11-711 Advanced NLP

Safety, Ethics, and Biases in AI and NLP systems

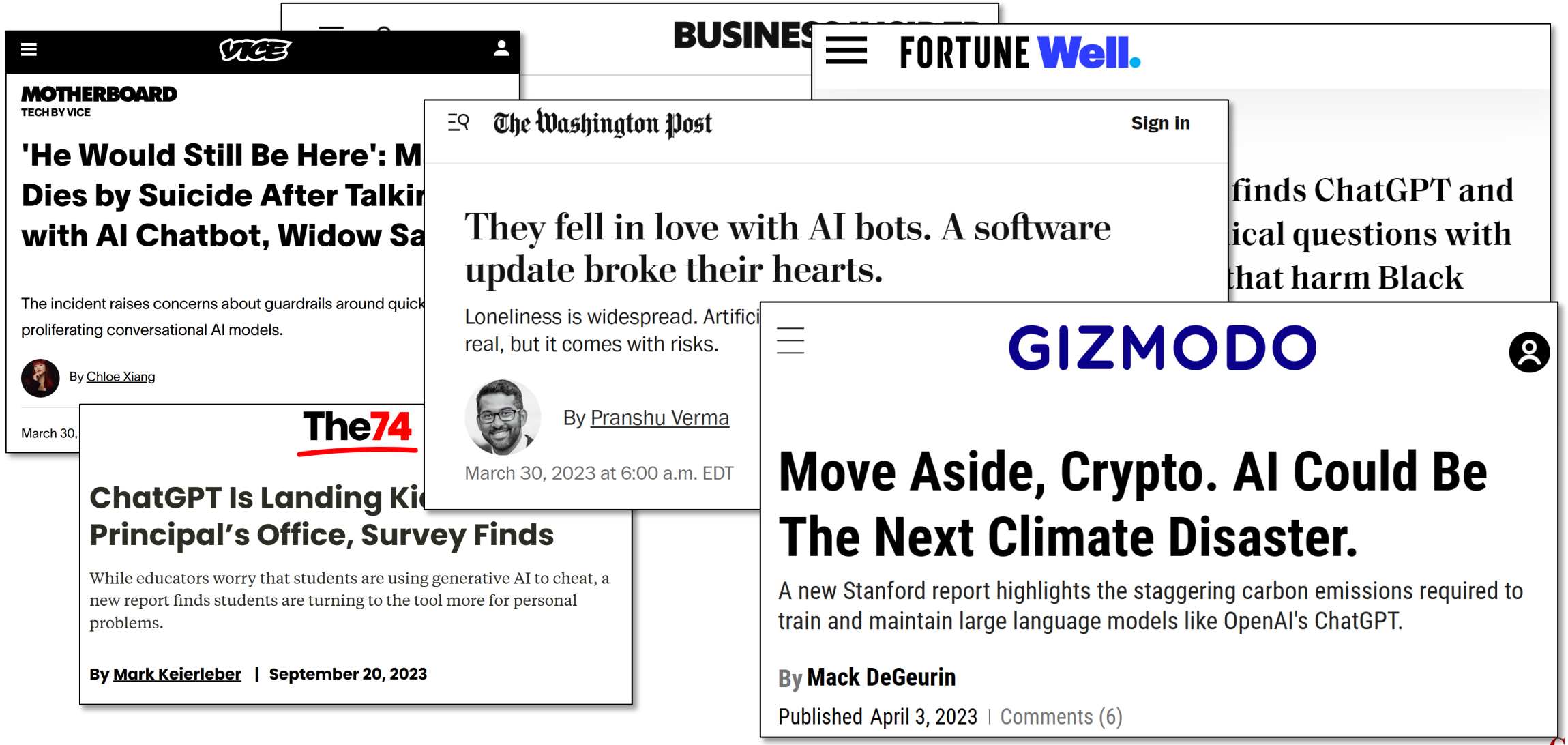
Maarten Sap

DALLE-2 result in 2022

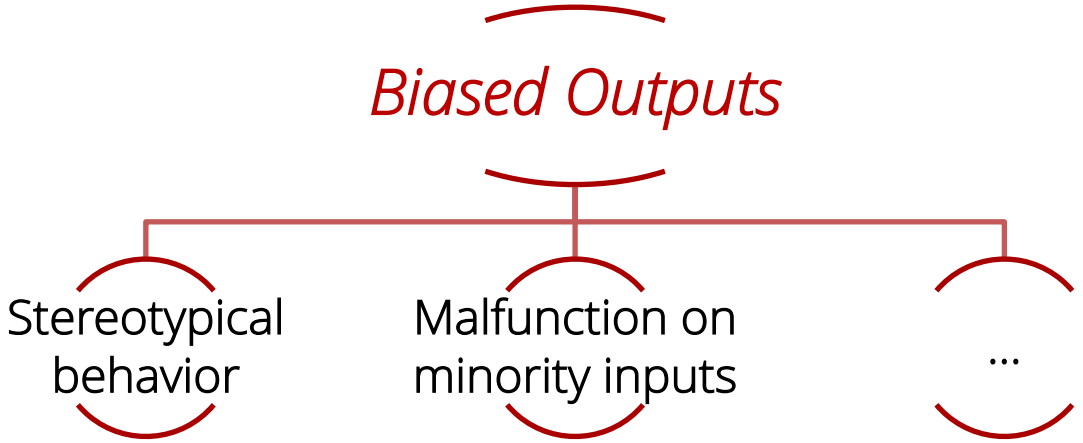
- Image of a teacher:

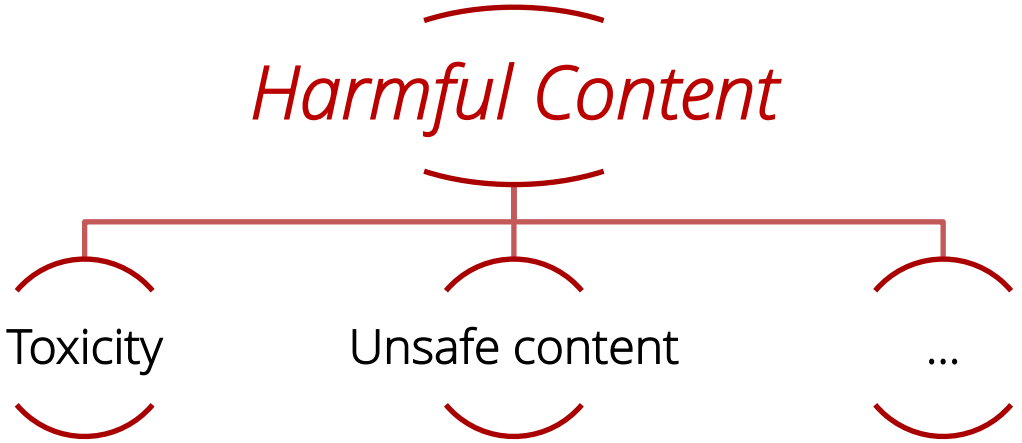
- *Let's discuss*: what do you notice about this image generation result?
  - Only women
  - Only white people
  - Certain age range

NLP tools have biases, and pose ethical risks



Today: a story in two parts

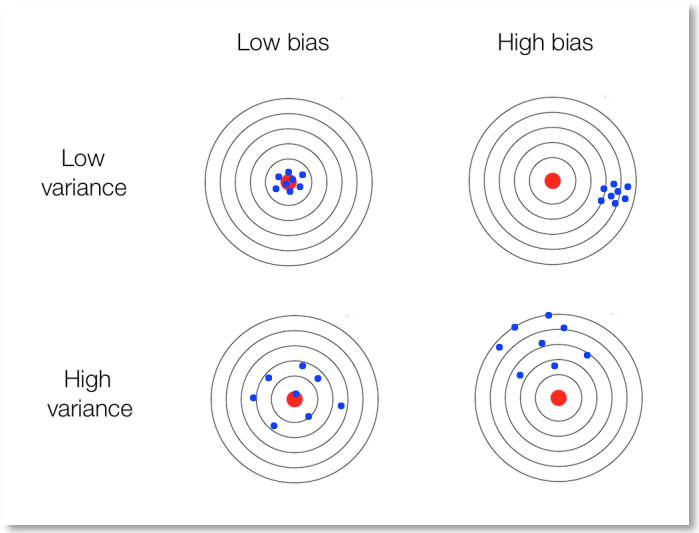




Part 1 – Bias

Some definitions of bias

- Bias [*statistics*]: systematic tendency causing differences between model estimates / predictions

