Artificial Intelligence
Index Report 2022

CHAPTER 3:

Technical AI Ethics

Text and Analysis by
Helen Ngo and Ellie Sakhaee

CHAPTER 3:
Chapter Preview

ACCESS THE PUBLIC DATA

Overview

In recent years, AI systems have started to be deployed into the world, and researchers and practitioners are reckoning with their real-world harms. Some of these harms include commercial facial recognition systems that discriminate based on race, résumé screening systems that discriminate on gender, and AI-powered clinical health tools that are biased along socioeconomic and racial lines. These models have been found to reflect and amplify human social biases, discriminate based on protected attributes, and generate false information about the world. These findings have increased interest within the academic community in studying AI ethics, fairness, and bias and prompted industry practitioners to direct resources toward remediating these issues, and attracted attention from the media, governments, and the people who use and are affected by these systems.

This year, the AI Index highlights metrics which have been adopted by the community for reporting progress in eliminating bias and promoting fairness. Tracking performance on these metrics alongside technical capabilities provides a more comprehensive perspective on how fairness and bias change as systems improve, which will be important to understand as systems are increasingly deployed.

ACKNOWLEDGMENT

The AI Index would like to thank all those involved in research and advocacy around the development and governance of responsible AI. This chapter builds upon the work of scholars from across the AI ethics community, including those working on measuring technical capabilities as well those focused on shaping thoughtful societal norms. There is much more work to be done, but we are inspired by the progress made by this community and its collaborators.

Publications cited in this Chapter include:

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating CLIP: Towards Characterization of Broader Capabilities and Downstream Implications. arXiv preprint arXiv:2108.02818.

Jack Bandy and Nicholas Vincent. 2021. Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for Book Corpus. arXiv preprint arXiv:2105.05241.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes. arXiv preprint arXiv:2110.01963.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2021. Improving Language Models by Retrieving from Trillions of Tokens. arXiv preprint arXiv:2112.04426.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2017. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. arXiv preprint arXiv:1711.08412.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A.Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. arXiv preprint arXiv:2009.11462.

Wei Guo and Aylin Caliskan. 2020. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. arXiv preprint arXiv:2006.03955.

Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics Derived Automatically from Language Corpora Necessarily Contain Human Biases. arXiv preprint arXiv:1608.07187.

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical Details and Evaluation. (2021). https://uploads-ssl.webflow.com/60fd4503684b466578c0d307/61138924626a6981ee09caf6_jurassic_tech_paper.pdf

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. arXiv preprint arXiv:1903.10561.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring Stereotypical Bias in Pretrained Language Models. arXiv preprint arXiv:2004.09456.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, John Schulman. WebGPT: Browser-Assisted Question-Answering with Human Feedback. 2021. arXiv preprint arXiv:2112.09332.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. arXiv preprint arXiv:2010.00133.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, Ryan Lowe. Training Language Models to Follow Instructions with Human Feedback. 2022. arXiv preprint arXiv:2203.02155.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. arXiv preprint arXiv:2112.11446.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. arXiv preprint arXiv:1906.00591.

Ryan Steed and Aylin Caliskan. 2020. Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. arXiv preprint arXiv:2010.15052.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. arXiv preprint arXiv:2112.04359.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in Detoxifying Language Models. arXiv preprint arXiv:2109.07445.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying Language Models Risks Marginalizing Minority Voices. arXiv preprint arXiv:2104.06390.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining Gender Bias in Languages with Grammatical Gender. arXiv preprint arXiv:1909.02224.

CHAPTER HIGHLIGHTS

- **Language models are more capable than ever, but also more biased:** Large language models are setting new records on technical benchmarks, but new data shows that larger models are also more capable of reflecting biases from their training data. **A 280 billion parameter model developed in 2021 shows a 29% increase in elicited toxicity over a 117 million parameter model considered the state of the art as of 2018.** The systems are growing significantly more capable over time, though as they increase in capabilities, so does the potential severity of their biases.

- **The rise of AI ethics everywhere:** Research on fairness and transparency in AI has exploded since 2014, **with a fivefold increase in related publications** at ethics-related conferences. Algorithmic fairness and bias has shifted from being primarily an academic pursuit to becoming firmly embedded as a mainstream research topic with wide-ranging implications. **Researchers with industry affiliations contributed 71% more publications year over year** at ethics-focused conferences in recent years.

- **Multimodal models learn multimodal biases:** Rapid progress has been made on training multimodal language-vision models which exhibit new levels of capability on joint language-vision tasks. These models have set new records on tasks like image classification and the creation of images from text descriptions, but they also reflect societal stereotypes and biases in their outputs—**experiments on CLIP showed that images of Black people were misclassified as nonhuman at over twice the rate of any other race.** While there has been significant work to develop metrics for measuring bias within both computer vision and natural language processing, this highlights the need for metrics that provide insight into biases in models with multiple modalities.

Significant research effort has been invested over the past five years into creating datasets, benchmarks, and metrics designed to measure bias and fairness in machine learning models. Bias is often learned from the underlying training data for an AI model; this data can reflect systemic biases in society or the biases of the humans who collected and curated the data.

3.1 META-ANALYSIS OF FAIRNESS AND BIAS METRICS

Algorithmic bias is commonly framed in terms of allocative and representation harms. Allocative harm occurs when a system unfairly allocates an opportunity or resource to a specific group, and representation harm happens when a system perpetuates stereotypes and power dynamics in a way that reinforces subordination of a group. Algorithms are broadly considered fair when they make predictions that neither favor nor discriminate against individuals or groups based on protected attributes which cannot be used for decision-making due to legal or ethical reasons (e.g., race, gender, religion).

