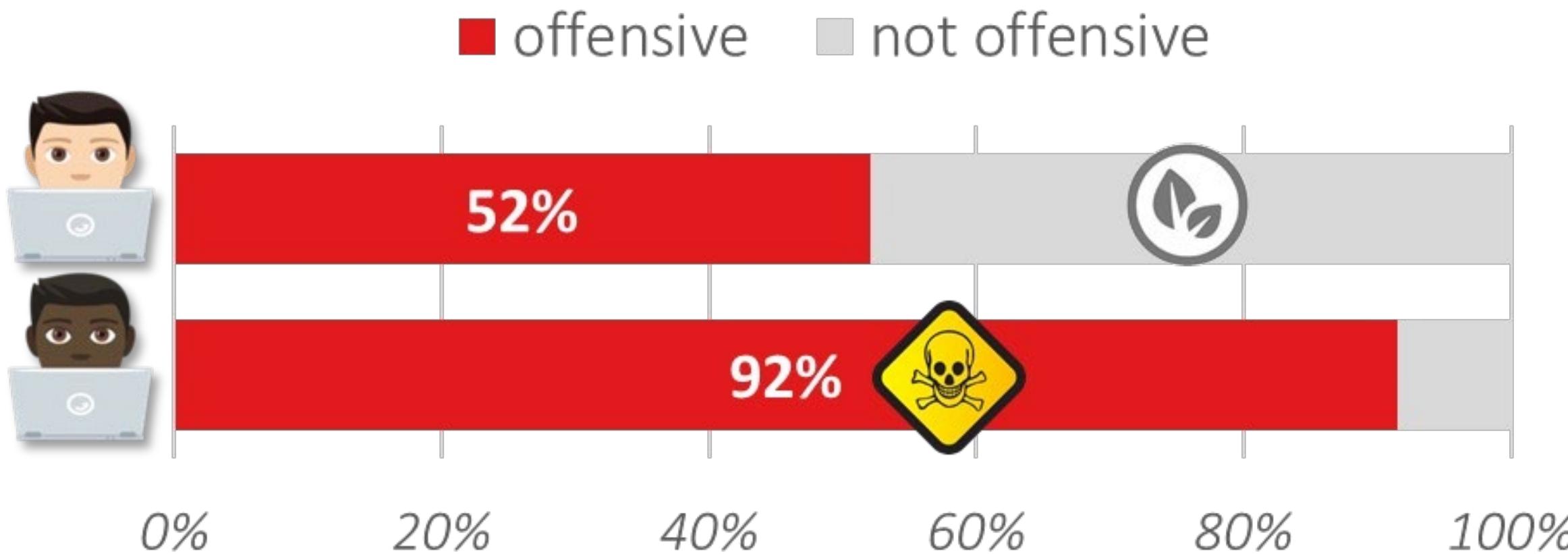


# Hate Speech or Toxic Language Detection



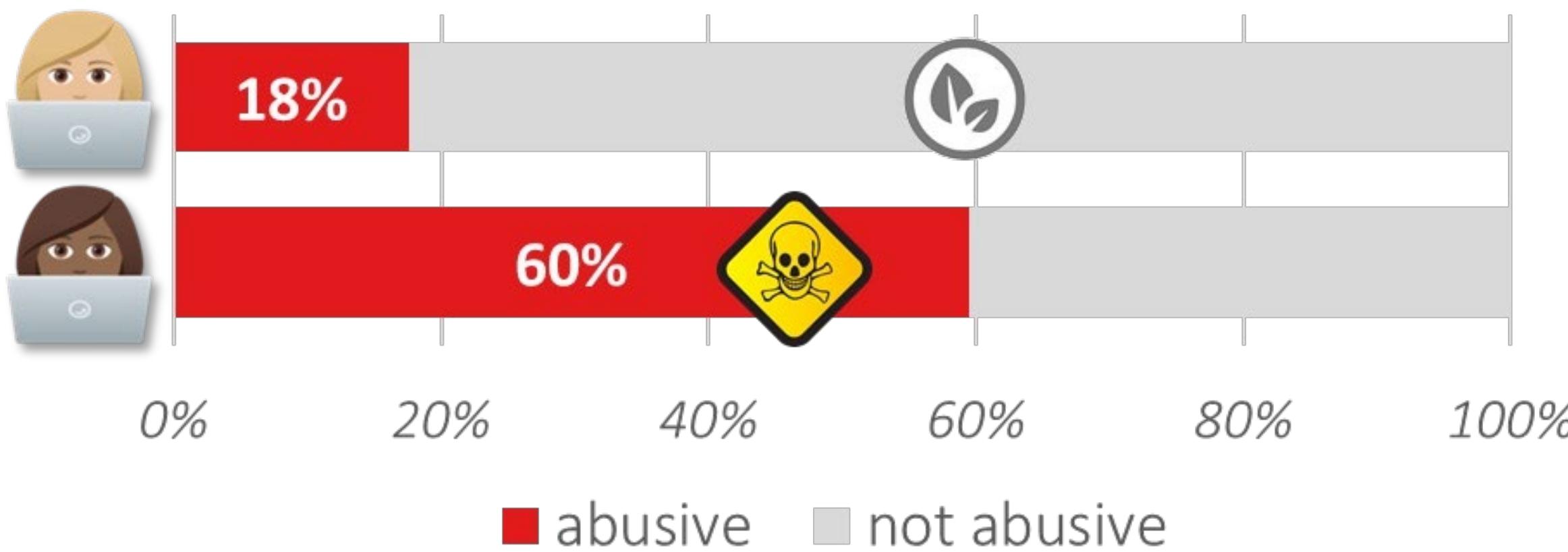
Goal: find and flag hateful or toxic content  
online, to make the internet less toxic

# Racial biases in two popular datasets [Sap et al 2019]



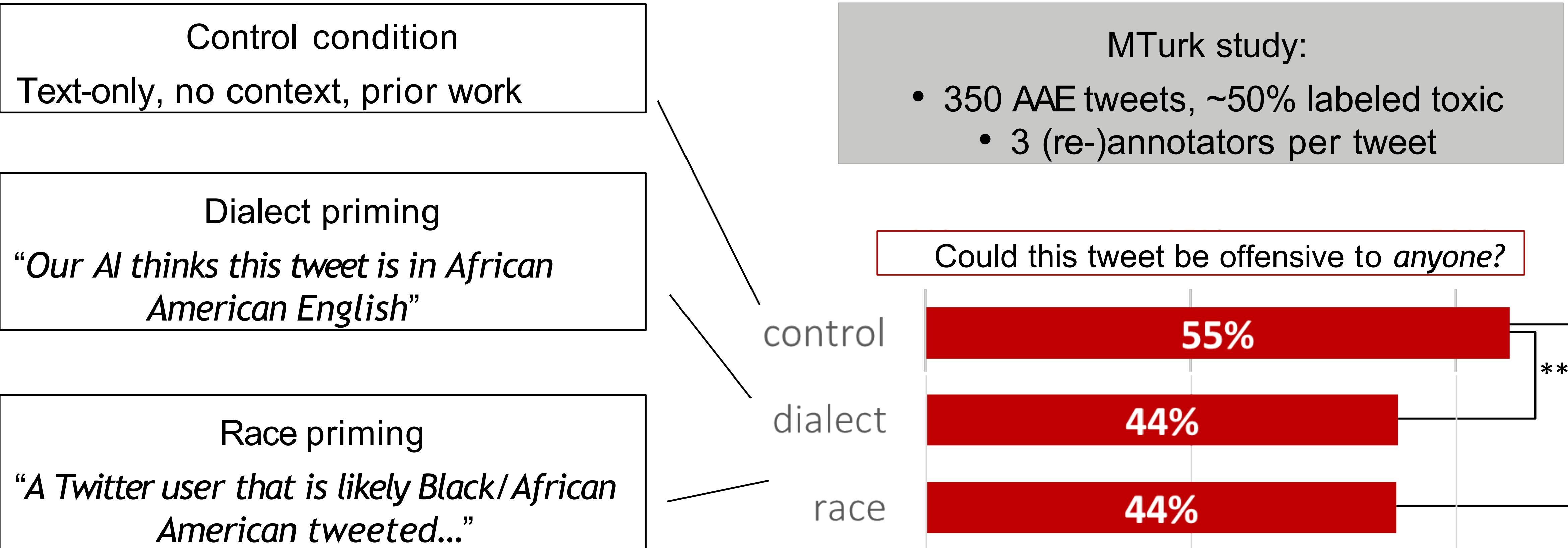
Twt-HATEBASE  
(Davidson et al., 2017)