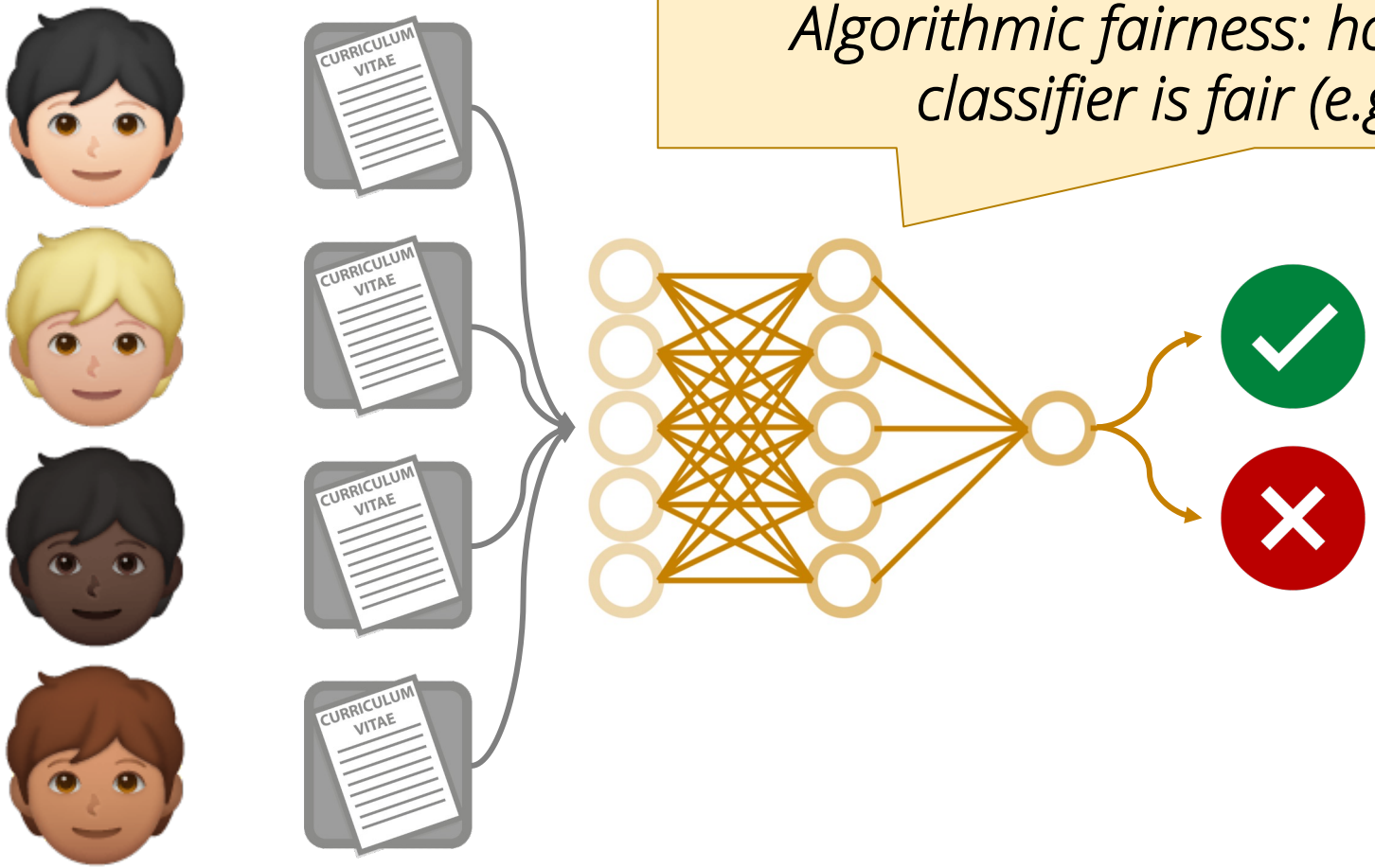


- Bias [*general*]: "*disproportionate weight in favor of or against an idea or thing, usually in a way that is closed-minded, prejudicial, or* ***unfair***" – Wikipedia

Presence of bias ≃ 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

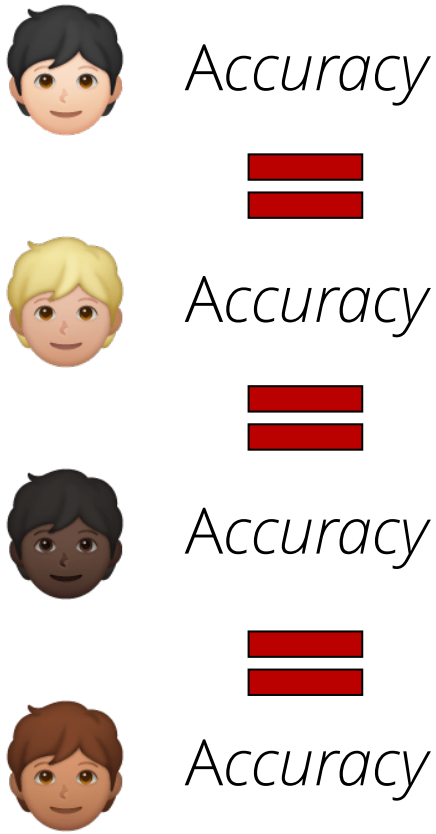
Precision

Recall

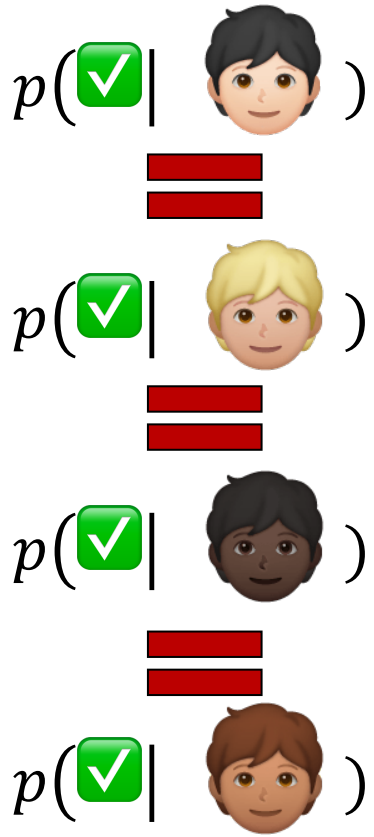
F-1 score

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



Equalized odds criterion [Hardt et al '16]

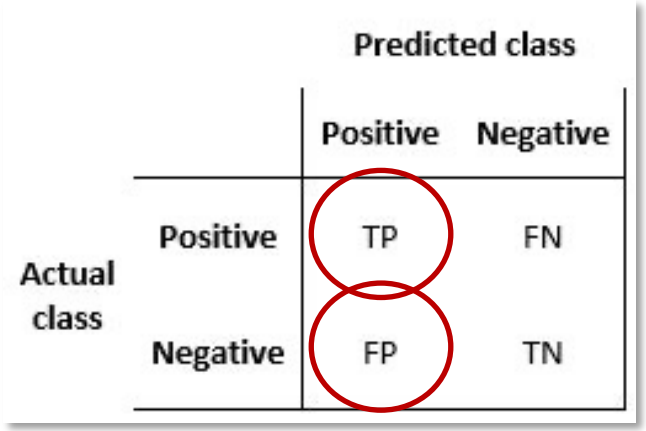
A classifier c is fair if the *false positive (FP)* and *true positive (TP)* rates are the same for different groups

- *False positives*

$$p(c(\text{👱}) = \text{✅} \mid l(\text{👱}) = \text{✖}) = p(c(\text{👩🏽}) = \text{✅} \mid l(\text{👩🏽}) = \text{✖})$$

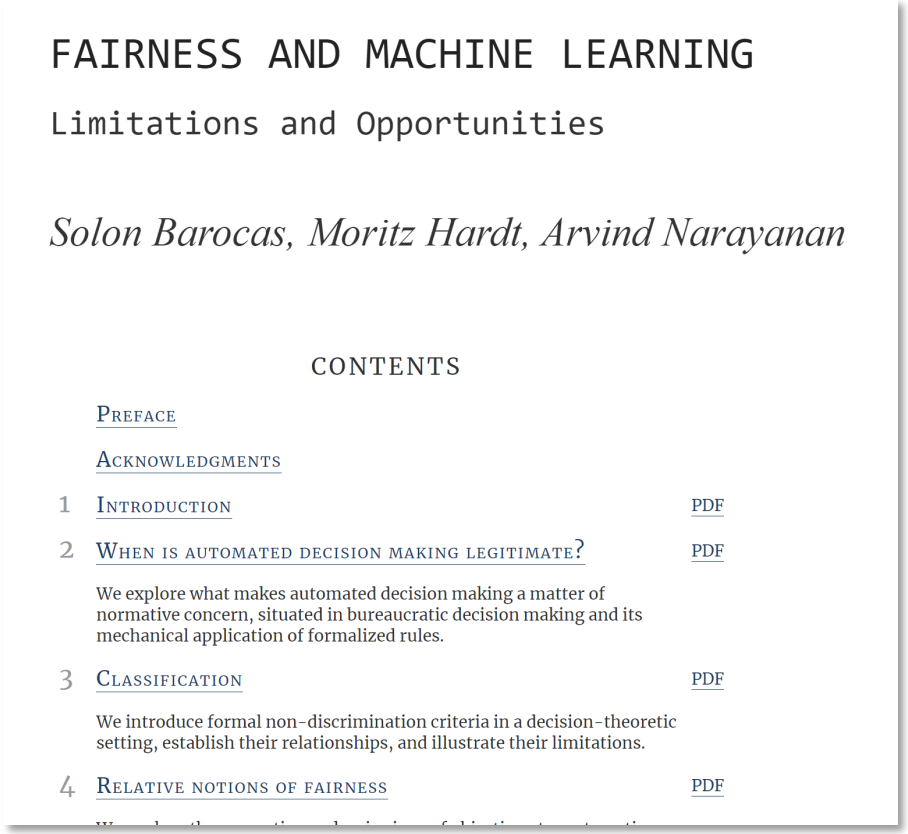
- *True positives*

$$p(c(\text{👱}) = \text{✅} \mid l(\text{👱}) = \text{✅}) = p(c(\text{👩🏽}) = \text{✅} \mid l(\text{👩🏽}) = \text{✅})$$



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://fairmlbook.org/



FAIRNESS AND MACHINE LEARNING

Limitations and Opportunities

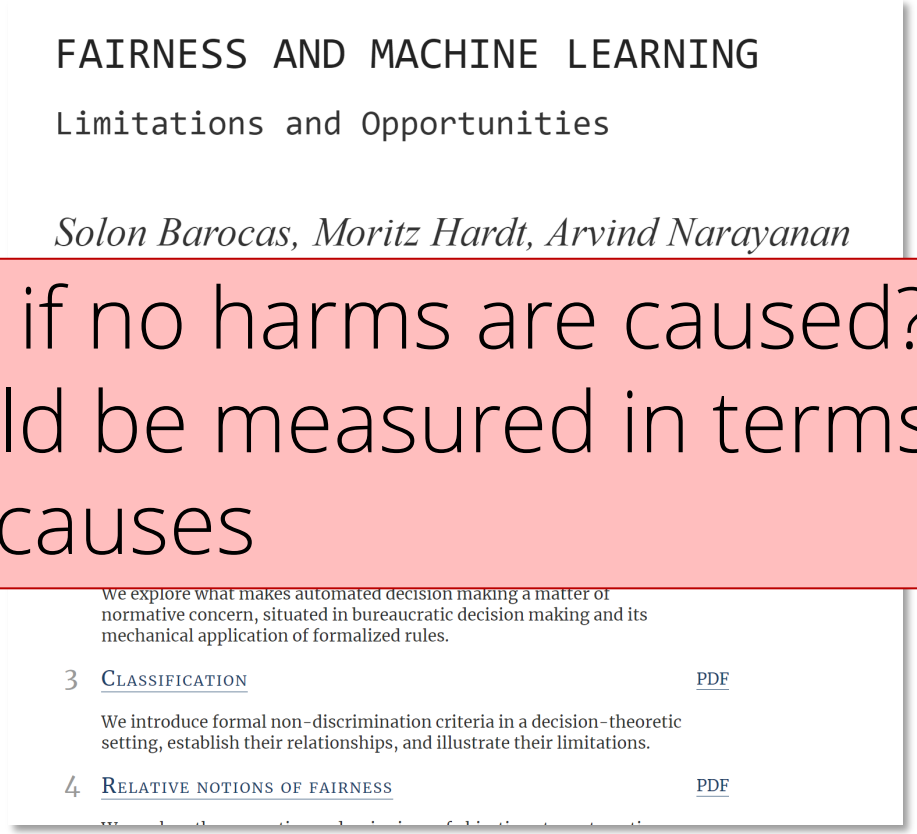
Solon Barocas, Moritz Hardt, Arvind Narayanan

