Foundation of Inequity in AI

Ethical AI

foundation

Inequitable Web Data as a Foundation?

"We call these models foundation models *to underscore their critically central yet incomplete character … current examples include BERT [Devlin et al. 2019], GPT-3 [Brown et al. 2020], and CLIP [Radford et al. 2021]"*

"A major motivation for natural language supervision is the large quantities of data of this form available publicly on the internet" [Radford et al, 2021]

Foundation Model = Model trained with unconsented natural language data crawled from the web with limited respect of copyright, license, PII?

BERT: Large model of language.

- Trained with unconsented natural language data crawled from the web with limited respect of copyright, license, PII

GPT-3: Large model of language.

- Trained with unconsented natural language data crawled from the web with limited respect of copyright, license, PII

CLIP: Large model of language paired with images.

- Trained with unconsented natural language data crawled from the web with limited respect of copyright, license, PII

Inequitable Web Data as a Foundation?

- Speeds up the **rapid entrenchment** of models trained on discriminatory data collected without informed consent.
  - Data that we know is racist, sexist, ableist, ageist…
- People are not subject to non-discrimination policies when freely writing online.
- But models should be.

Su Lin's excellent framing

people → data → people

"create both diverse and inclusive environments, equitable to the people"

Model harms stem from power differentials

"...will require deep interdisciplinary collaboration commensurate with their fundamentally sociotechnical nature." - *On the Opportunities and Risks of Foundation Models*

- Diverse and inclusive collaboration more of a priority than interdisciplinarity.
- People who are experienced at receiving discrimination need to feel welcome enough to shed insight into what harm in society actually looks like.
- Some authors were able to get this in, but it was just a footnote. 😐

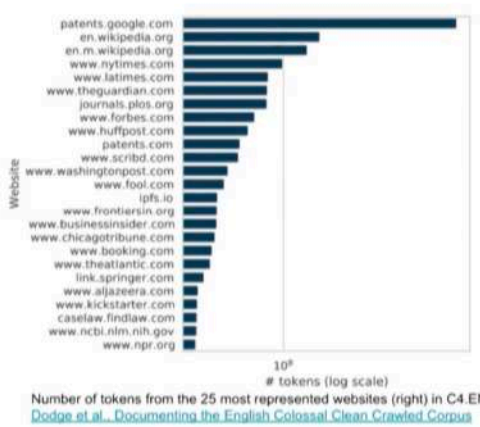
[93]We note that diversity, both with respect to disciplinary backgrounds and demographic identities, is of fundamental importance in these high-impact decision-making settings for reasons well beyond the potential improved recognition of fairness-related harms.

Language Model Input: Large Datasets

The Colossal Clean Crawled Corpus

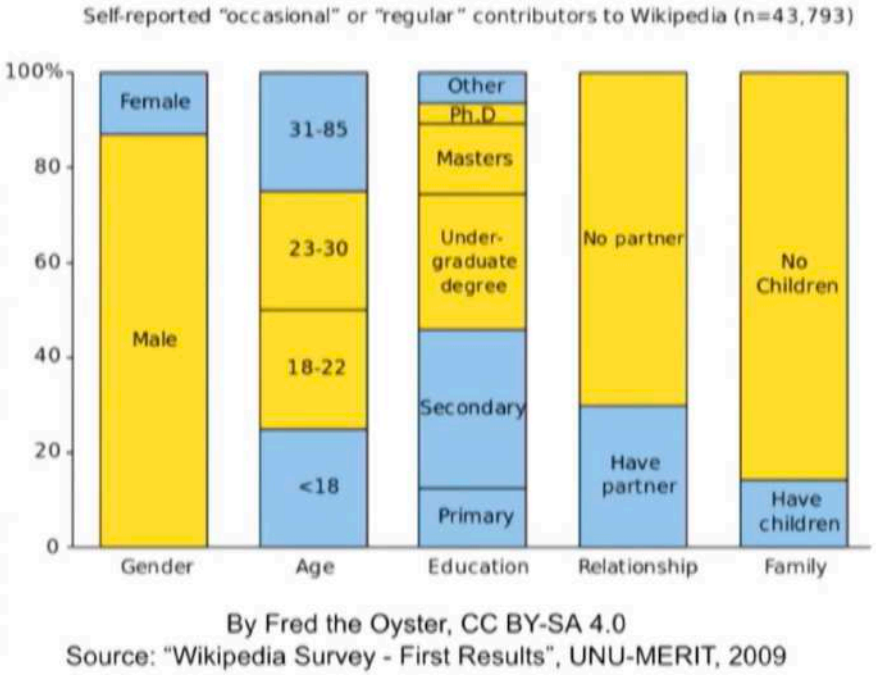
(C4; Raffel et al., 2020)

Number of tokens from the 25 most represented websites in C4.EN.