Both datasets have **biases w.r.t. AAE tweets**



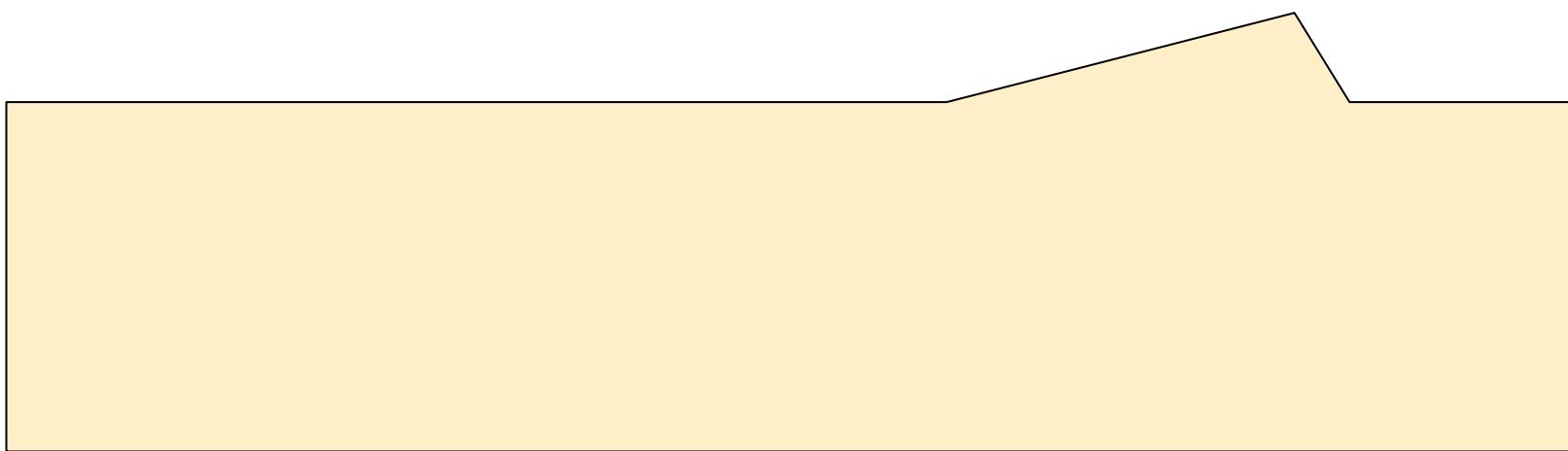
Twt-BOOTSTRAP  
(Founta et al., 2018)

# Enhancing the labeling interface [Sap et al 2019]

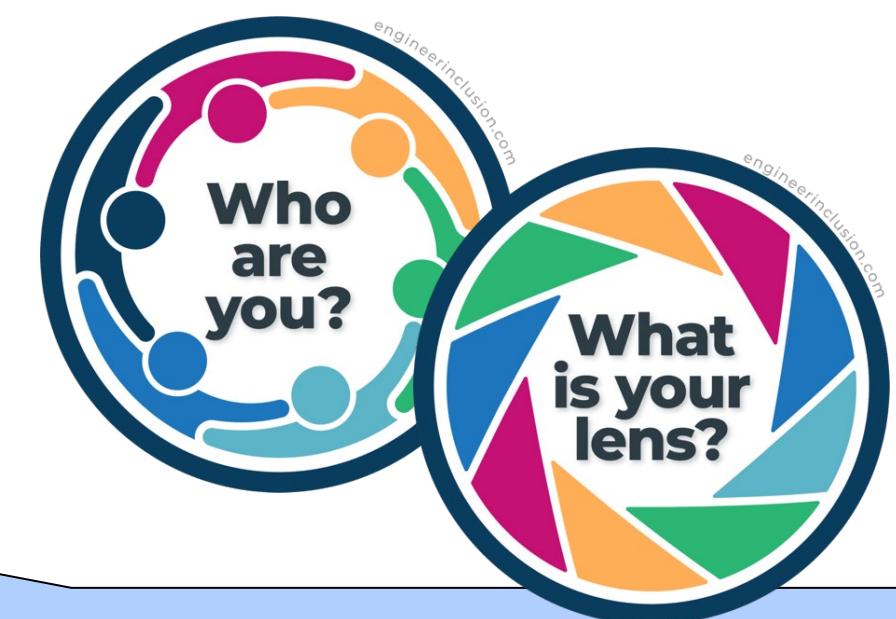


Takeaway: adding social context to labeling mitigated bias

# Why did these biases occur? Why didn't NLP system designers think about these issues beforehand?



The world itself is biased



System designers have our own biases because of their *positionality*, i.e., set of perspectives that we hold due to our lived experiences and identity.

Positionality affects all our choices (e.g., assuming 1-1 mapping between languages and gendered pronouns, assuming toxicity looks the same in different dialects)

# Debiasing AI systems

*Is it even possible?*

# DALLE-2 vs. Gemini



- DALLE-2 generated images were shown to have social biases, later fixed by adding identity keywords to the input prompts (e.g., prompt+“Asian”; [Spakes 2022](#))
  - Gemini generations also shown to have skews
- *Let's discuss:* what do you think of this? How are these two generations different?

# Limits of debiasing

- Gender debiasing doesn't work
  - Breaks down for non-binary genders, racial categories or other social identity types
- Intrinsic debiasing ≠ actual debiasing
  - Finetuning often reintroduces biases
  - Out-of-distribution data often still show biases
- *Real world vs. ideal world*: is reflecting the (biased) status quo the goal? or do we want to build a more fair or just world?
- Justice and fairness go beyond data & model fairness