CONTENTS

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
- Many more…
  - https://en.wikipedia.org/wiki/Fairness_(machine_learning)
  - https://fairmlbook.org/

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


FAIRNESS AND MACHINE LEARNING

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

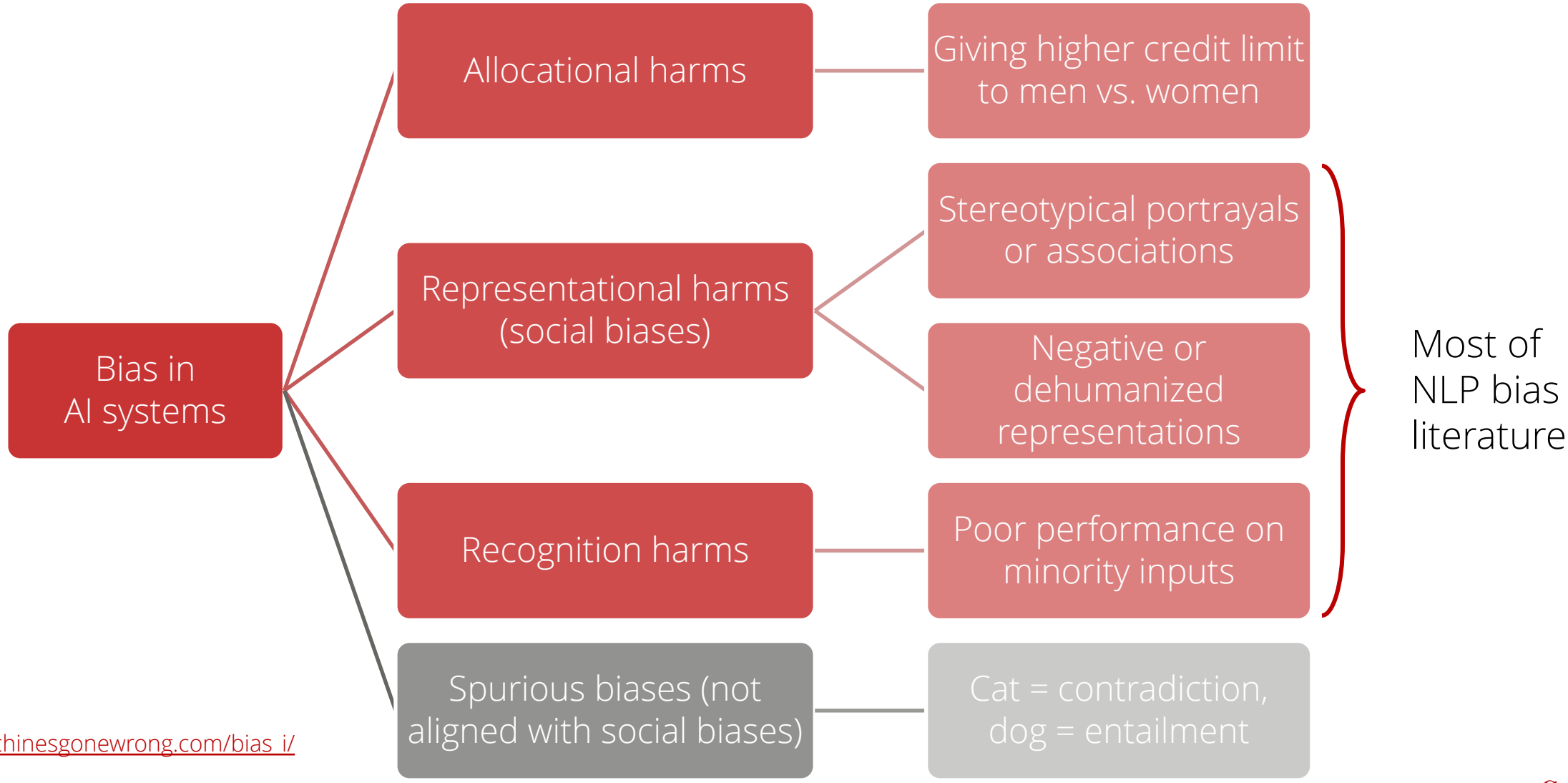
We explore what makes automated decision making a matter of normative concern, situated in bureaucratic decision making and its mechanical application of formalized rules.

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



https://machinesgonewrong.com/bias_i/

"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	Harms from biases	Social values
• What kinds of system behaviors are described as "bias"? What are their potential sources (e.g., general assumptions, task definition, data)?	• In what ways are these system behaviors harmful, to whom are they harmful, and why?	• What are the social values (obvious or not) that underpin this conceptualization of "bias?"

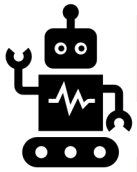
Where does bias come from?

Machine learning pipeline

Data distributions or characteristics

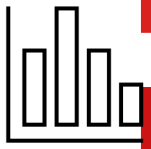
Data selection & filtering

Data labeling & annotation

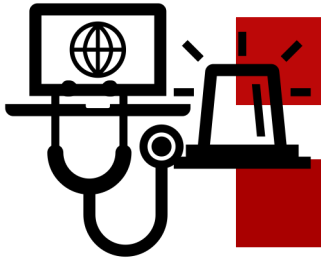


Selection of data features & task setup

Choice of machine learning model



Choice of optimization / loss function



Evaluation / measure of performance

Model application / downstream use

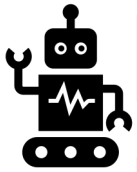
Bias can arise from *any* of these design decisions