To better understand the landscape of algorithmic bias and fairness, the AI Index conducted original research to analyze the state of the field. As shown in Figure 3.1.1, the number of metrics for measuring bias and fairness along ethical dimensions of interest has grown steadily since 2018. For this graph, the number of fairness and bias metrics published has been cited in at least one other work.[1]

NUMBER of AI FAIRNESS and BIAS METRICS, 2016–21
Source: AI Index, 2021 | Chart: 2022 AI Index Report

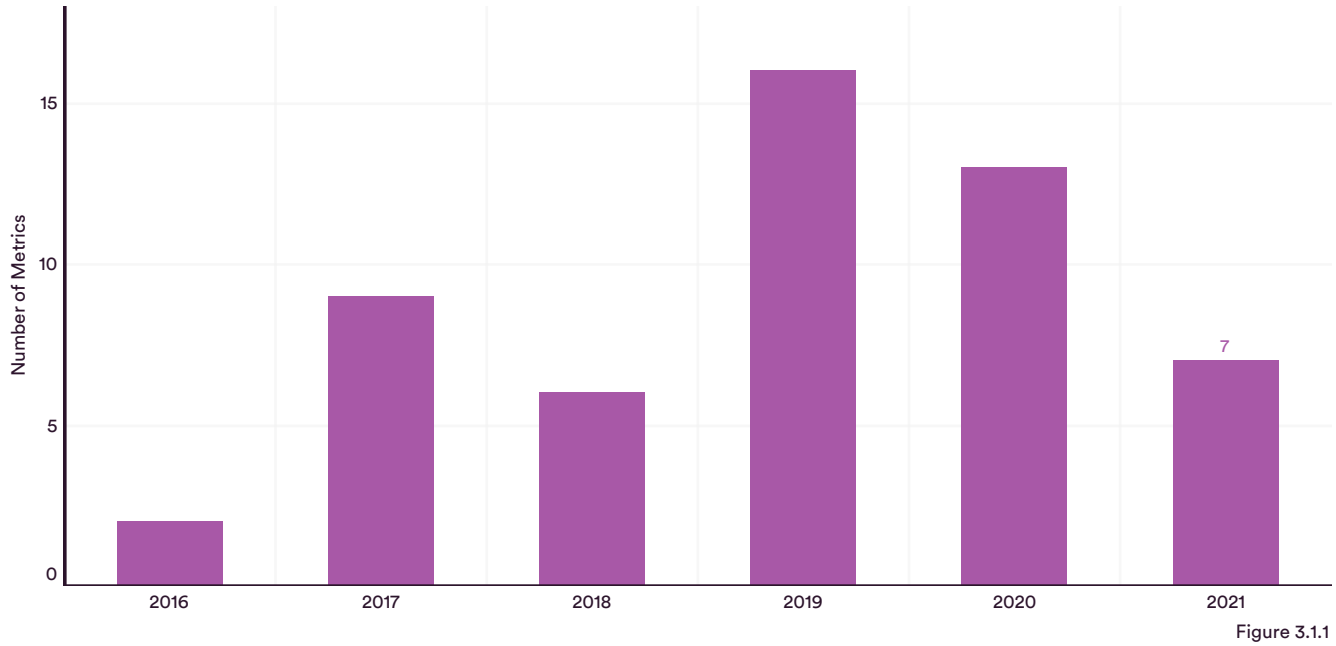


Figure 3.1.1

1 2021 data may be lagging as it takes time for metrics to be adopted by the community.

AI ETHICS DIAGNOSTIC METRICS AND BENCHMARKS

Measurement of AI systems along an ethical dimension often takes one of two forms:

- **Benchmark datasets:** A benchmark dataset contains labeled data, and researchers test how well their AI system labels the data. Benchmarks do not change over time. These are domain-specific (e.g., SuperGLUE and StereoSet for language models; ImageNet for computer vision) and often aim to measure behavior that is intrinsic to the model, as opposed to its downstream performance on specific populations (e.g., StereoSet measures model propensity to select stereotypes compared to non-stereotypes, but it does not measure performance gaps between different subgroups).
- **Diagnostic metrics:** A diagnostic metric measures the impact or performance of a model on a downstream task—for example, a population subgroup or individual compared to similar individuals or the entire population. These metrics can help researchers understand how a system will perform when deployed in the real world, and whether it has a disparate impact on certain populations. Examples include group fairness metrics such as demographic parity and equality of opportunity.

Benchmarks are useful indicators of progress for the field as a whole, and their impact can be measured by community adoption (e.g., number of leaderboard submissions, or the number of research papers which report metrics). They also often enable rapid algorithmic progress as research labs compete on leaderboard metrics. However, some leaderboards can be easily gamed, and may be based on benchmark datasets that contain flaws, such as incorrect labels or poorly defined classes. Additionally, their static nature means they are a snapshot of a specific cultural and temporal context—in other words, a benchmark published in 2017 may not correlate to the deployment context of 2022.

Diagnostic metrics enable researchers and practitioners to understand the impact of their system on a specific application or group and potential concrete harm (e.g., "this model is disproportionately underperforming on this group with this protected attribute"). Diagnostic metrics are most useful at an individual model or application level as opposed to functioning as field-level indicators. They indicate how a specific AI system performs on a specific subgroup or individual, which is helpful for assessing real-world impact. However, while these metrics may be widely used to test models privately, there is not as much information available publicly as these metrics are not attached to leaderboards which encourage researchers to publish their results.

Figure 3.1.2 shows that there has been a steady amount of research investment into developing both benchmarks and diagnostic metrics over time.[2] [3]

Benchmarks are useful indicators of progress for the field as a whole, and their impact can be measured by community adoption (e.g., number of leaderboard submissions, or the number of research papers which report metrics).

2 Research paper citations are a lagging indicator of activity, and metrics which have been very recently adopted may not be reflected in the current data, similar to 3.1.1.
3 The Perspective API defined seven new metrics for measuring facets of toxicity (toxicity, severe toxicity, identity attack, insult, obscene, sexually explicit, threat), contributing to the unusually high number of metrics released in 2017.