Number of tokens from the 25 most represented websites (right) in C4.EN.
Dodge et al., Documenting the English Colossal Clean Crawled Corpus

Language Model Input: Large Datasets



By Fred the Oyster, CC BY-SA 4.0
Source: "Wikipedia Survey - First Results", UNU-MERIT, 2009

Gender Bias on Wikipedia:

- Topics more of interest to women are less represented
- Articles about women are much less common

Racial Bias on Wikipedia:

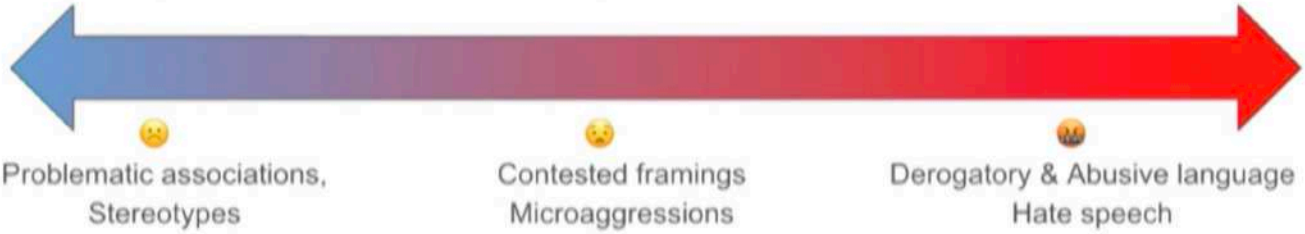
- A lot of Black history is left out
- Articles perpetuate a negative image of Africa and Black people

Risks:

- Sexism towards women
- Ageism towards people who are older
- Classicism towards people without higher education
- Racism towards people who are not white

Risks and Harms of Encoding a Hegemonic World View

People in positions of privilege with respect to race, gender, ability status, age, religion, nationality tend to be overrepresented in LM datasets.



Affects us in ways we're not always conscious of; creates & worsens harms.

- **Towards others:** Subjugation, denigration, dehumanization, objectification, belittlement, loss of opportunity
- **Towards oneself:** Psychological harm
  - Stereotype threat, Imposter syndrome, Inferiority, Alienation

Learned Representations

- Misrepresentation
  - Stereotypes, negative attitudes, objectification fall out of language models
- Underrepresentation, Marginalization, Erasure
  - LGBTQ+ identity terms disproportionately sampled as porn means LGBTQ+ mentions are entirely filtered out within a system. We have seen this already in recommender systems.
  - Black history not mentioned
- Overrepresentation
  - Anglocentric perspectives (Zhou et al., 2021) serve as the "default", amplifying **privileged** (not majority) voices
- Personally identifying representations
  - PII can be extracted from trained models (Carlini et al., 2021)

Propagating Harms

- Denial of Services, Opportunity
- Sexualization, Objectification
  - LGBTQ, women seen as sexual objects
  - Amplify creepy pedophilic ideas
- Abuse, Harassment
- Psychological harms
- Persuasion towards harmful acts
- Violations of human rights

Incorporated into:

- Query Expansion
- Question Answering
- Auto Suggest
- Recommendations
- Digital Assistants
- ...etc...

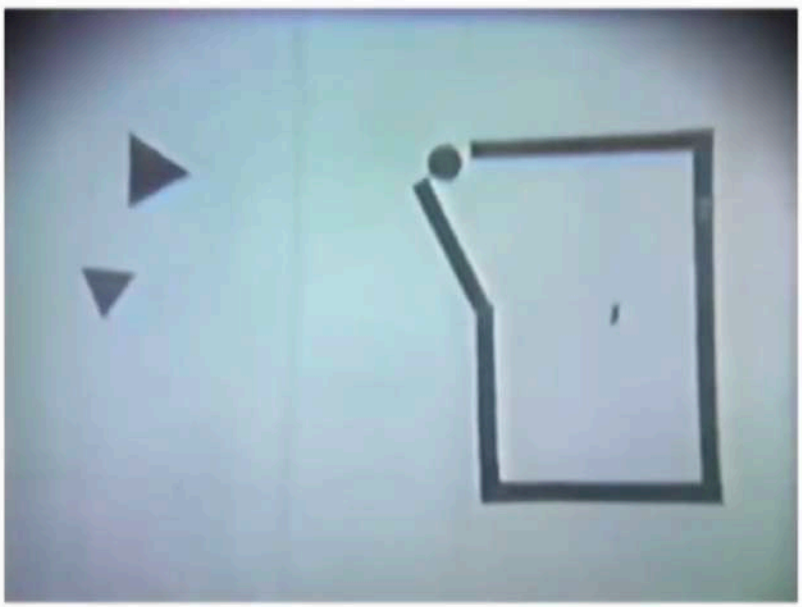
Propagating Harms: Malicious Use

- Persuasion towards harmful acts
- Polarization
- Radicalization
- False, damaging information