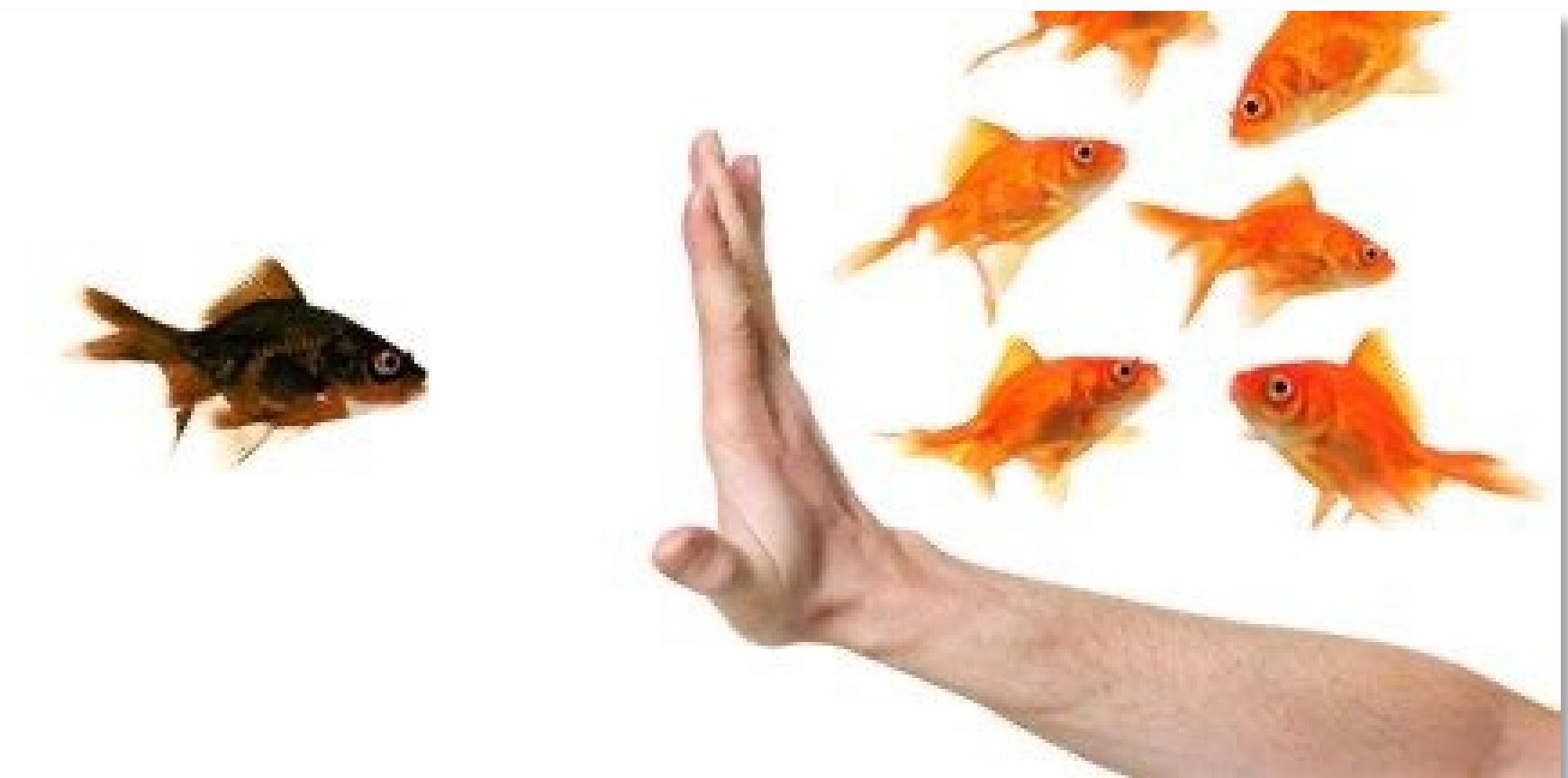


“Lipstick on a pig” paper,  
Gonen & Goldberg 2019

**On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations**

Yang Trista Cao<sup>\*†1</sup>, Yada Pruksachatkun<sup>\*2</sup>, Kai-Wei Chang<sup>2,3</sup> Rahul Gupta<sup>2</sup>  
Varun Kumar<sup>2</sup>, Jwala Dhamala<sup>2</sup>, Aram Galstyan<sup>2,4</sup>

<sup>1</sup>University of Maryland, College Park



A lot of people have understood that we need to have more diverse datasets, but unfortunately, I felt like that's kind of where the understanding has stopped. It's like '*let's diversify our datasets. And that's kind of ethics and fairness, right?*' But you can't ignore social and structural problems.



*Timnit Gebru, PhD*

# Socio-technical view on bias & fairness

- You can have an “fair” NLP/ML model (e.g., facial recognition system)
  - 95% accuracy/error rate on white & Black faces
- But if the system is used by law enforcement, bias creeps in w.r.t. who the system is used on
  - Black people more often arrested, due to racial biases
- Actual error rates are a function of deployment
- Algorithm’s fairness ≠ fairness of treatment