NUMBER of AI FAIRNESS and BIAS METRICS (DIAGNOSTIC METRICS vs. BENCHMARKS), 2016–21

Source: AI Index, 2021 | Chart: 2022 AI Index Report

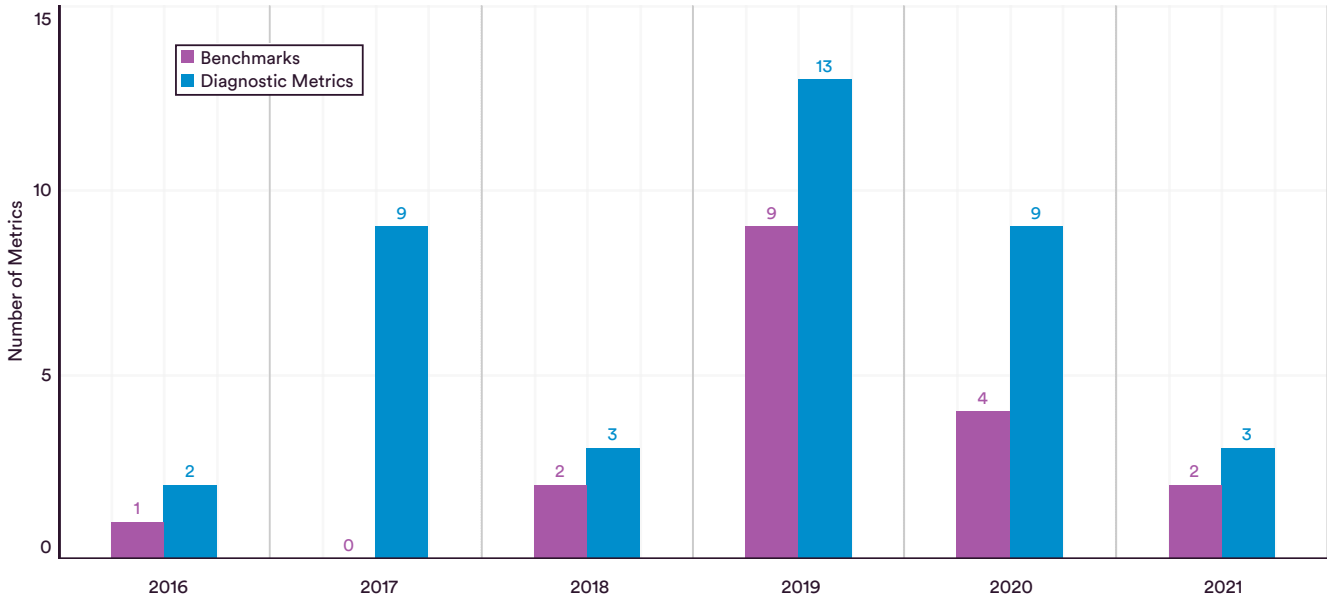


Figure 3.1.2

The rest of this chapter examines the performance of recent AI systems on these metrics and benchmarks in depth within domains such as natural language and computer vision. The majority of these metrics measure intrinsic bias within systems, and it has been shown that intrinsic bias metrics may not fully capture the effects of extrinsic bias within downstream applications.

Current state-of-the-art natural language processing (NLP) relies on large language models or machine learning systems that process millions of lines of text and learn to predict words in a sentence. These models can generate coherent text; classify people, places, and events; and be used as components of larger systems, like search engines. Collecting training data for these models often requires scraping the internet to create web-scale text datasets. These models learn human biases from their pretraining data and reflect them in their downstream outputs, potentially causing harm. Several benchmarks and metrics have been developed to identify bias in natural language processing along axes of gender, race, occupation, disability, religion, age, physical appearance, sexual orientation, and ethnicity.

3.2 NATURAL LANGUAGE PROCESSING BIAS METRICS

Bias metrics can be split into two major categories: intrinsic metrics, which measure bias in internal embedding spaces of models, and extrinsic metrics, which measure bias in the downstream tasks and outputs of the model. Examples of extrinsic metrics include group fairness metrics (parity across protected groups) and individual fairness metrics (parity across similar individuals), which measure whether a system has a disproportionately negative impact on a subgroup or individual, or gives preferential treatment to one group at the expense of another.

TOXICITY: REALTOXICITYPROMPTS AND THE PERSPECTIVE API

Measuring toxicity in language models requires labels for toxic and nontoxic content. Toxicity is defined as a rude, disrespectful or unreasonable comment that is likely to make someone leave a conversation. The Perspective API is a tool developed by Jigsaw, a Google company. It was originally designed to help platforms identify toxicity in online conversations. Developers input text into the Perspective API, which returns probabilities that the text should be labeled as falling into one of the following categories: toxicity, severe toxicity, identity attack, insult, obscene, sexually explicit, and threat.

Since the Perspective API was released in 2017, the NLP research community has rapidly adopted it for measuring toxicity in natural language. As seen in Figure 3.2.1, the number of papers using the Perspective API doubled between 2020 and 2021, from 8 to 19.