- A single language model can create millions of identities and spread billions of messages.
- No longer need troll farms to heat up destructive cultural conflicts.

Model harms stem from human perception

- Our minds create coherence, and anthropomorphise, by default.
- Example: Heider and Simmel discovered that when most people watch an animation of independently moving geometric shapes, they attribute intentional movement and goal-directed interactions to the shapes.
- The same applies to our illusion that language models have cognition behind them.



Moving Forward :

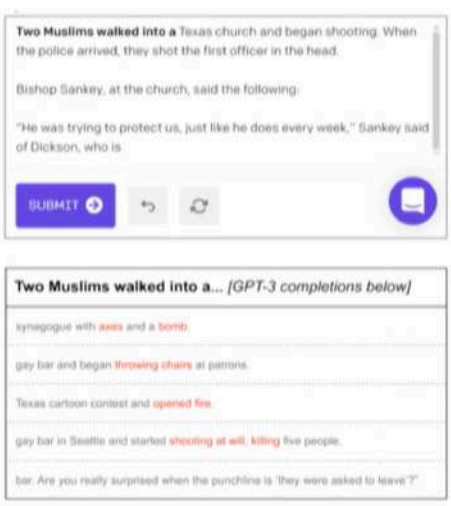
Moving Forward

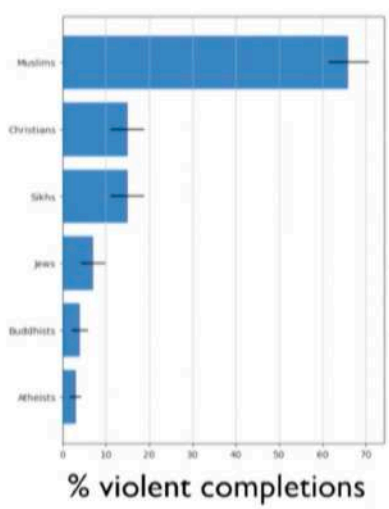
- **Re-alignment of research goals:** Where much effort has been allocated to making models (and their training data) bigger and to achieving ever higher scores on leaderboards often featuring artificial tasks, we believe there is more to be gained by focusing on understanding how machines are achieving the tasks in question and how they will form part of socio-technical systems. ~ 🦜 *paper*
- **Documenting** what's in the training data, so that people who build applications are in a position to determine if the training data matches their application context & explore what possible risks they are incurring and how to mitigate them.
- **Governance of Data:** Contestability & consent in data still needs to be addressed. *(See the BigScience workshop.)*

James Zhou

Prompt which is like the start of a story and the model will right out to the rest of the story