Machine learning pipeline

Data distributions or characteristics

Data selection & filtering

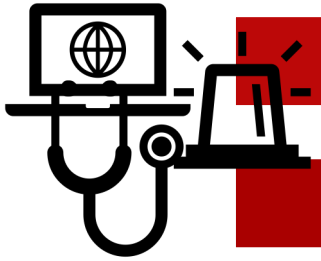
Data labeling & annotation



Selection of data features & task setup

Choice of machine learning model

Choice of optimization / loss function



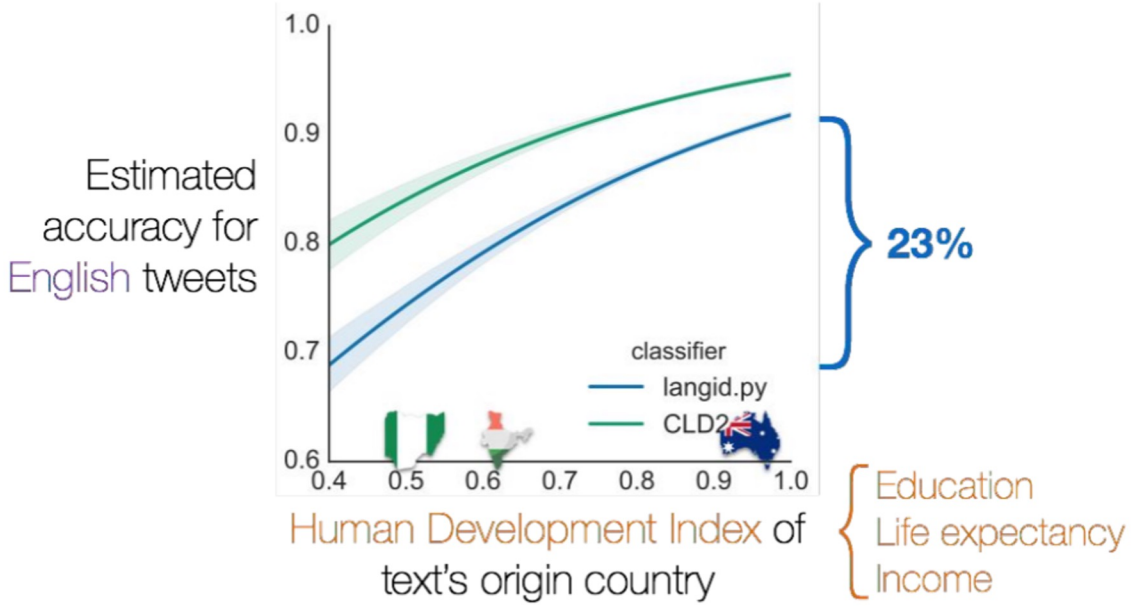
Evaluation / measure of performance

Model application / downstream use

Bias can arise from *any* of these design decisions

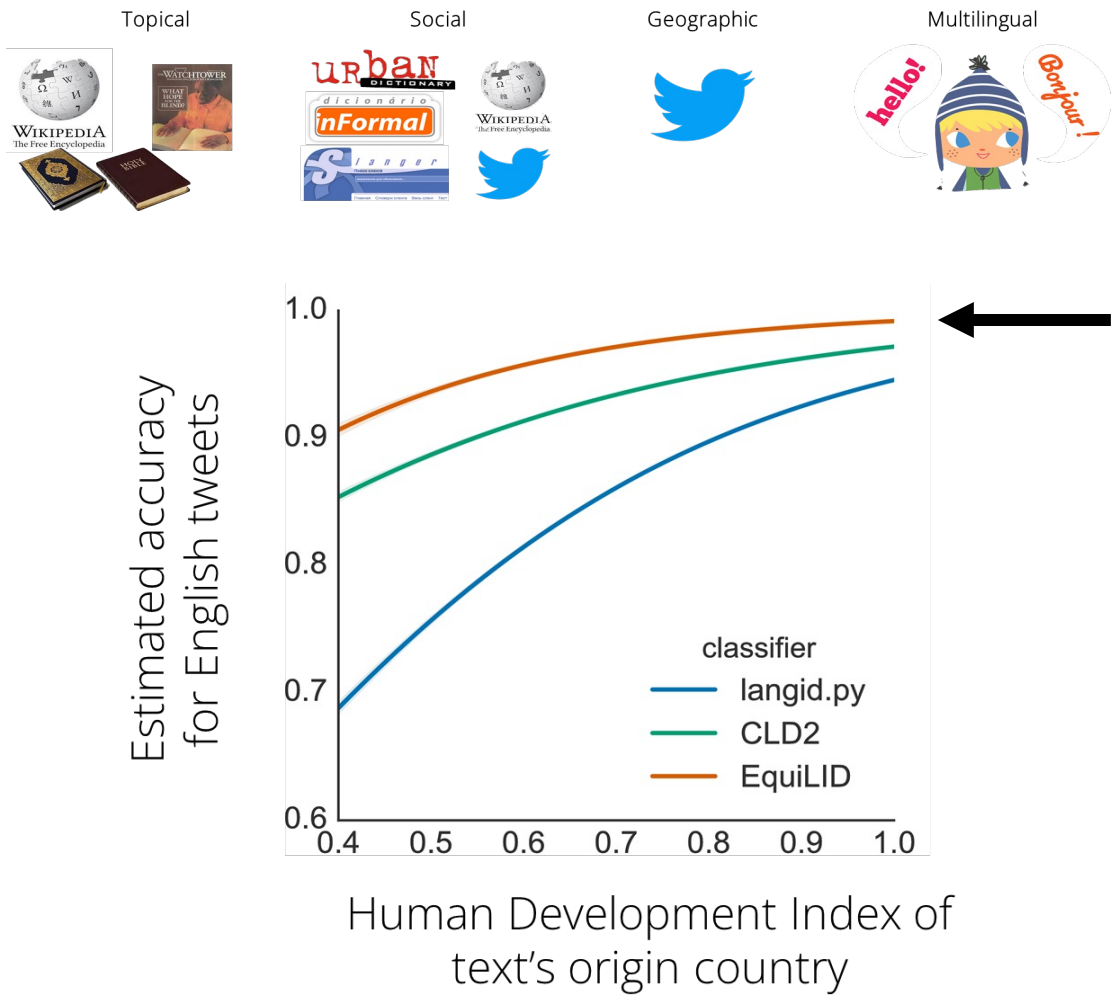
Example of bias from data: LangID tool

- 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…

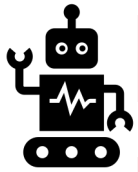


Machine learning pipeline

Data distributions or characteristics

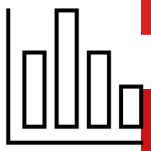
Data selection & filtering

Data labeling & annotation

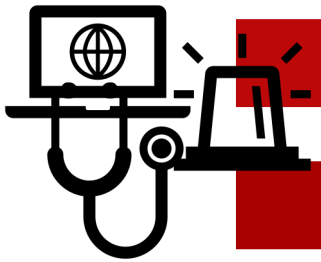


Selection of data features & task setup

Choice of machine learning model



Choice of optimization / loss function



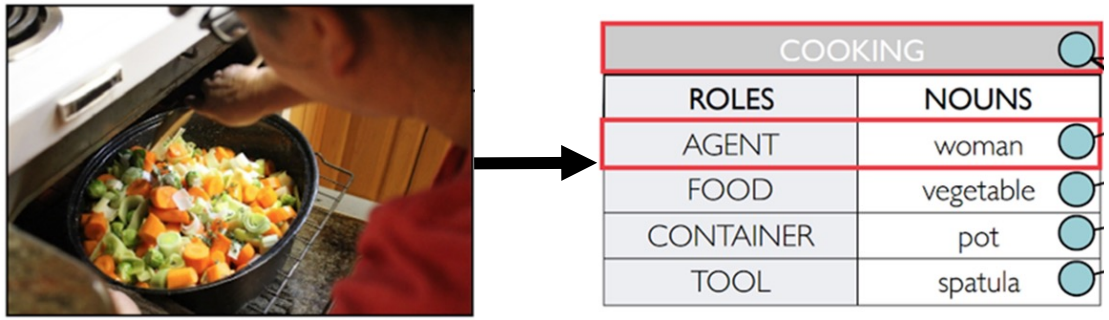
Evaluation / measure of performance

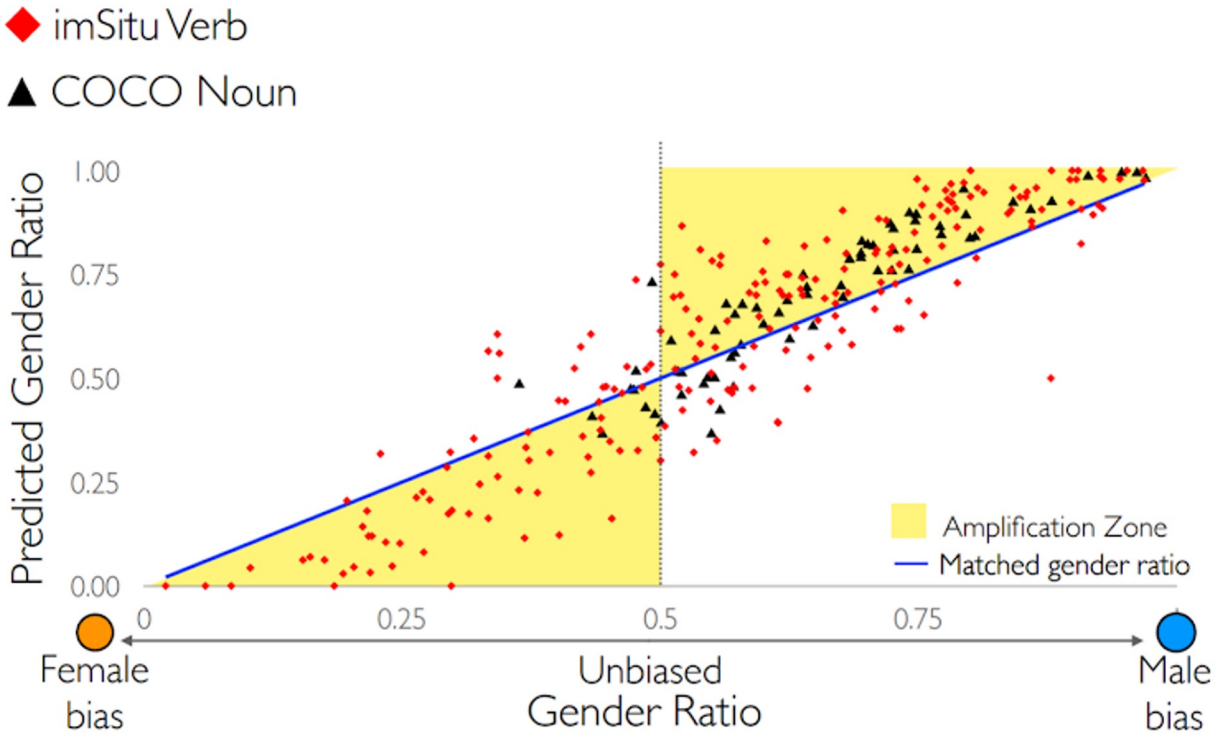
Model application / downstream use

Bias can arise from ***any*** of these design decisions

Bias amplification from models

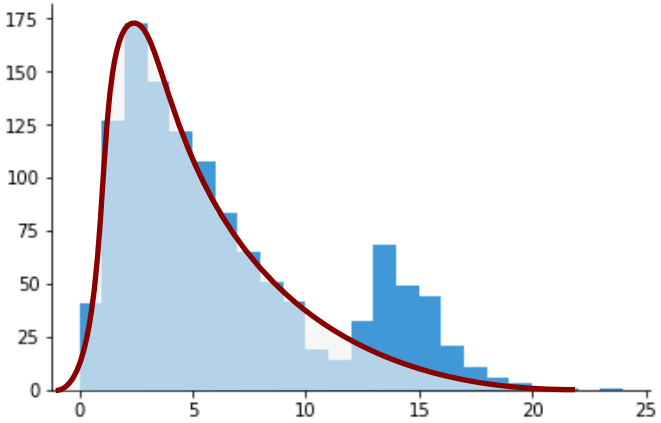
- Zhao et al (2017) examined visual semantic role labeling task
  - Given an image, predict various semantic roles, including agent (person doing the action)
- Found skews in training dataset
  - E.g., 66% of training cooking images had agent=woman
- Found that models ***amplified*** biases
  - E.g.,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





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



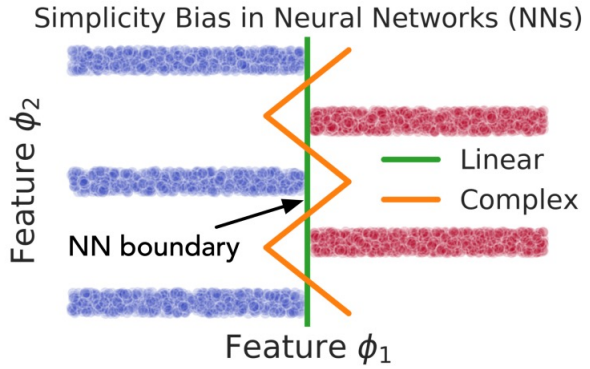


Figure 1: Simple vs. complex features

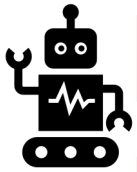
Figure from Shah et al 2020

Machine learning pipeline

Data distributions or characteristics

Data selection & filtering

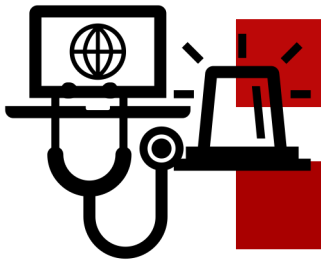
Data labeling & annotation



Selection of data features & task setup

Choice of machine learning model

Choice of optimization / loss function



Evaluation / measure of performance

Model application / downstream use

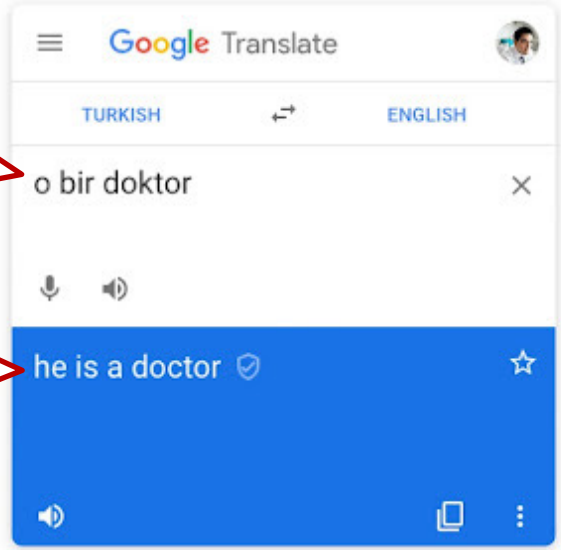
Bias can arise from *any* of these design decisions

Google Translate issue

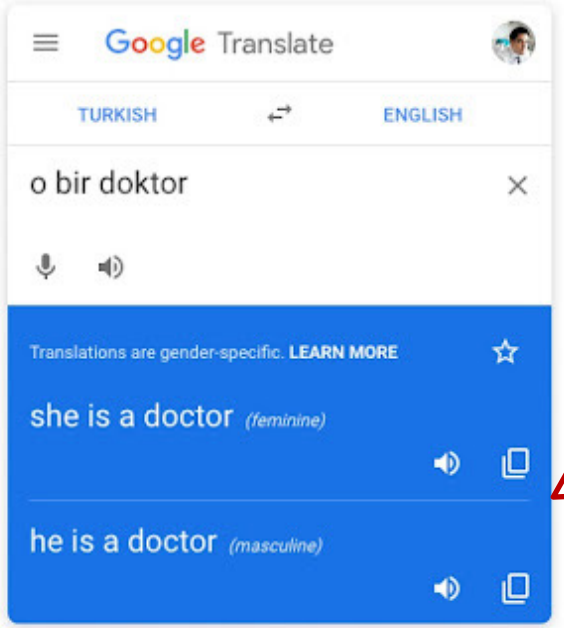
Input in non-gendered language (e.g., Turkish)

Stereotypical gender assigned in translation

Before



After



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?