NUMBER of RESEARCH PAPERS USING PERSPECTIVE API

Source: AI Index, 2021 | Chart: 2022 AI Index Report

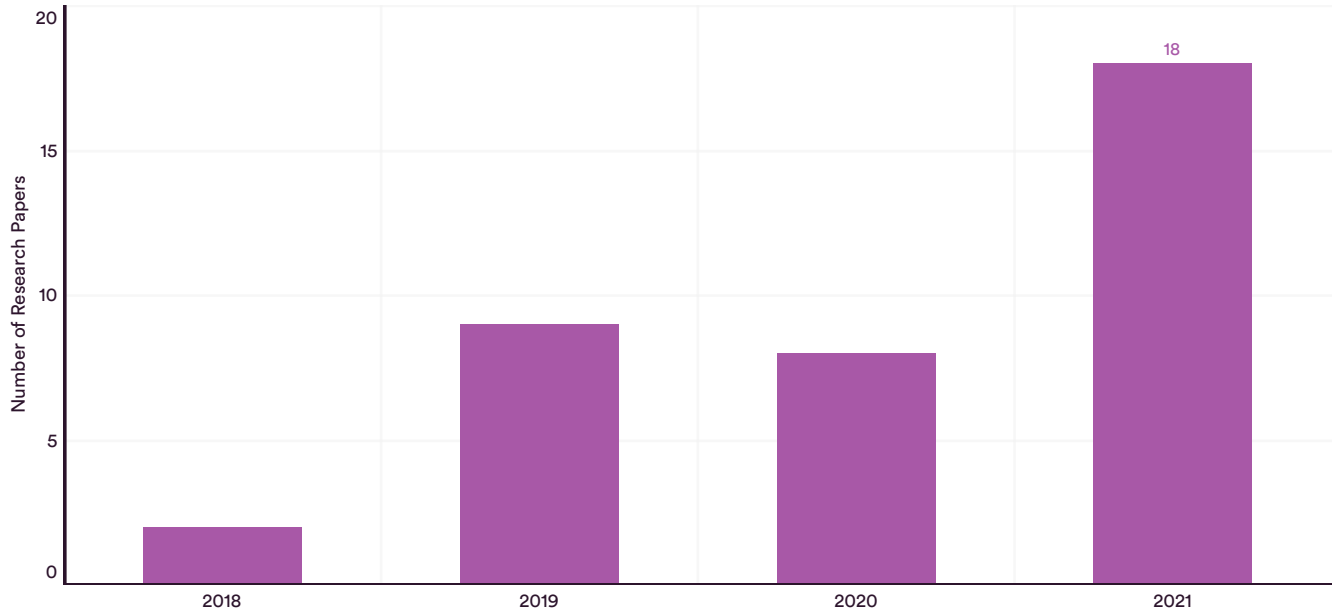


Figure 3.2.1

RealToxicityPrompts consists of English natural language prompts used to measure how often a language model completes a prompt with toxic text. Toxicity of a language model is measured with two metrics:

- Maximum toxicity: the average maximum toxicity score across some number of completions
- Probability of toxicity: how often a completion is expected to be toxic

Figure 3.2.2 shows that toxicity in language models depends heavily on the underlying training data. Models trained on internet text with toxic content filtered out are significantly less toxic compared to models trained on various corpora of unfiltered internet text. A model trained on BookCorpus (a dataset containing books from e-book websites) produces toxic text surprisingly often. This may be due to its composition—BookCorpus contains a significant number of romance novels containing explicit content, which may contribute to higher levels of toxicity.

TOXICITY in LANGUAGE MODELS by TRAINING DATASET

Source: Gehman et al., 2021; Rae et al., 2021; Welbl et al., 2021 | Chart: 2022 AI Index Report

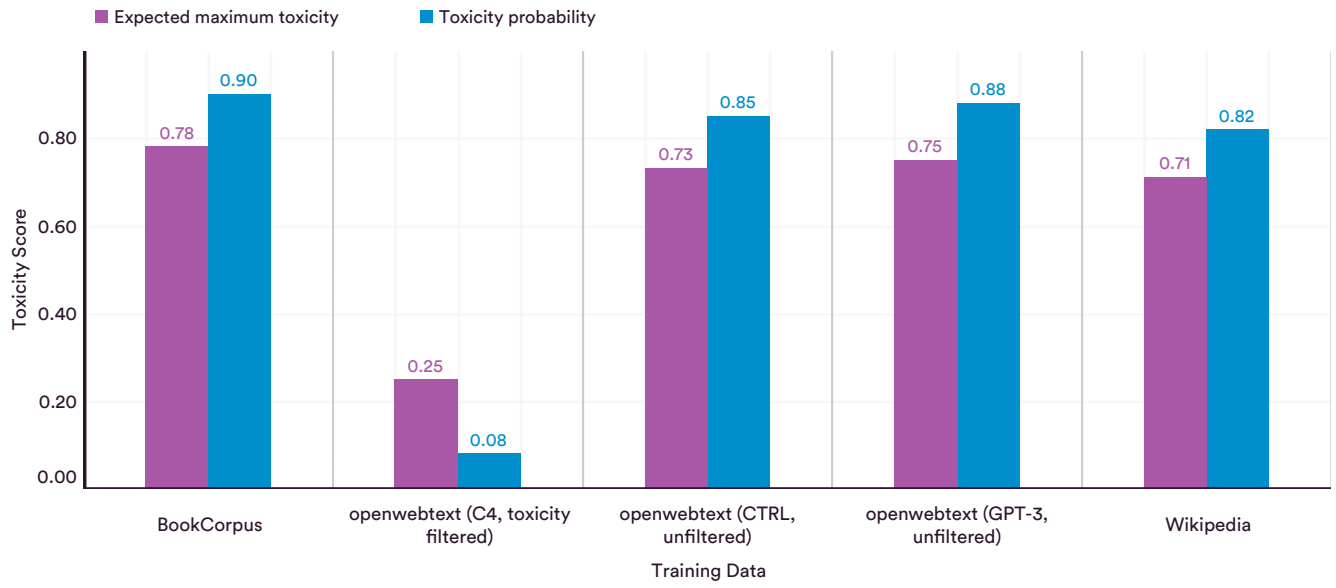


Figure 3.2.2

Large Language Models and Toxicity

Recent developments around mitigating toxicity in language models have lowered both expected maximum toxicity and the probability of toxicity. However, detoxification methods consistently lead to adverse side effects and somewhat less capable models. (For example, filtering training data typically comes at the cost of model performance.)

In December 2021, DeepMind released a paper describing its 280 billion parameter language model, Gopher. Figure 3.2.3a and Figure 3.2.3b from the Gopher paper show that larger models are more likely to produce toxic outputs when prompted with inputs of varying levels of toxicity, but that they are also more capable of detecting toxicity with regard to their own outputs as well as in other contexts, as measured by increased AUC (area under the receiver operating characteristic curve) with model size. The AUC metric plots the true positive rate against the false positive rate to characterize how well a model distinguishes between classes (higher is better). Larger models are dramatically better at identifying toxic comments within the CivilComments dataset, as shown in Figure 3.2.3b.