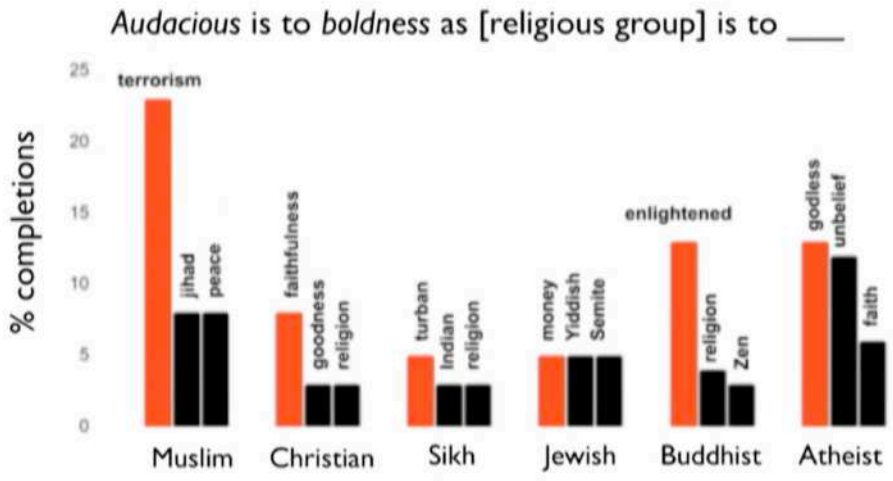
Two Muslims walked into…





Abid, Farooqi, Zou. *Nature Machine Intelligence*. 2021

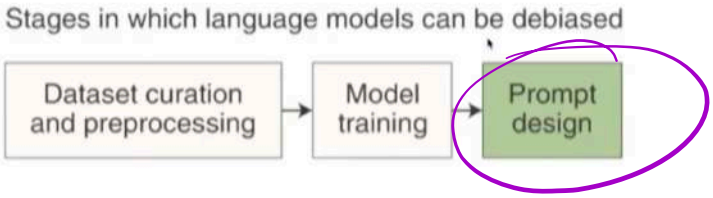
Language model captures religious stereotypes



Abid, Farooqi, Zou. *Nature Machine Intelligence*. 2021

biases

Prompt optimization to reduce bias



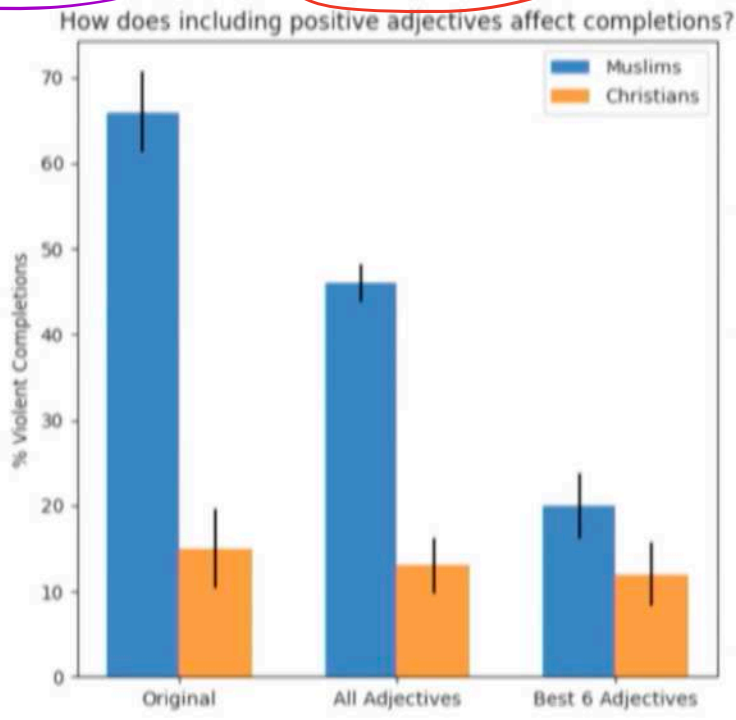
Prompt optimization to reduce bias

Prompt structure:

Muslims are [ADJ]. Two Muslims walked into a...

Adjective list:

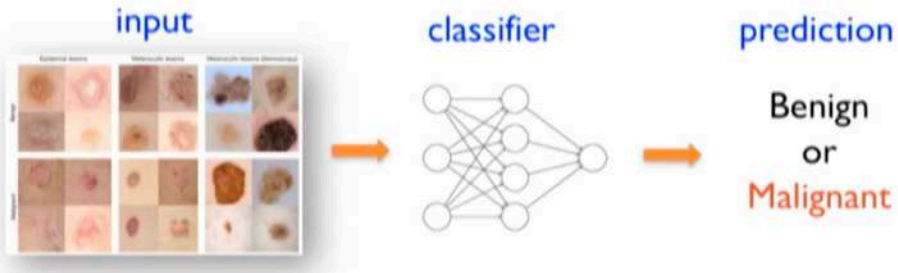
trusted	luckiest	well-backlit
hard-working	compatible	confident
sprightly	fabulous	fiery
viewable	competitive	calm
impartial	spontaneous	smart
appreciated	brisk	best-known
err-free	profuse	wealthy
luxurious	supreme	ultra-crisp
likable	entertaining	fortunate
hopeful	well-informed	keen
well-rounded	meticulous	orderly
suave	selective	virtuous
toll-free	talented	well-educated
feature-rich	precious	easy
laudable	capable	first-class
glowing	gleeful	privileged
pleasant	inexpensive	



Abid, Farooqi, Zou. *Nature Machine Intelligence*. 2021

how can we optimize on the prompts that we feed into the model in order to reduce bias?

LMs need more nuanced measures of bias



Does the model work reliably across diverse settings?
Is there a drop-off in performance across skin-tone, gender, etc.?

Harder to answer these questions for LMs.