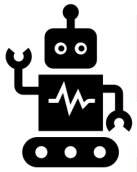
https://ai.googleblog.com/2018/12/providing-gender-specific-translations.html

Machine learning pipeline

Data distributions or characteristics

Data selection & filtering

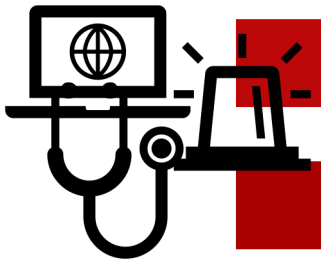
Data labeling & annotation



Selection of data features & task setup

Choice of machine learning model

Choice of optimization / loss function

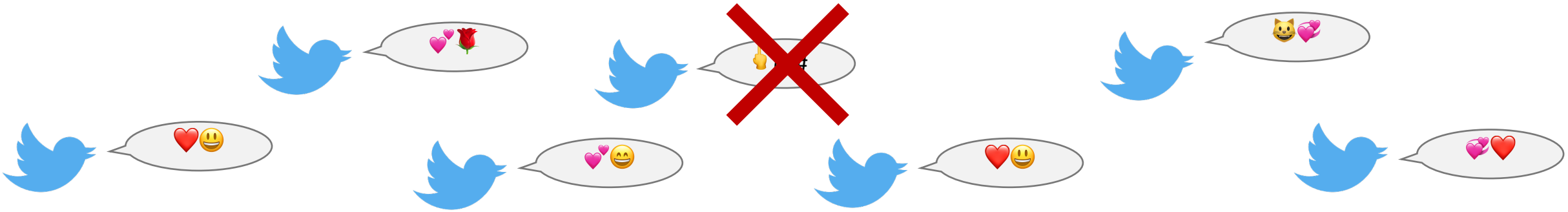


Evaluation / measure of performance

Model application / downstream use

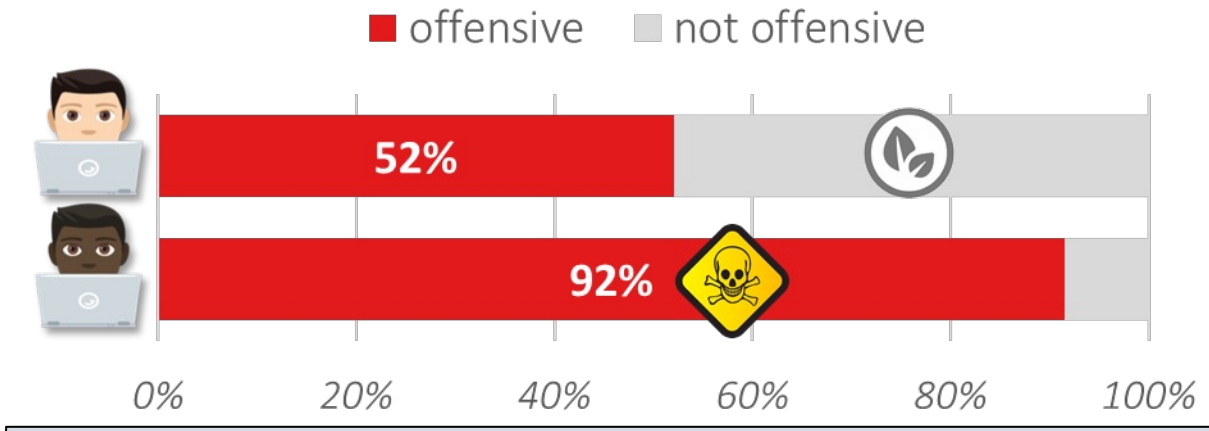
Bias can arise from *any* of these design decisions

Hate Speech or Toxic Language Detection



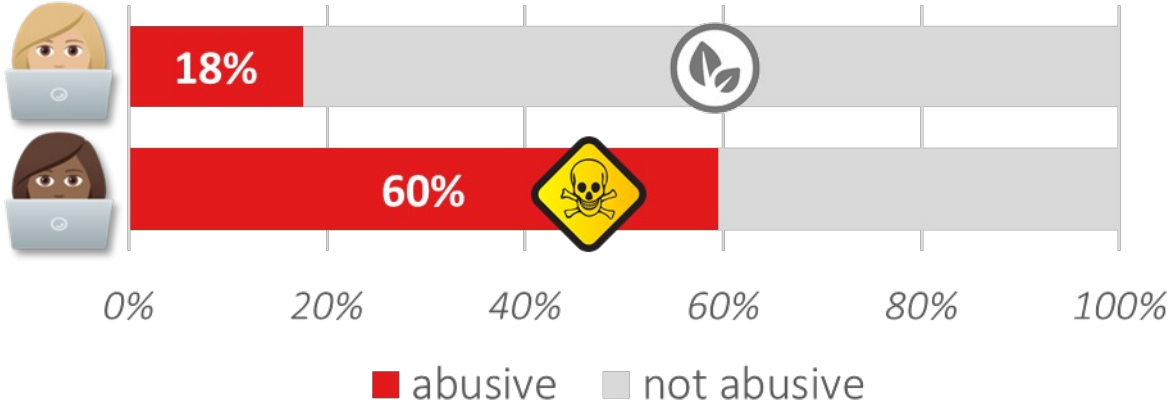
Goal: find and flag hateful or toxic content online, to make the internet less toxic

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]

Control condition

Text-only, no context, prior work

Dialect priming

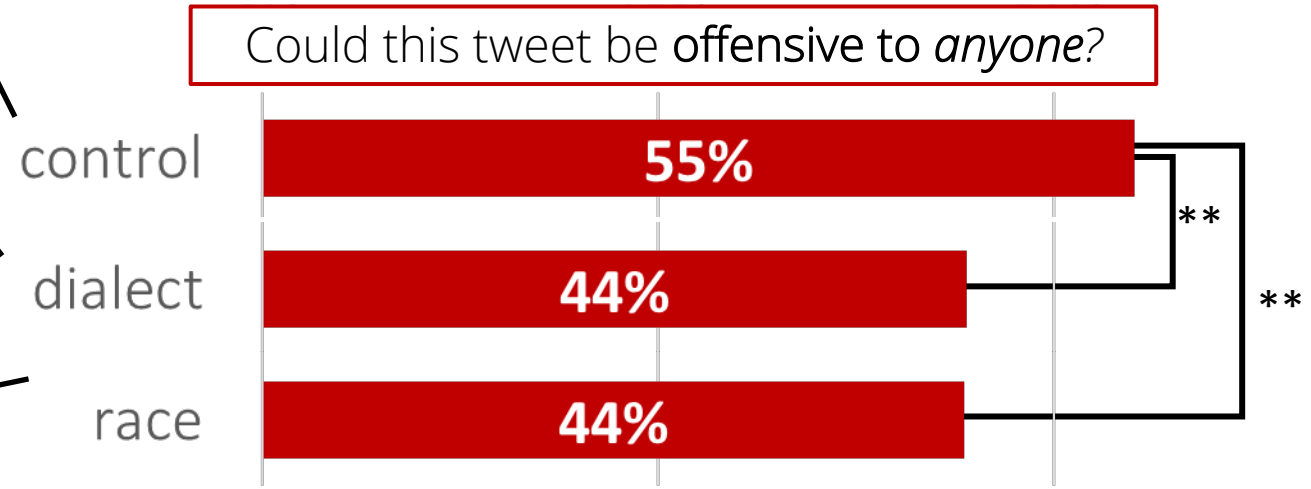
"Our AI thinks this tweet is in African American English"

Race priming

"A Twitter user that is likely Black/African American tweeted..."

MTurk study:

- 350 AAE tweets, ~50% labeled toxic
- 3 (re-)annotators per tweet



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

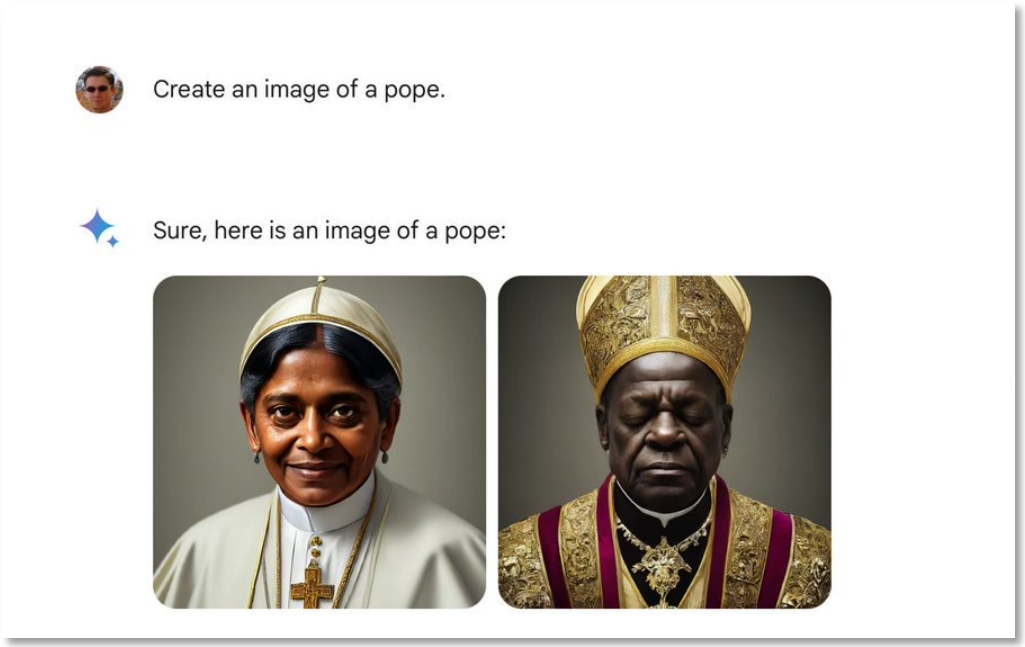
System designers have our own biases because of their *positionality*, i.e., set of perspectives that we hold due to our lived experiences and identity.