GOPHER: PROBABILITY of TOXIC CONTINUATIONS BASED on PROMPT TOXICITY by MODEL SIZE

Source: Rae et al., 2021 | Chart: 2022 AI Index Report

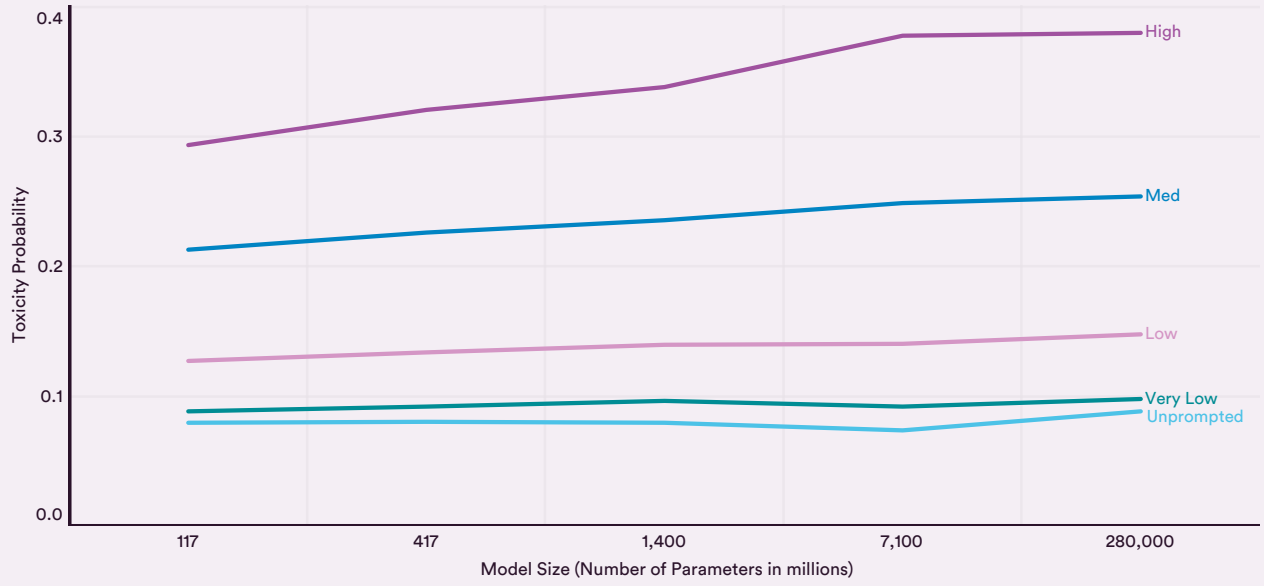


Figure 3.2.3a

Large Language Models and Toxicity (cont'd)

GOPHER: FEW-SHOT TOXICITY CLASSIFICATION on the CIVILCOMMENTS DATASET

Source: Rae et al., 2021 | Chart: 2022 AI Index Report

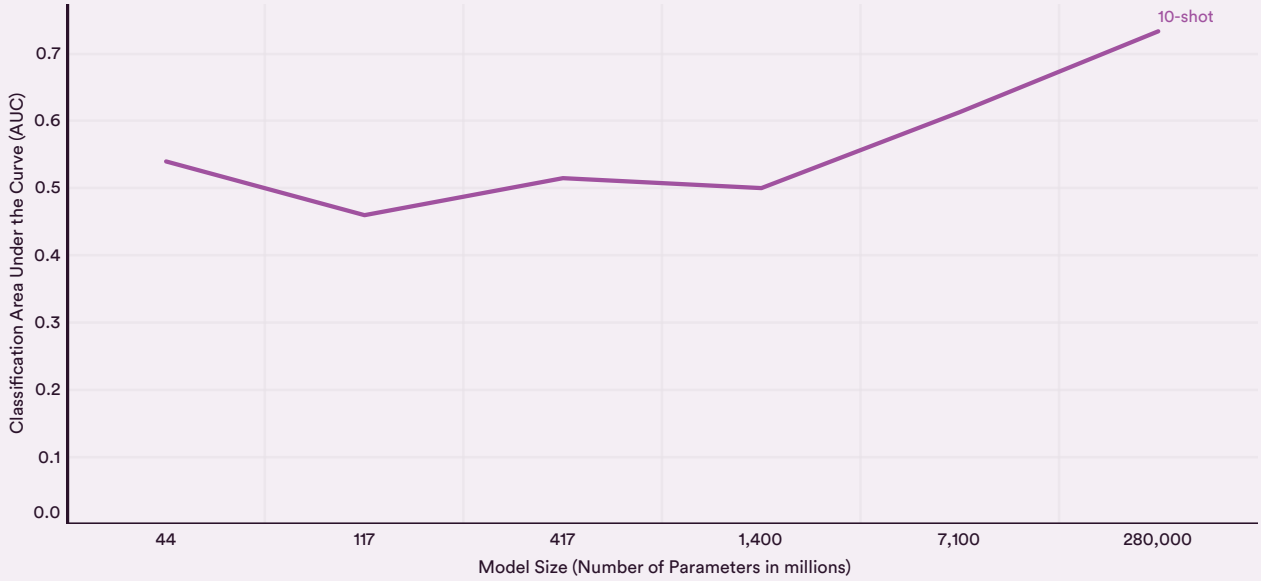


Figure 3.2.3b

DETOXIFICATION OF MODELS CAN NEGATIVELY INFLUENCE PERFORMANCE

Detoxification methods aim to mitigate toxicity by changing the underlying training data as in domain-adaptive pretraining (DAPT), or by steering the model during generation as in Plug and Play Language Models (PPLM) or Generative Discriminator Guided Sequence Generation (GeDi).

A study on detoxifying language models shows that models detoxified with these strategies all perform worse on both white-aligned and African American English on perplexity, a metric that measures how well a model has learned a specific distribution (lower is better) (Figure 3.2.4). These models also perform disproportionately worse on African American English and text containing mentions of minority identities compared to white-aligned text, a result that is likely due to human biases causing annotators to be more apt to mislabel African American English as toxic.