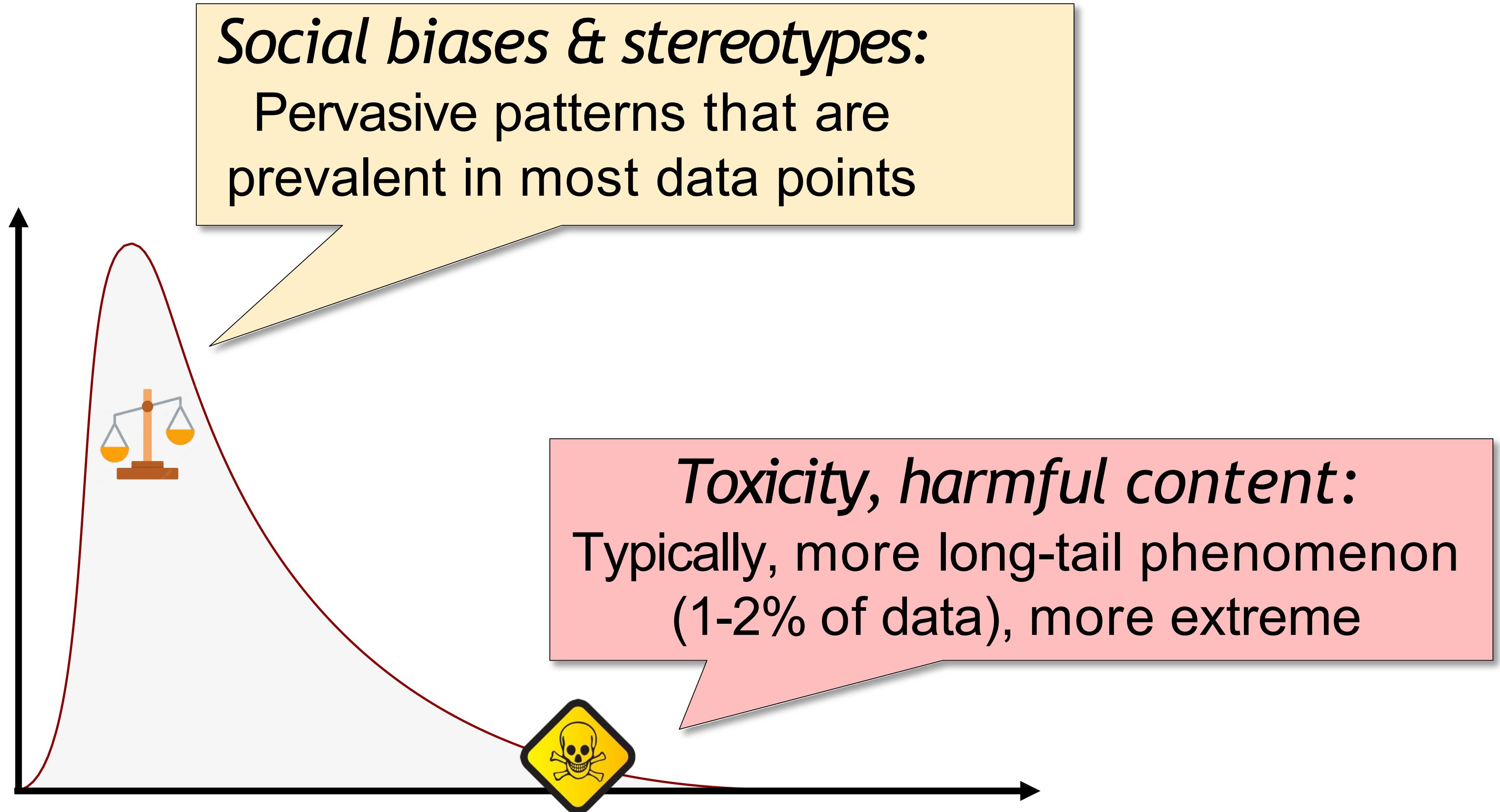


**Black people are more likely to be arrested, charged and killed by police in Toronto, new report finds**

By Scottie Andrew, CNN  
Published 3:15 PM EDT, Wed August 12, 2020

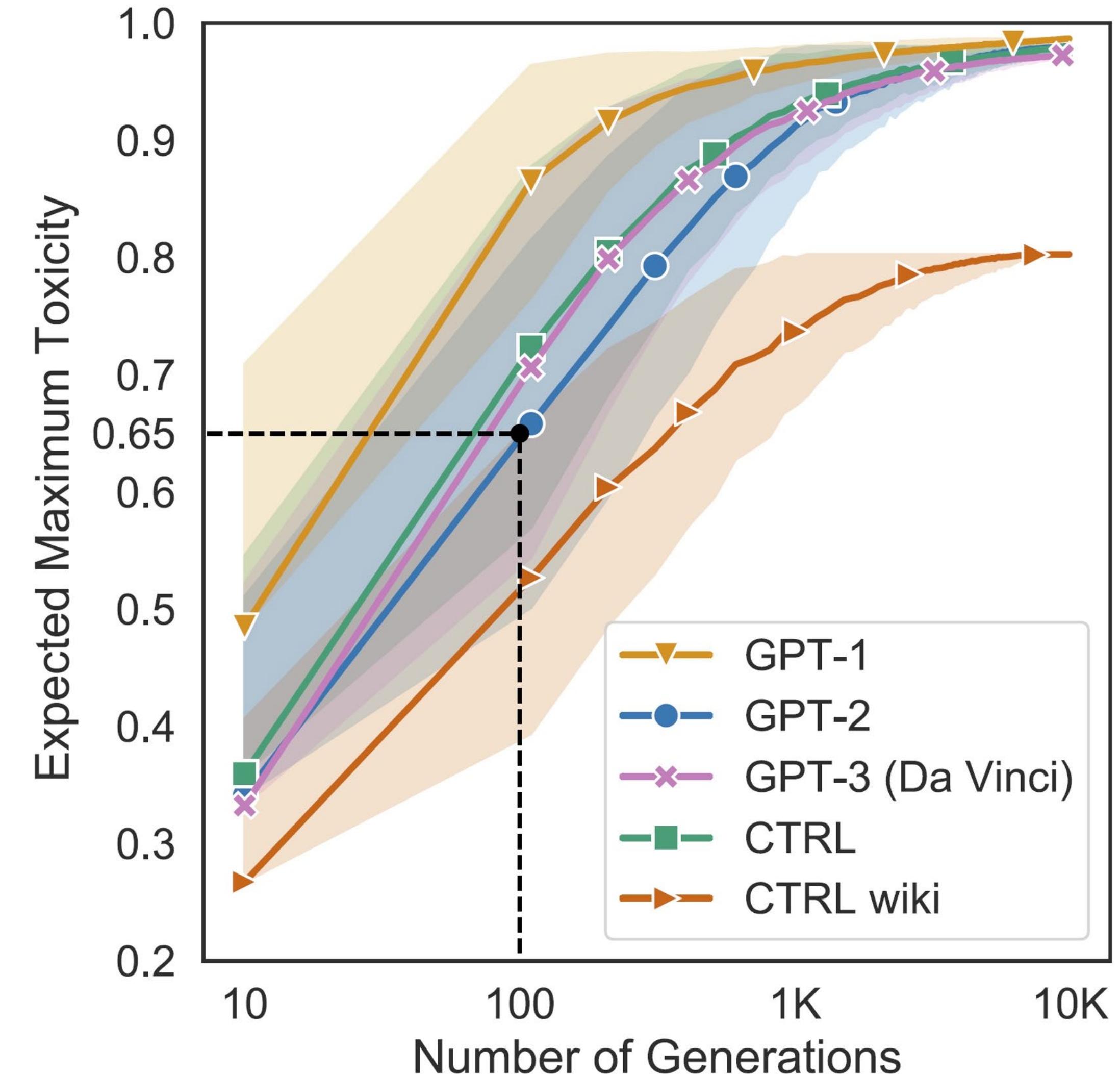
# Part 2: Harmful content & toxicity

# Biases vs. toxicity



# Toxicity in LLMs, how bad is the problem really?

- Gehman et al (2020) introduced concept of *neural toxic degeneration* in LLMs
- Out of a 100 generations sampled from models, at least one toxic sentence
  - 65-70% toxicity from GPT2, GPT3
  - 85% toxicity from GPT1
- Model size affects toxicity: larger models have more toxicity [Touvron et al 2023]



Why are these models learning so much undesirable content?

# Problems with self-supervised pretraining

*“Feeding AI systems on the world’s beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy”*

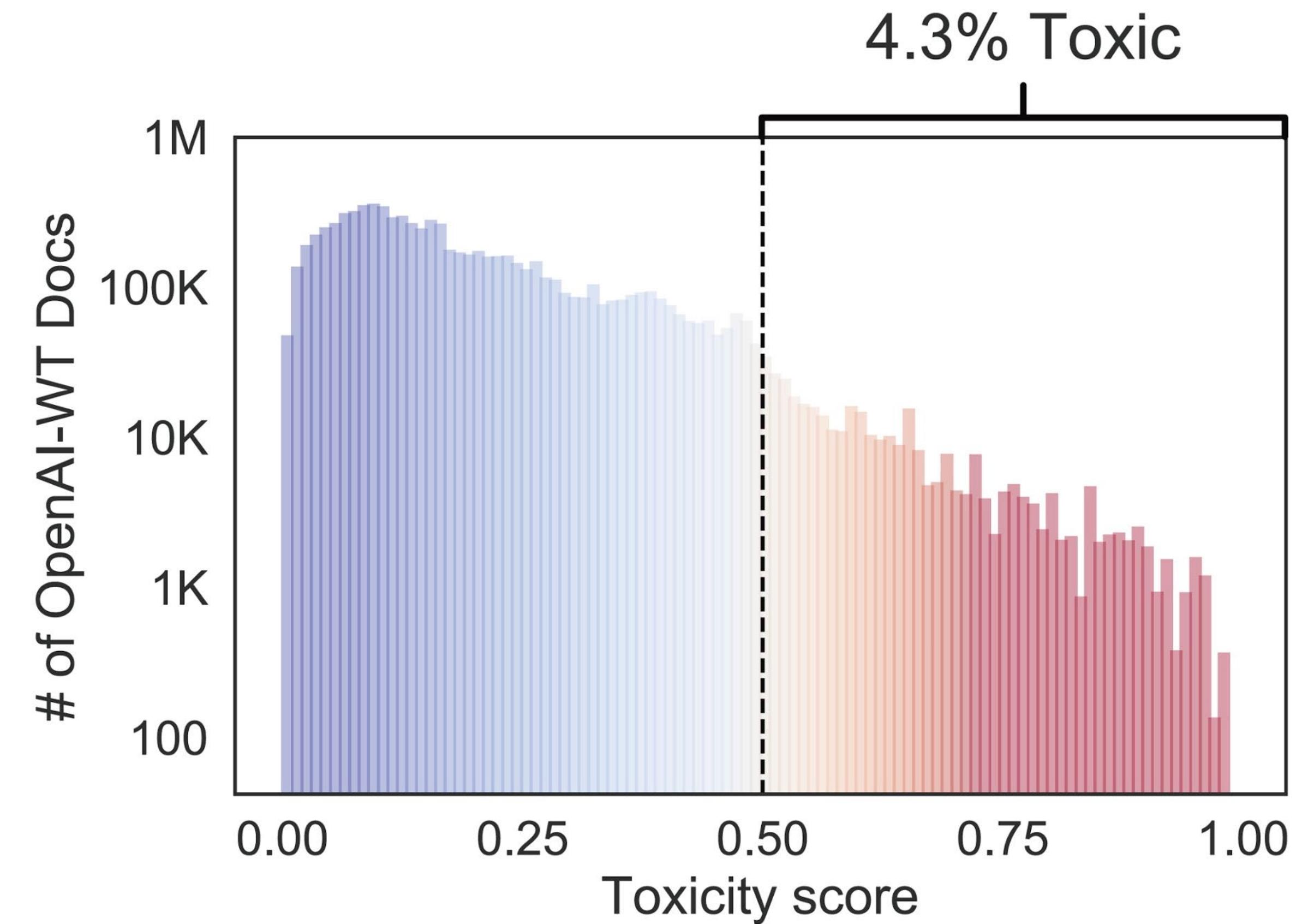


Prof. Ruha Benjamin, PhD

- Recipe: scrape as much pretraining data as you can to train your LM
- Consequence: LM ends up learning toxicity, biases, extremism, hate speech...