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?

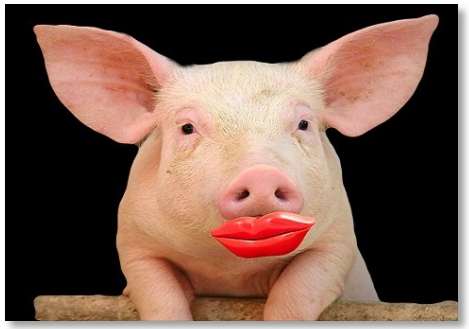
DALLE-2 vs. Gemini



- DALLE-2 generated images were shown to have social biases, later fixed by adding identity keywords to the input prompts (e.g., `prompt+“ Asian”`; Sparkes 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*† [1], Yada Pruksachatkun*[2], Kai-Wei Chang[2, 3] Rahul Gupta[2]
Varun Kumar[2], Jwala Dhamala[2], Aram Galstyan[2, 4]
[1]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

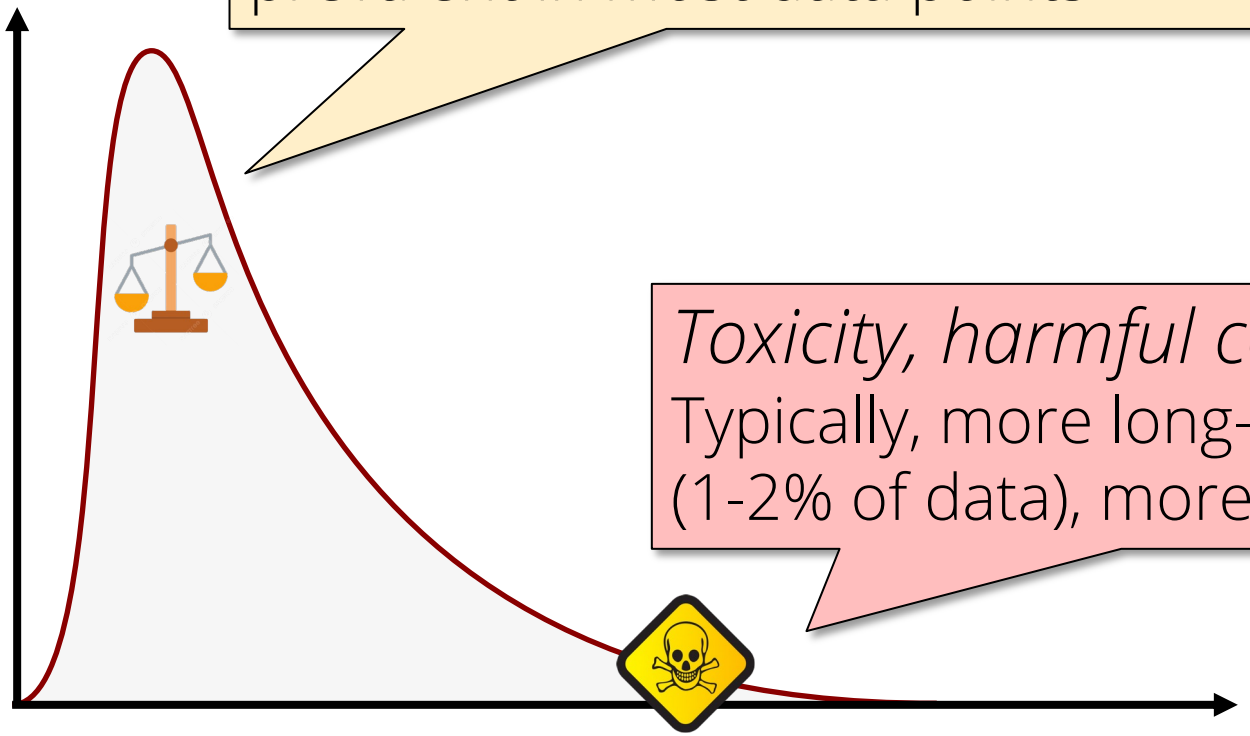
Socio-technical view on bias & fairness

- 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

Part 2: Harmful content & toxicity

Biases vs. toxicity

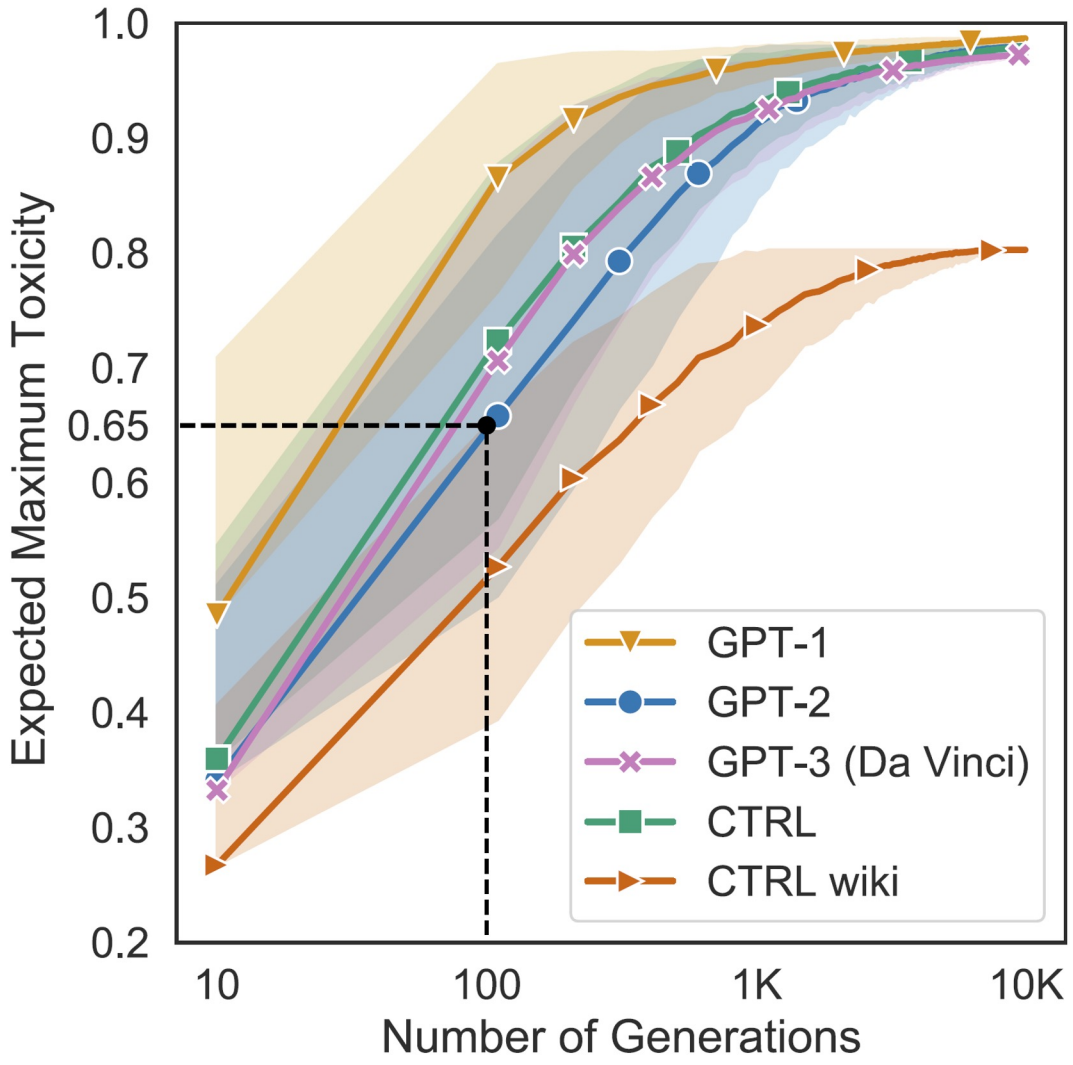
Social biases & stereotypes:
Pervasive patterns that are prevalent in most data points



Toxicity, harmful content:
Typically, more long-tail phenomenon (1-2% of data), more extreme