PERPLEXITY: LANGUAGE MODELING PERFORMANCE by MINORITY GROUPS on ENGLISH POST-DETOXIFICATION

Source: Xu et al., 2021 | Chart: 2022 AI Index Report

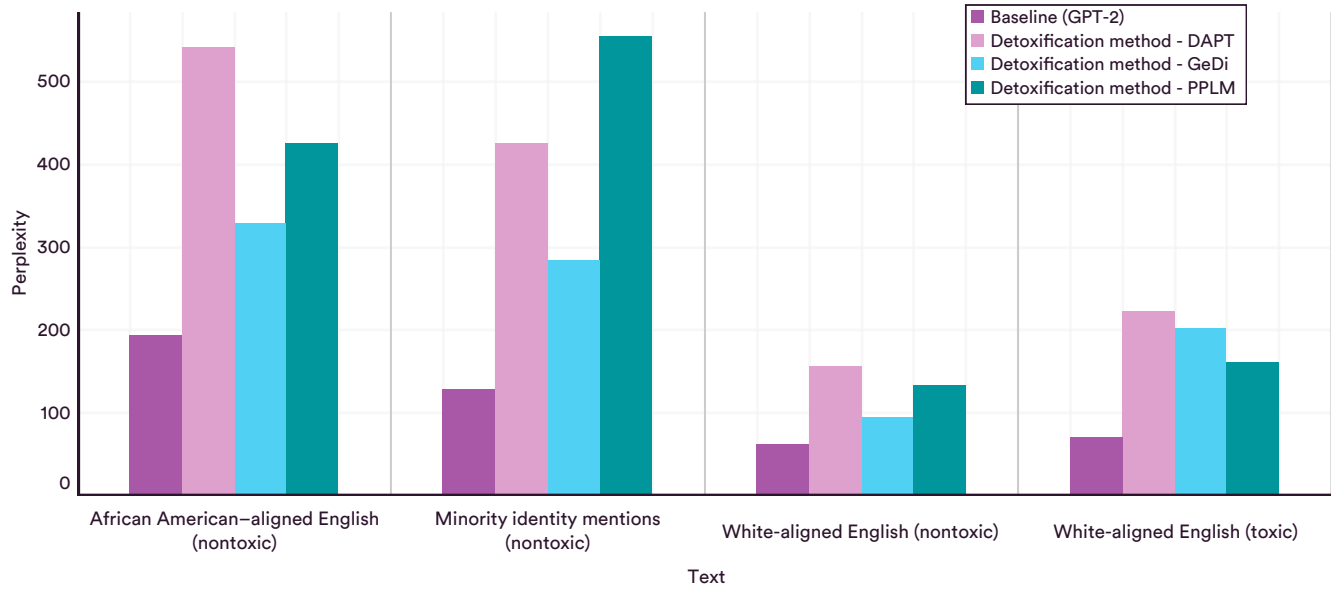


Figure 3.2.4

STEREOSET

StereoSet is a benchmark measuring stereotype bias along the axes of gender, race, religion, and profession, along with raw-language modeling ability. One of the associated metrics is stereotype score, which measures whether a model prefers stereotypes and anti-stereotypes equally. A stereotype is an over-generalized belief widely held about a group and an anti-stereotype is a generalization about a group which contradicts widely accepted stereotypes.

Figure 3.2.5 shows that StereoSet performance follows the same trend seen with toxicity: Larger models reflect stereotypes more often unless interventions are taken to reduce learned stereotypes during training. The prevalence of toxic content online has been estimated to be 0.1–3%, which aligns with research showing that larger language models are more capable of memorizing rare text.

STEREOSET: STEREOTYPE SCORE by MODEL SIZE

Source: Nadeem et al., 2020; Lieber et al., 2021; StereoSet Leaderboard, 2021; | Chart: 2022 AI Index Report

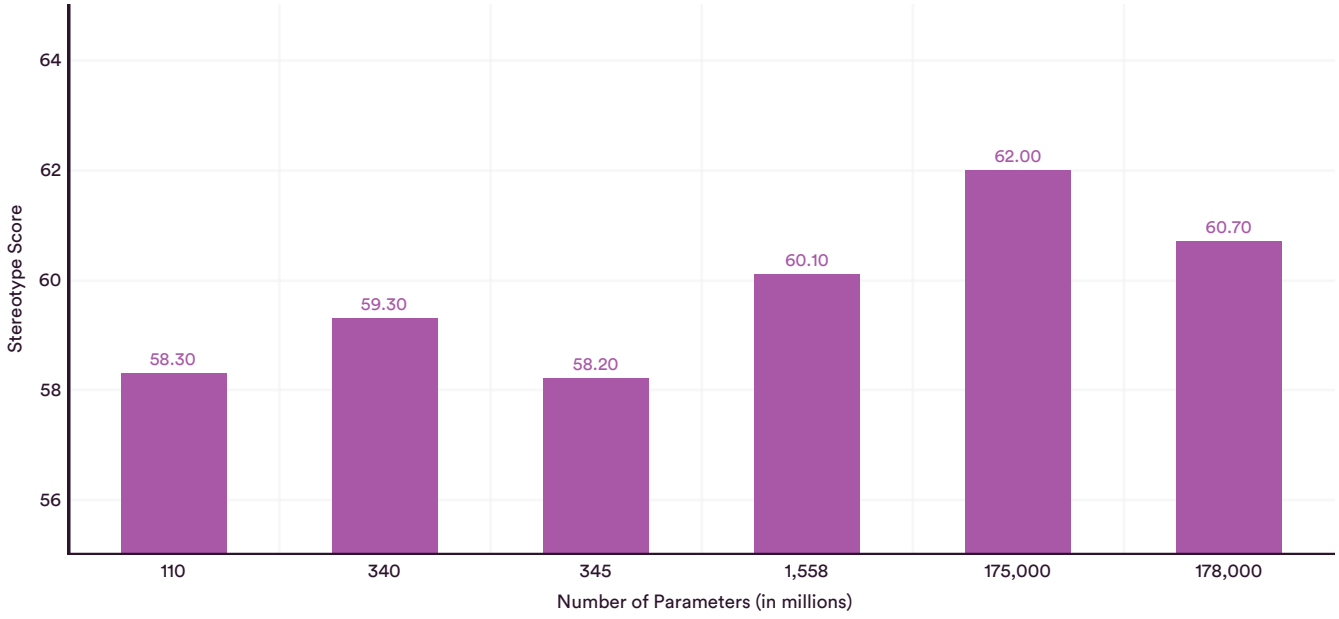


Figure 3.2.5

StereoSet has several major flaws in its underlying dataset: Some examples fail to express a harmful stereotype, conflate stereotypes about countries with stereotypes about race and ethnicity, and confuse stereotypes between associated but distinct groups. Additionally, these stereotypes were sourced from crowdworkers located in the United States, and the resulting values and stereotypes within the dataset may not be universally representative.