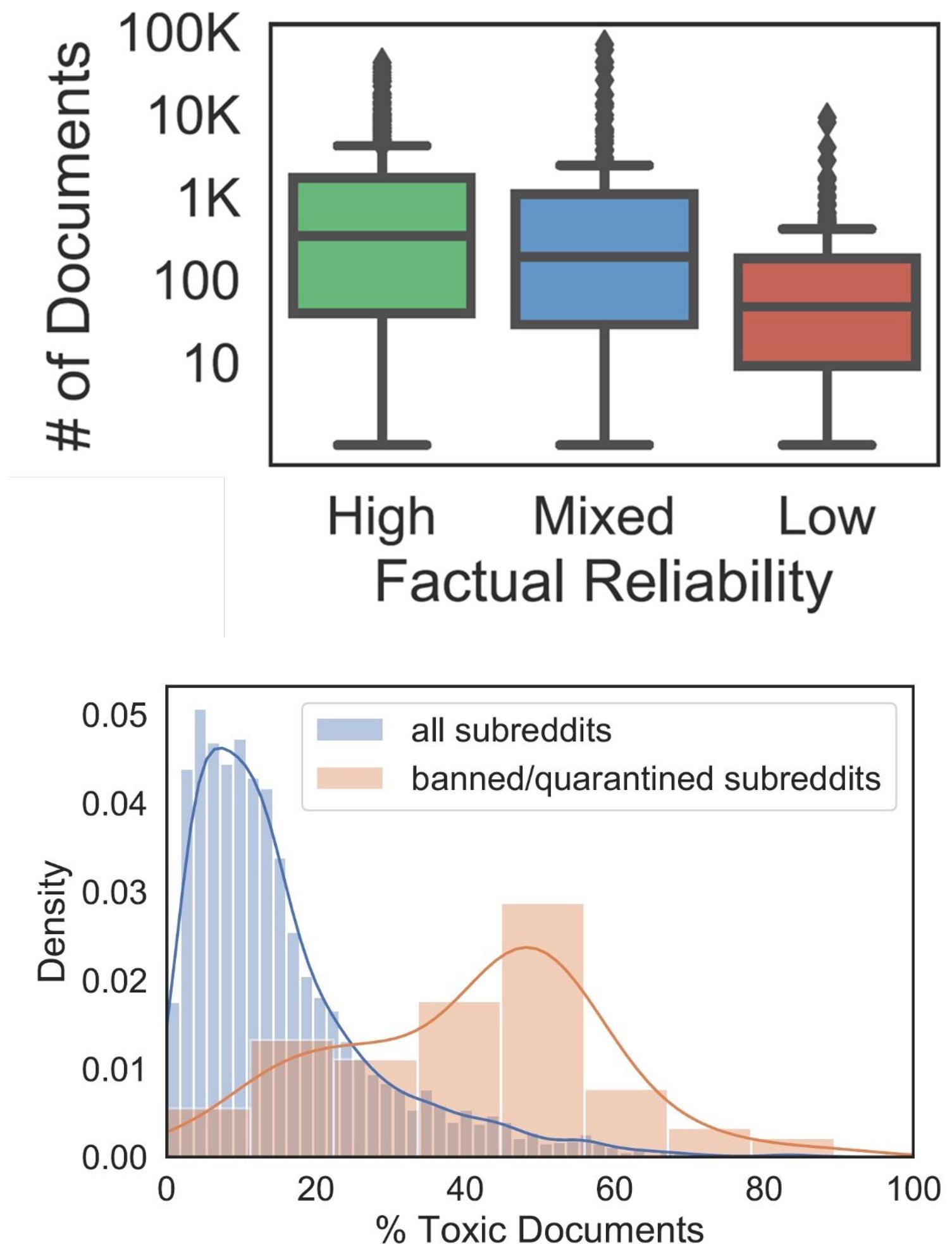
# Toxicity in GPT-2's pretraining data

- Gehman et al (2020) accessed the actual GPT-2 training corpus (OpenAI-WT)
  - 8 million documents, 38Gb of text
  - Outbound links from Reddit posts with Karma $\geq 3$
- Scored it with PerspectiveAPI toxicity
- Found >4% of documents (340,000) are toxic



# Fake news in GPT-2's pretraining data

- Also looked at sources of documents in training data
- Cross-referencing sources of documents with known factual reliability categorization
  - >272K (3.4%) docs from low/mixed reliability sources
- Examining source where document is shared
  - >200K (3%) docs linked from banned/quarantined subreddits, which typically are more toxic docs
- Important to examine training data
  - Can only do that if publicly released!
- *So... need approaches to safeguard your model against this undesirable content, knowledge, and text.*



# How to safeguard your LLMs

# Overview – LLM safeguarding

Safeguards from training data

- Filtering out toxic training data

Safeguards from input prompt classification

- Topic-based filters
- Toxic content detection

Safeguards from instruction-tuning & RLHF

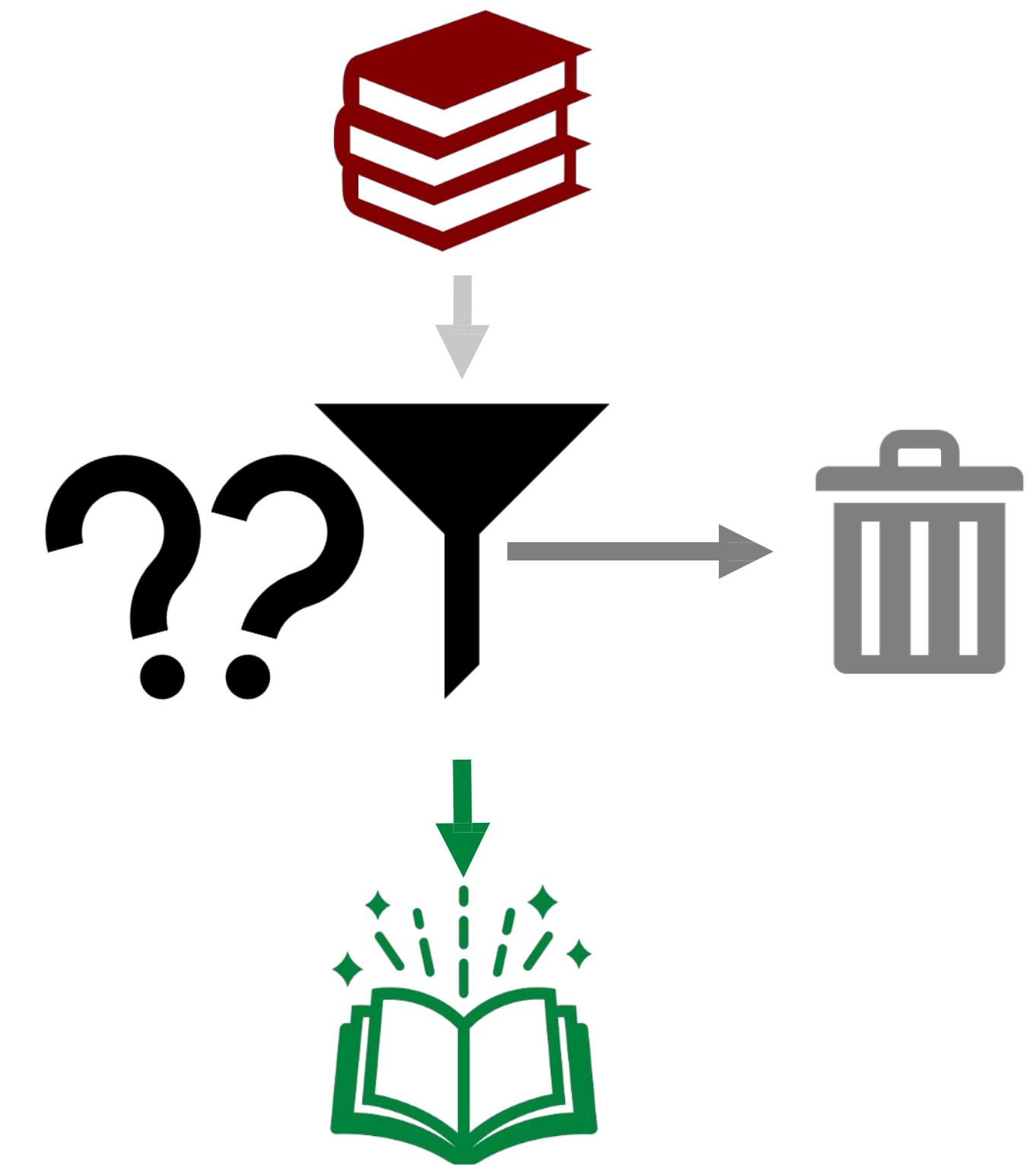
- Write demonstrations for refusing to answer
- RLHF models to prefer non-toxic generations

Safeguards at the output level

- Generate-then-classify
- Controllable text generation

# Dataset filtering

- *Argument:* if you don't want your model to generate toxicity/hate speech, do not train it on such data (garbage in, garbage out)
- *Approach:* data filtering to ensure “high quality”
  - How do you know what is “high quality” ?
    - GPT-2: Reddit “Karma” score as signal
    - T5, BERT: “blocklist” of “bad words”
    - GPT-3: “quality” classifier



# Blocklist of “bad” words

- [“List of Dirty, Naughty, Obscene, or Otherwise Bad Words”](#) originally by Shutterstock employees
  - Meant to prevent words in autocomplete settings
- Has been used by most companies creating LLMs
  - BERT, T5, GPT-2, etc.
- If document contains a “bad” word, remove it from training data
  - F\*ck, sh\*t, sex, vagina, viagra, n\*gga, f\*g, b\*tch, etc.
- *Let's discuss:* what are issues with this?
  - Strong risk of over-deleting bio, legal, minority content