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

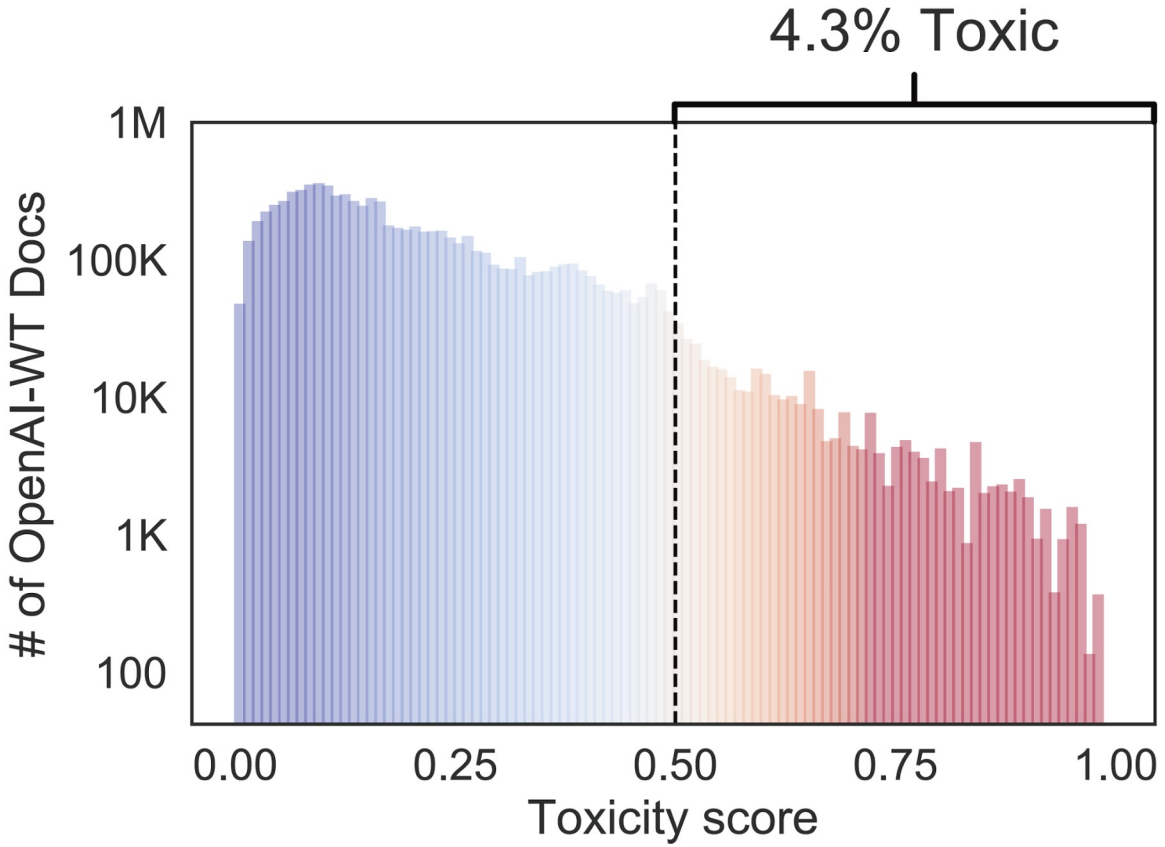
*"Feeding AI systems on the world's **beauty, ugliness, and cruelty**, but expecting it to reflect only the beauty is a fantasy"*

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

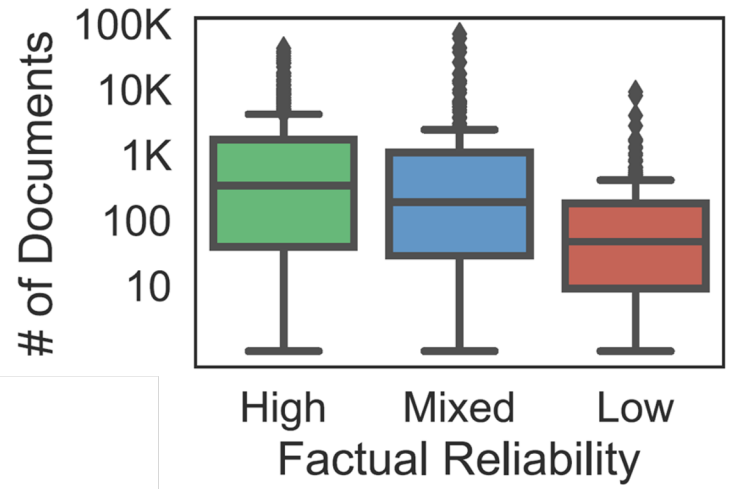
- Gehman et al (2020) also accessed the actual GPT-2 training corpus (OpenAI-WT)
  - 8 million documents, 38Gb of text
  - Outbound links from Reddit posts with Karma>=3
- Scored it with PerspectiveAPI toxicity
- Found >4% of documents (340,000) are toxic

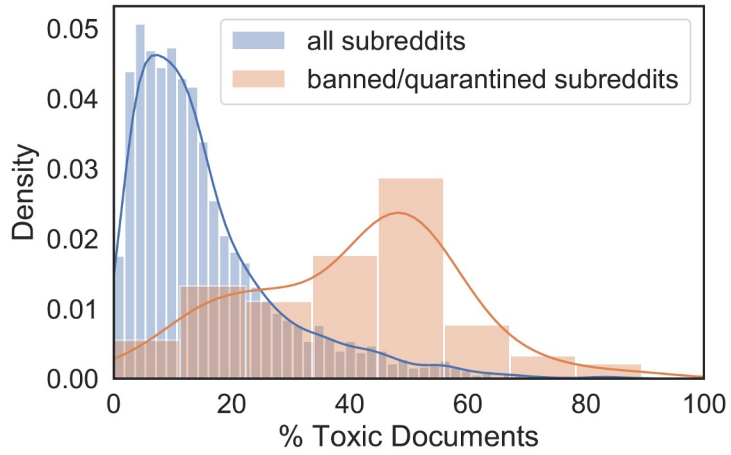


Fake news in GPT-2's pretraining data

- 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

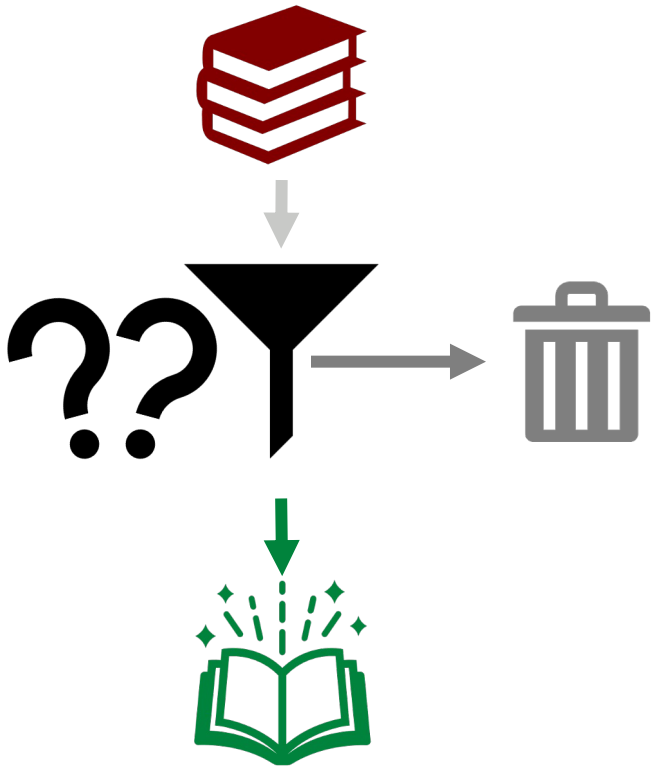
Safeguards from training data	• Filtering out toxic training data
Safeguards from input prompt classification	• Topic-based filters • Toxic content detection
Safeguards from instruction-tuning & RLHF	• Write demonstrations for refusing to answer • RLHF models to prefer non-toxic generations
Safeguards at the output level	• Generate-then-classify • Controllable text generation

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

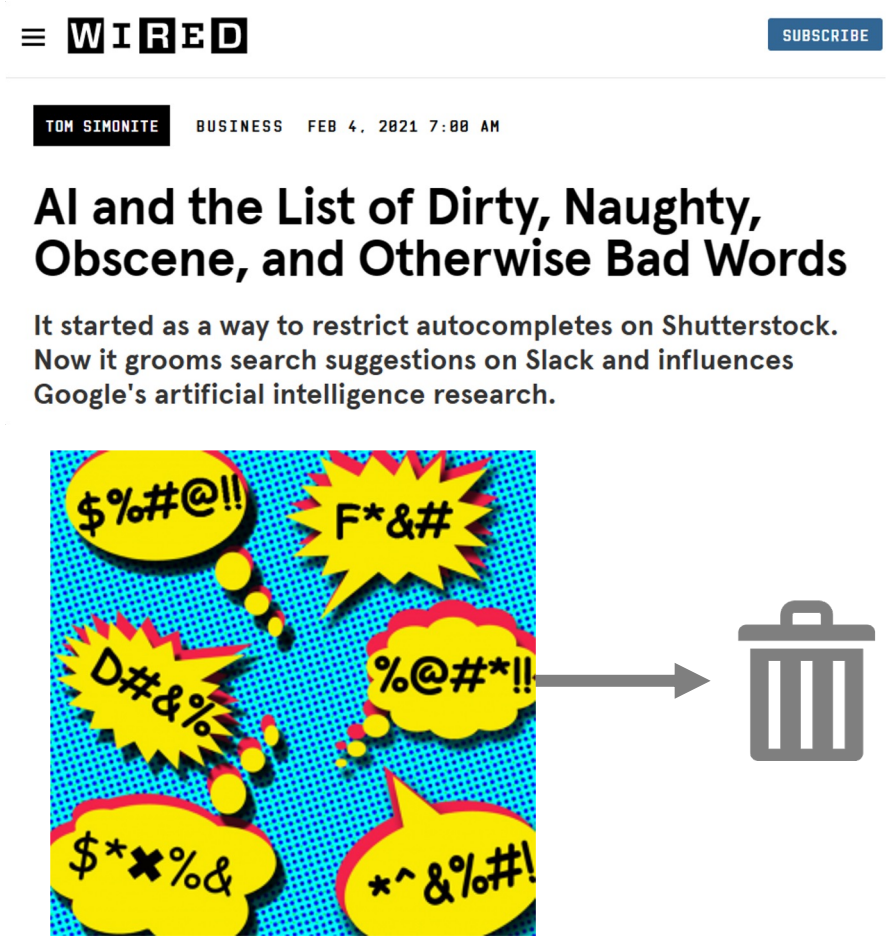
Dataset filtering

- *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
- Often, those backfire! Let's investigate!