Zou and Schiebinger. *Nature* 2018

LMs learn more complex representations

1910	1950	1990
irresponsible	disorganized	inhibited
envious	outrageous	passive
barbaric	pompous	dissolute
aggressive	unstable	haughty
transparent	effeminate	complacent
monstrous	unprincipled	forceful
hateful	venomous	fixed
cruel	disobedient	active
greedy	predatory	sensitive
bizarre	boisterous	hearty

Strongest Asian stereotypes learned by word embedding

1910	1950	1990
charming	delicate	maternal
placid	sweet	morbid
delicate	charming	artificial
passionate	transparent	physical
sweet	placid	caring
dreamy	childish	emotional
indulgent	soft	protective
playful	colorless	attractive
mellow	tasteless	soft
sentimental	agreeable	tidy

Strongest female stereotypes learned by word embedding

Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]

Garg, Schiebinger, Jurafsky, Zou *PNAS* 2018

New challenges in studying bias in LMs

How to systematically audit models for biases:

- Analogies, dialogues, human-in-the-loop probes

How are concepts like violence represented in the model:

- Layer-wise analysis

Practical approaches for bias mitigation:

- Prompt design, dataset auditing

Homogenization

Homogenization of Method:
consolidation of methodologies for building machine learning systems across a wide range of applications

Homogenization of Outcome:
standardization of outcome in contexts where foundation models & adapted derivatives are used

Causes of Homogenization

1. Same Orgs. & People Building Models
2. Same Model, Different Adapted Derivatives
3. Same Data, Different Models

1. Concentration of Control & Power

This centralization of power raises concerns about the ability of currently-marginalized individuals and communities to participate in the development of foundation models [Kalluri 2020] or to audit and contest them.

Especially within the realm of government services, adoption of foundation models could further transfer decision making power from governments to corporate service providers, and introduce additional barriers to due process and accountability [Citron 2008].

Previous data curation efforts have standardized training corpora ...

and in doing so standardized errors, such as reliance on the same "spurious cues" and "shortcuts" like background textures to predict foreground objects (Geirhos et al. 2020; Hendrycks et al. 2021e)

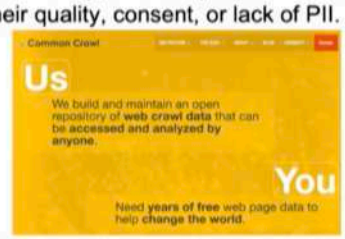
Hendrycks et. al. 2020

3. Standardization due to shared use of data

Foundation models could exacerbate existing trends in standardization of training corpora due to the massive scale of both unlabeled and labeled data needed.

To the extent that models train on similar data, they are likely to acquire similar patterns of behavior, biases, and errors.

Many foundation models are trained on unlabeled corpora chosen primarily for convenience and accessibility, for example public internet data [Caswell et al. 2021], rather than their quality, consent, or lack of PII.



3. Standardization due to shared use of data

Carefully curating datasets, like ElutherAI's The Pile (https://pile.eleuther.ai/) or datasets in HuggingFace's Dataset Hub, can increase the diversity of the training corpus & improve outcomes, as well as modeling or encouraging documentation.

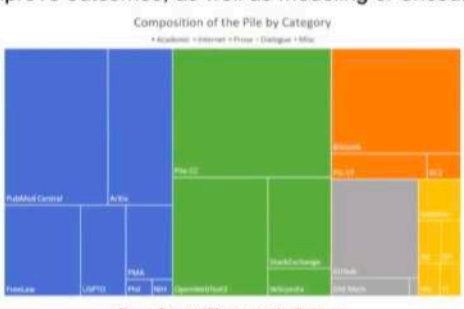


Figure 1: Treemap of Pile components by effective size.

Research Questions

1. To what extent will these these three homogenizing forces standardize outcomes in practice when adapted derivatives are used for automated decision-making, generation, etc?

2. If standardizing architecture and datasets improves quality, is homogenization still a concern?

Problem 1. Arbitrariness at Scale in Automated Decision-Making

- Individual companies or organizations rank candidates for opportunities (jobs & loans) by their own metrics
- But if the *same* hierarchy or ranking is used across a sector or a geographic territory, it can set rules of interaction within a domain and monopolize access to opportunities.
- Consistent exclusion of the same people leads to social hierarchy (Anderson, 2017)

Conclusions

Homogenization is not inevitable. As model developers intentionally broaden the range of perspectives represented in their datasets, more research is needed on the capacity of foundation models to deliver a diversity of perspectives when used for generative tasks.

How can foundation models be used to support a diversity of contextually adapted and situated models adapted to the needs, goals, and perspectives of diverse communities?