= WIRED

TOM SIMONITE BUSINESS FEB 4, 2021 7:00 AM

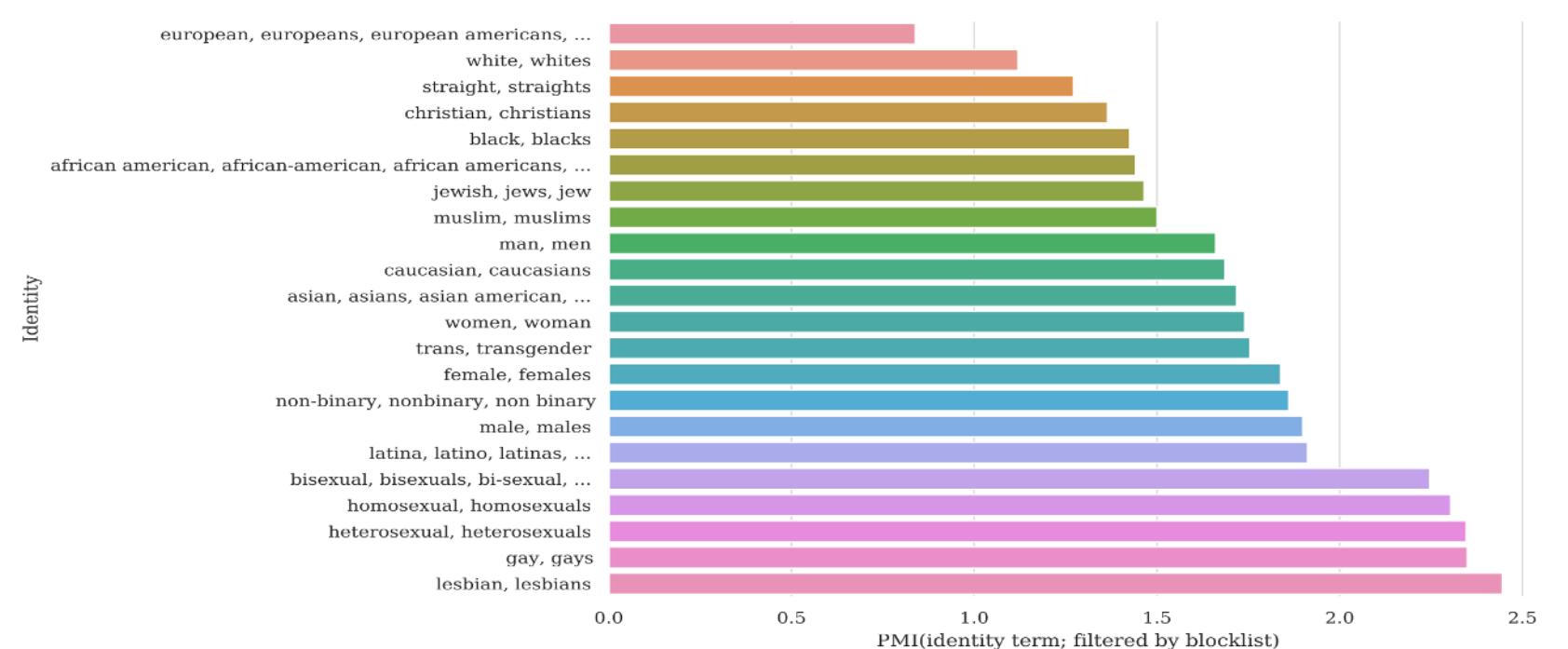
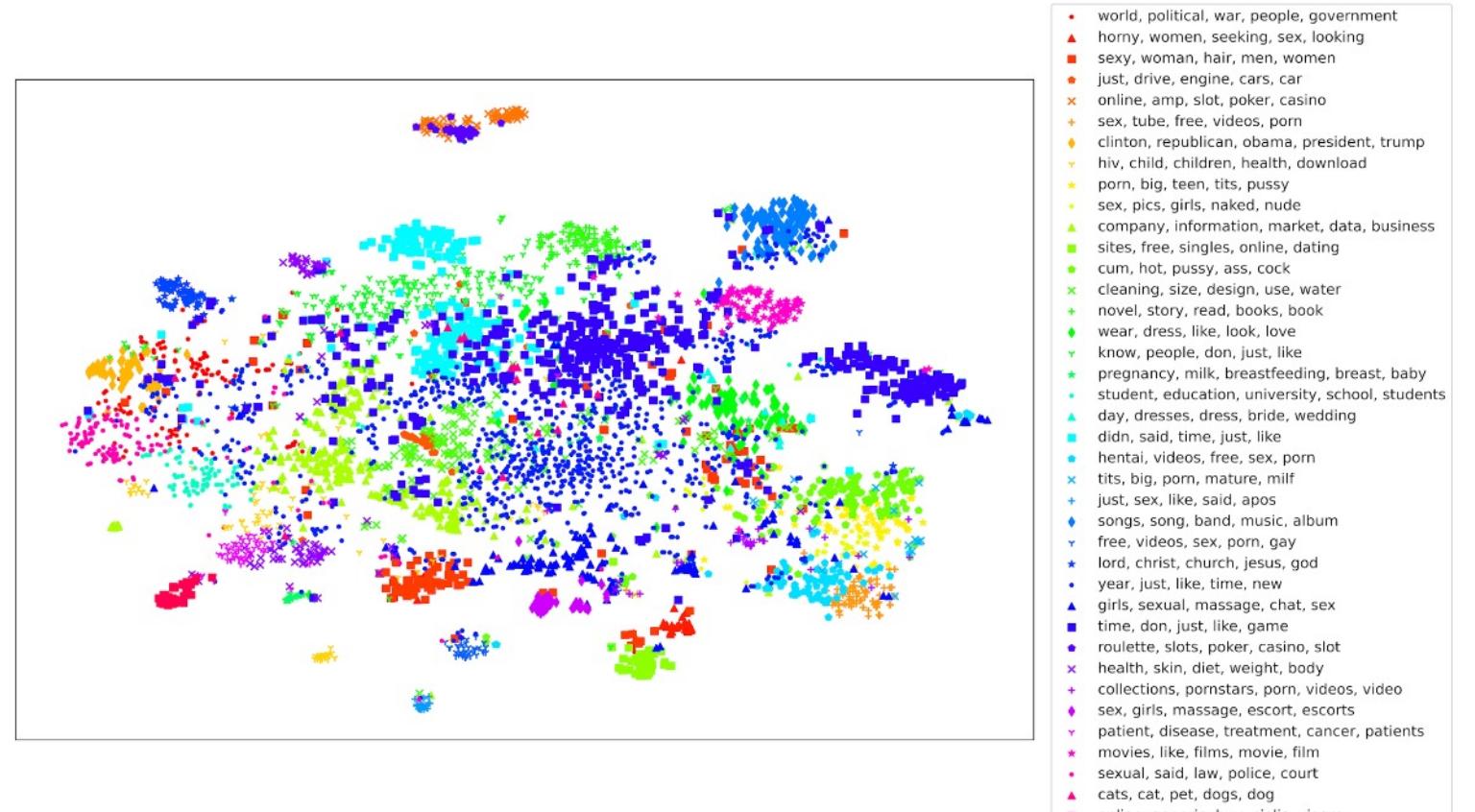
## AI and the List of Dirty, Naughty, Obscene, and Otherwise Bad Words

It started as a way to restrict autocompletes on Shutterstock. Now it grooms search suggestions on Slack and influences Google's artificial intelligence research.



# Effect of “bad word” blocklist filtering

- Dodge et al examined the effect of blocklist filtering on the C4 corpus
- When looking at 100k documents that were excluded due to “bad words”
  - Found only 31% related to porn/explicit sex
  - Remaining was biology, medicine, legal
- Also examined the effect on which minority identities were removed
  - Found queer/LGBTQ identity terms removed more
- Examined dialects removed due to “bad words”
  - Found AAE, Hispanic English more likely to be removed