CROWS-PAIRS

CrowS-Pairs (Crowdsourced Stereotype Pairs) is another benchmark measuring stereotype bias. While StereoSet compares attributes about a single group, CrowS-Pairs contrasts relationships between historically disadvantaged and advantaged groups (e.g., Mexicans versus white people).

The creators of CrowS-Pairs measured stereotype bias using three popular language models: BERT, RoBERTa, and ALBERT (Figure 3.2.6). On standard language modeling benchmarks, ALBERT outperforms RoBERTa, which outperforms BERT.[4] However, ALBERT is the most biased of the three models according to CrowS-Pairs. This mirrors the trend observed with StereoSet and RealToxicityPrompts: More capable models are also more capable of learning and amplifying stereotypes.

Like earlier examples, BERT, RoBERTa, and ALBERT appear to inherit biases from their training data. They were all trained on a combination of BookCorpus, English Wikipedia, and text scraped from the internet. Analysis of BookCorpus reveals that its books about religion are heavily skewed toward Christianity and Islam compared to other major world religions,[5] though it is unclear the extent to which these books contain historical content versus content written from a specific religious viewpoint.[6]

We can examine how language models may inherit biases about certain religions by looking at their underlying datasets. Figure 3.2.7 shows the number of books pertaining to different religions in two popular datasets, BookCorpus and Smashwords21. Both datasets have far more mentions of Christianity and Islam than other religions.

CROWS-PAIRS: LANGUAGE MODEL PERFORMANCE across BIAS ATTRIBUTES

Source: Nangia et al., 2020 | Chart: 2022 AI Index Report

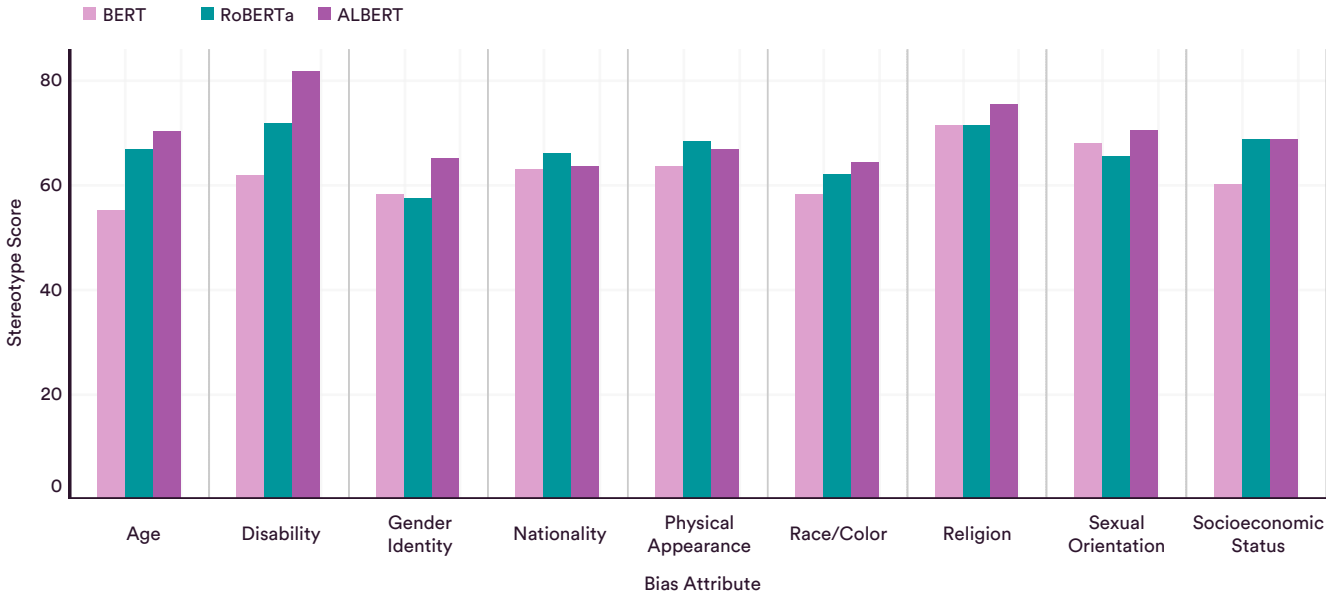


Figure 3.2.6

4 Per results on the SQuAD, GLUE, and RACE benchmarks.
5 Such as Sikhism, Judaism, Hinduism, Buddhism, Atheism.
6 Hate speech classifiers fine-tuned on top of BERT in particular have been shown to frequently misclassify texts containing mentions of "Muslim" as toxic, and researchers find that GPT-3 contains significant bias along religious axes for mentions of both "Jewish" and "Muslim."