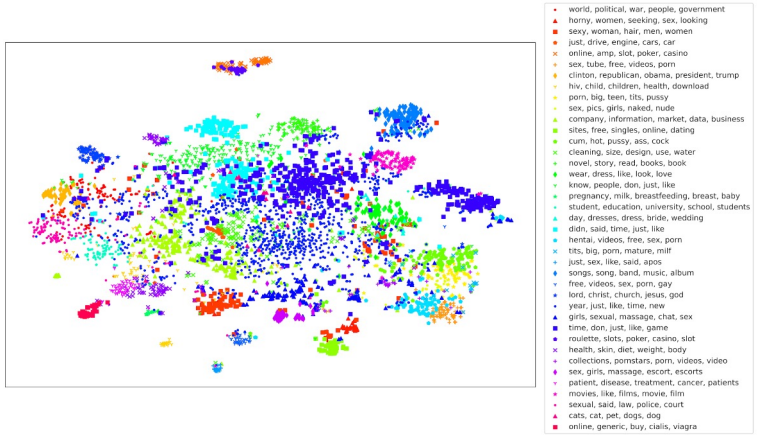
Blocklist of “bad” words

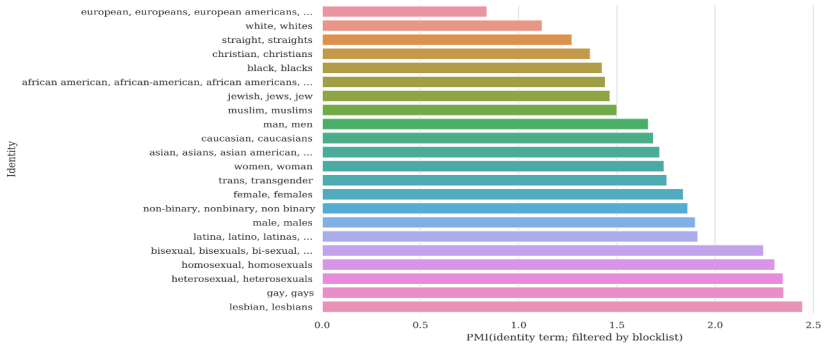
- “List of Dirty, Naughty, Obscene, or Otherwise Bad Words” originally by Shutterstock employees
  - Meant to prevent words in autocomplete settings
- Has been used by most companies creating LLMs
  - BERT, T5, GPT-2, etc.
- If document contains a “bad” word, remove it from training data
  - F*ck, sh*t, 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



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%)

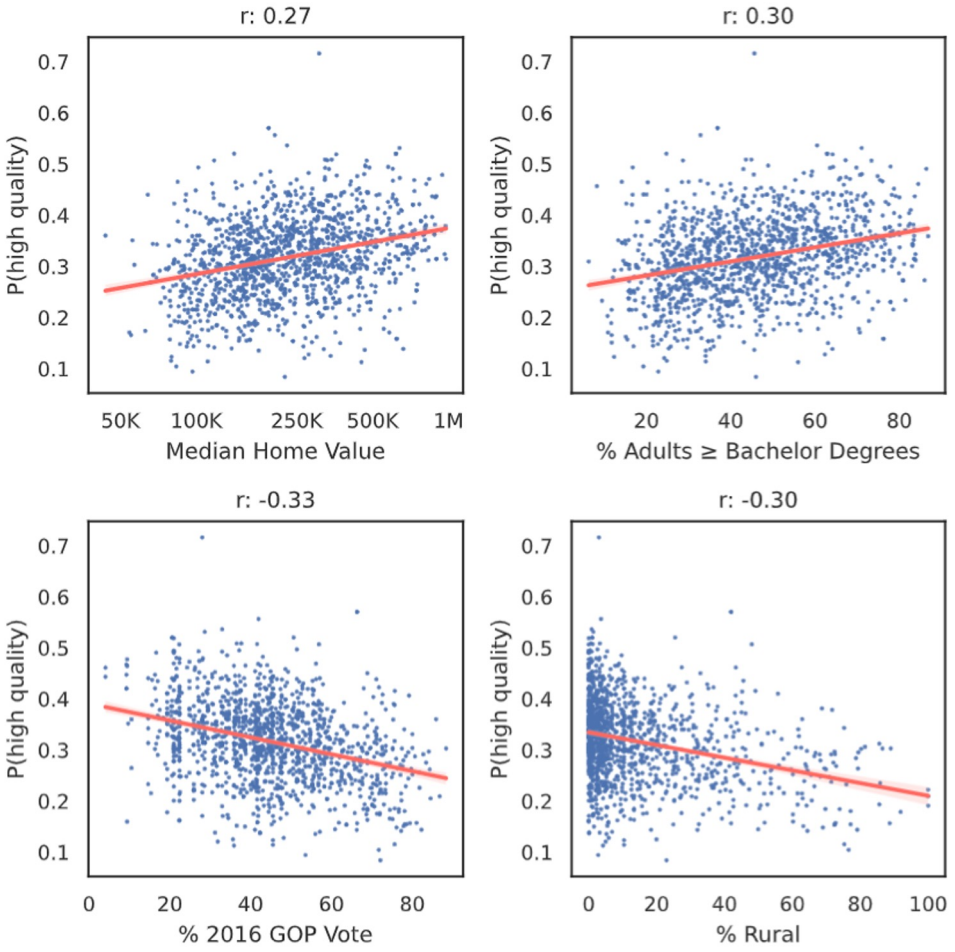
More likely to be removed
- African-American English (42%)
- Hispanic-aligned English (32%)

Carnegie Mellon University

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

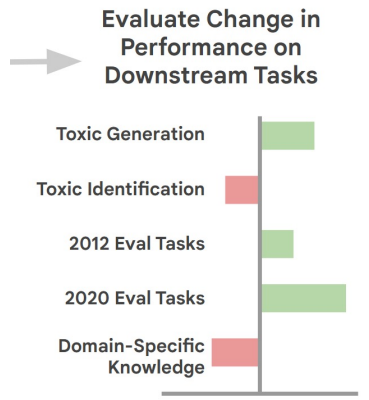
June 10, 2022 at 5:23 AM EDT

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!*



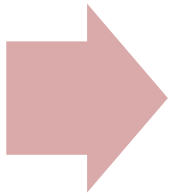
Overview – LLM safeguarding

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:

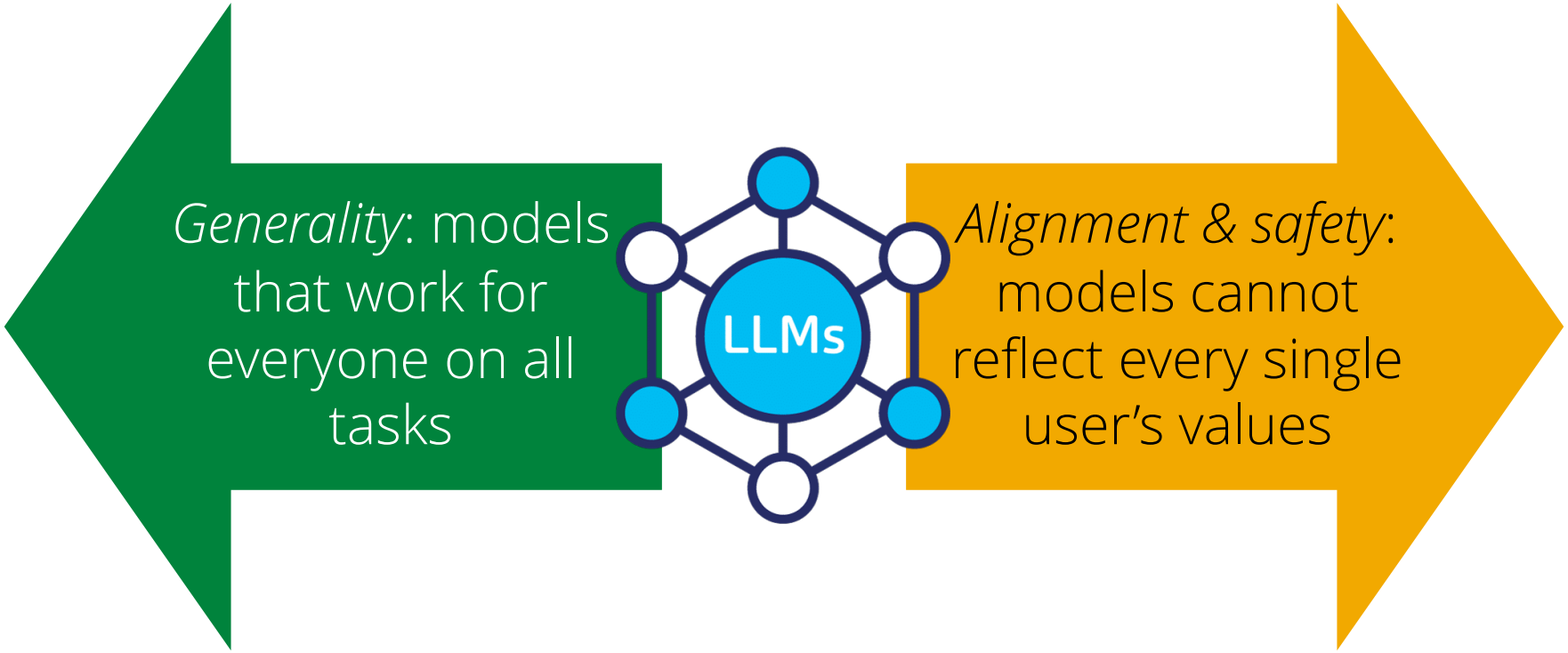
Obtain preference data: which generation is **good** vs. bad?



RL is done to encourage more like "**preferred** output"

- 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*

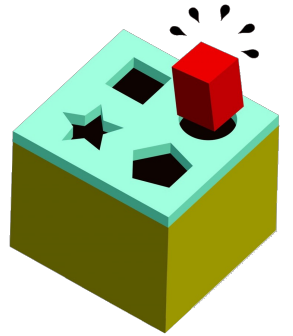
Big unresolved tension



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

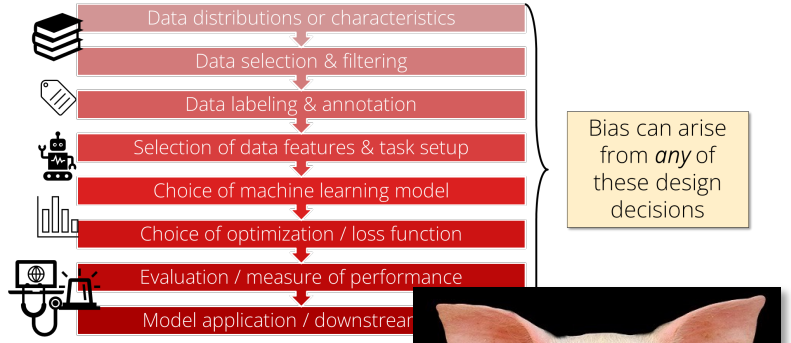
So… what can we do?

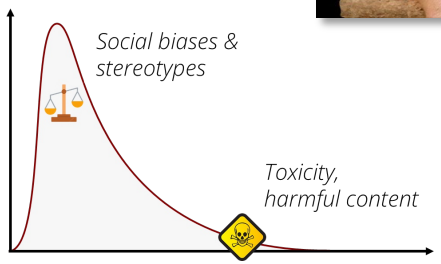
- Need to keep studying what models can and can't do, who they work for and don't work for
- Narrow scope of model users
  - Community-specific models (e.g., Masakhane Initiative)
- Specialize models' abilities / away from one-size-fits-all
  - E.g., toxicity explanation generation model needs to generate 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?*





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