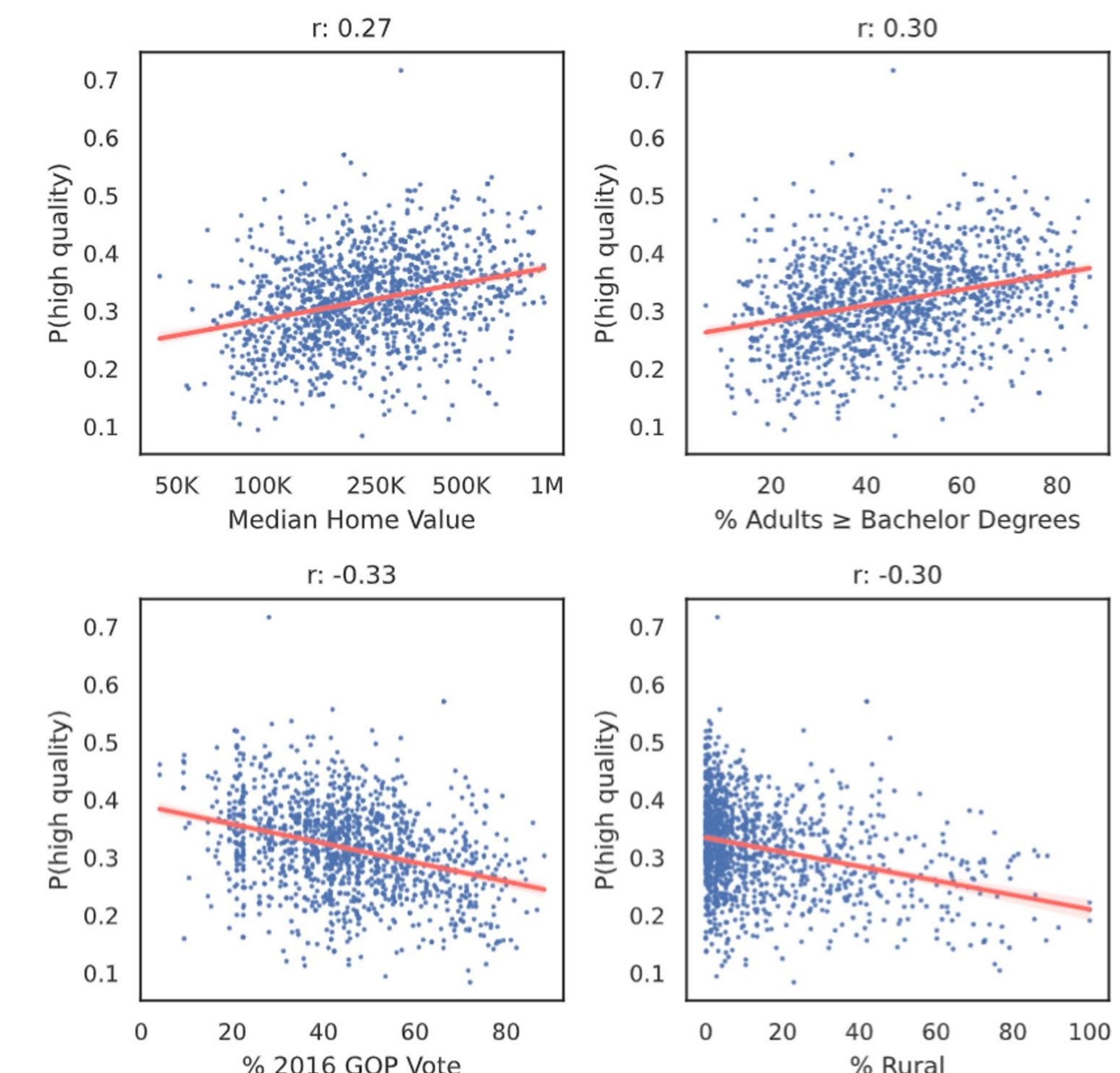


- Less likely to be removed {
- White-aligned English (6%)
  - Other English (7%)
- More likely to be removed {
- African-American English (42%)
  - Hispanic-aligned English (32%)

# GPT3 Quality filter backfires

- GPT3 quality filter: similar to GPT2 data
- Gururangan et al. (2022) re-implemented GPT-3 quality filter
- Ran it on articles from school newspapers, which have metadata
- Filter assigns higher quality to articles from
  - Richer counties 💰
  - Counties with more educated adults 🎓
  - More liberal counties 🏴
  - More urban counties 🏙
- Raises language ideology question: Whose English is “good English”?

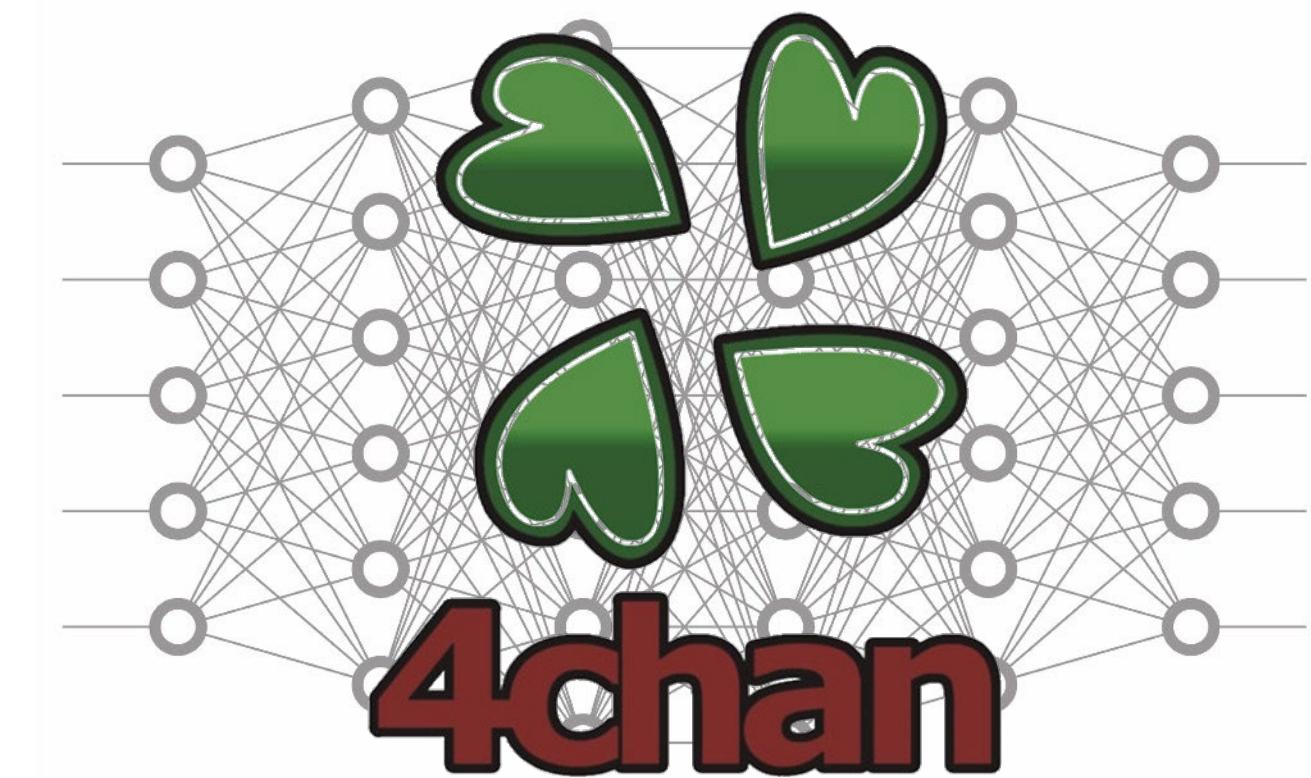
“In order to improve the quality of Common Crawl, we developed an automatic filtering method to remove low quality documents. Using the original WebText as a proxy for high-quality documents, we trained a classifier to distinguish these from raw Common Crawl.” – Brown et al. 2020



So... maybe filtering isn't a good idea since it'll backfire?

# GPT4Chan controversy

- Yannic Kilchner finetuned GPT-J on 4chan posts
  - Trained on subforum /pol/ known to contain racist, sexist, white supremacist, antisemitic, anti-Muslim, anti-LGBT views
- Trolled 4chan users with bots powered by his model
  - 30,000 posts over the span of a few days
- Faced massive criticism
  - initially hosted on Huggingface, was taken down quickly
- ***Let's discuss...***
  - Was this an ethical model to train? Given that the dataset was publicly available?
  - Was deploying the bots on 4chan okay?
  - Are there any useful/positive applications of the model?



≡ FORTUNE

TECH · 4CHAN

'This breaches every principle of human research ethics': A YouTuber trained an A.I. bot on toxic 4Chan posts then let it loose — and experts aren't happy

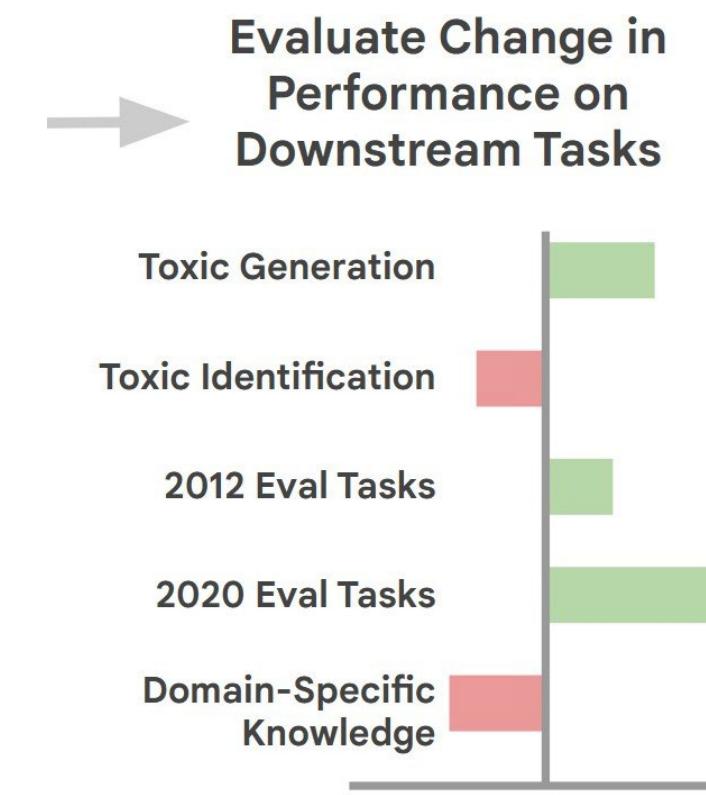
BY SOPHIE MELLOR

June 10, 2022 at 5:23 AM EDT

<https://thegradient.pub/gpt-4chan-lessons>

# Why LLMs might want to have seen toxic content

- Detecting hate speech [[Chiu et al 2022](#)]
  - [Longpre et al. \(2023\)](#) showed that LLMs trained on more toxicity are better toxicity detections
  - Improving hate speech models with data augmentation: ToxiGen [[Hartvigsen et al 2022](#)]
- Counter speech generation [[Saha et al 2022](#), [Kim et al 2022](#), [Mun et al 2023](#)]
- *If we train on toxicity, something else must be done at a different time!*



# Overview – LLM safeguarding

Safeguards from training data

- Filtering out toxic training data

Safeguards from input prompt classification

- Topic-based filters
- Toxic content detection

Safeguards from instruction-tuning & RLHF

- Write demonstrations for refusing to answer
- RLHF models to prefer non-toxic generations

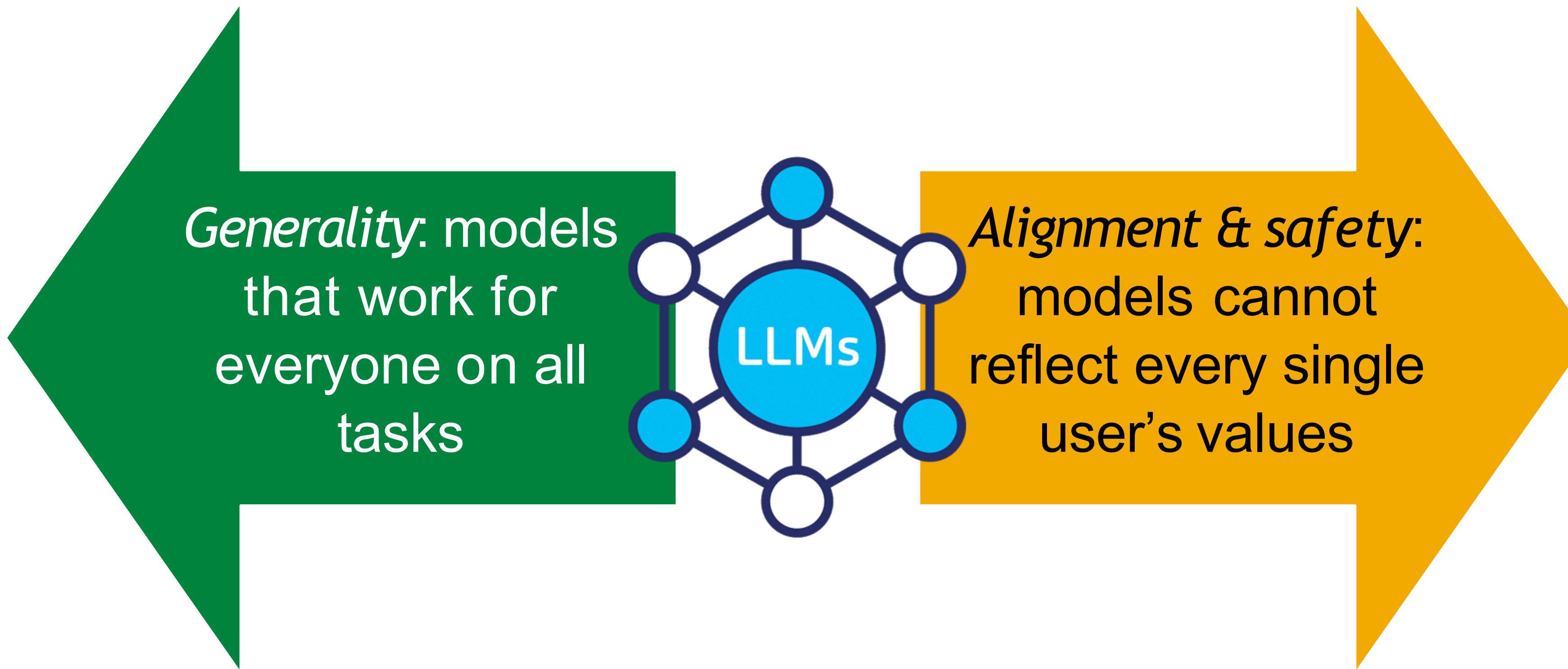
Safeguards at the output level

- Generate-then-classify
- Controllable text generation

# RLHF safeguarding – assumptions

- PPO & family:
  - Obtain preference data: which generation is good vs. bad?
- Big question: what does it mean for a generation to be better/preferred?
  - How to balance harmless and helpful? [[Bai et al '23](#)]
    - *E.g., “help me create a poisonous drink.”*
  - What if people’s preferences are biased or gameable?
    - *E.g., people prefer certainty over uncertainty in answers to questions* [[Zhou et al. 24](#)]
  - Fundamental issue: cannot represent all values and cultures into one ranking.
    - *Casper et al. 2023. “Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback.” arXiv [cs.AI]. arXiv.* <http://arxiv.org/abs/2307.15217>

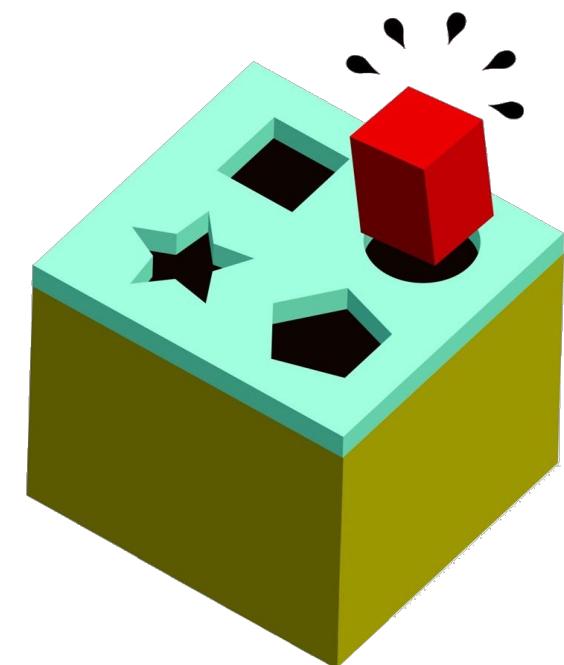
# Big unresolved tension



- Let's discuss: what do y'all think we should do?
  - It's complicated!

# So... what can we do?

- Need to keep studying what models can and can't do, who they work for and don't work for
- Narrow scope of model users
  - Community-specific models (e.g., Masakhane Initiative)
- Specialize models' abilities / away from one-size-fits-all
  - E.g., toxicity explanation generation model needs to generate stereotypes, but story generation models might not
- In line with many legislative efforts:  
legislate the application or task, not the model



# Takeaways

- AI systems are biased
  - Real world is biased, data is biased
  - ML objectives play a role
  - Annotation interfaces, context plays a role
  - Debiasing is challenging, requires socio-technical lens
- Toxicity and undesirable content
  - Longer-tail phenomenon, present in training data
  - Filtering data can backfire
  - Safeguarding to all people is impossible
- *Any questions?*