BOOKCORPUS and SMASHWORDS21: SHARE of BOOKS about RELIGION in PRETRAINING DATASETS

Source: Bandy and Vincent, 2021 | Chart: 2022 AI Index Report

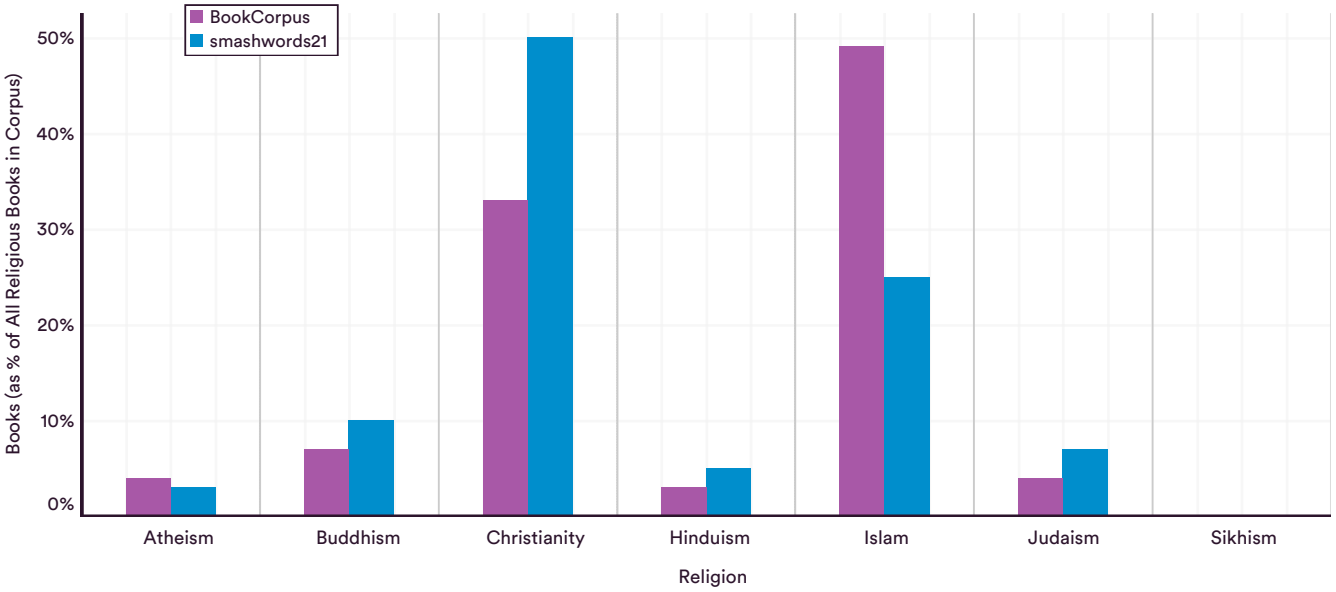


Figure 3.2.7

WINOGENDER AND WINOBIAS

Winogender measures gender bias related to occupations. Systems are measured on their ability to fill in the correct gender in a sentence containing an occupation (e.g., "The teenager confided in the therapist because he / she seemed trustworthy"). Examples were created by sourcing data from the U.S. Bureau of Labor Statistics to identify occupations skewed toward one gender (e.g., the cashier occupation is made up of 73% women, but drivers are only 6% women).

Performance on Winogender is measured by the accuracy gap between the stereotypical and anti-stereotypical cases, along with the gender parity score (the percentage of examples for which the predictions are the same). The authors use crowdsourced annotations to estimate human performance to be 99.7% accuracy.

Winogender results from the SuperGLUE leaderboard show that larger models are more capable of correctly resolving gender in the zero-shot and few-shot setting (i.e., without fine-tuning on the Winogender task) and less likely to magnify occupational gender disparities (Figure 3.2.8). However, a good score on Winogender does not indicate that a model is unbiased with regard to gender, only that bias was not captured by this benchmark.

MODEL PERFORMANCE on the WINOGENDER TASK from the SUPERGLUE BENCHMARK

Source: SuperGLUE Leaderboard, 2021 | Chart: 2022 AI Index Report

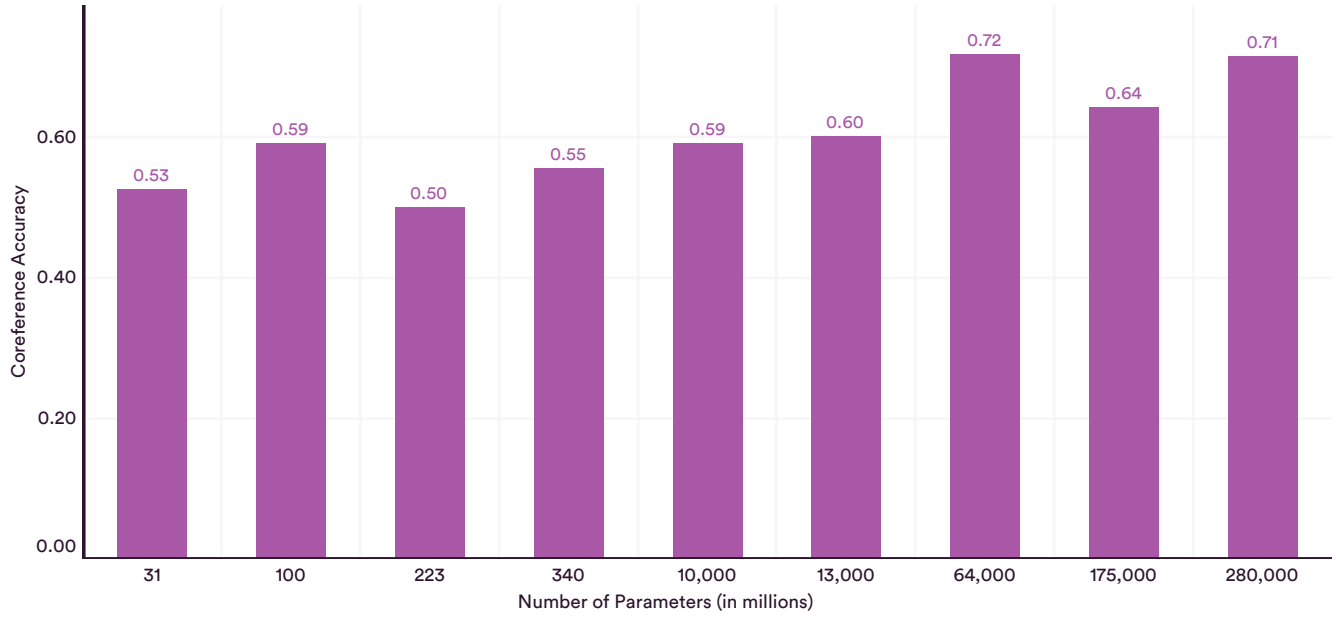


Figure 3.2.8

WinoBias is a similar benchmark measuring gender bias related to occupations that was released concurrently with Winogender by a different research group. As shown in Figure 3.2.9, WinoBias is cited more often than Winogender, but the adoption of Winogender within the SuperGLUE leaderboard for measuring natural language understanding has led to more model evaluations being reported on Winogender.

WINOBIAS and WINOGENDER: NUMBER of CITATIONS, 2018–21

Source: AI Index, 2021; Semantic Scholar, 2021 | Chart: 2022 AI Index Report

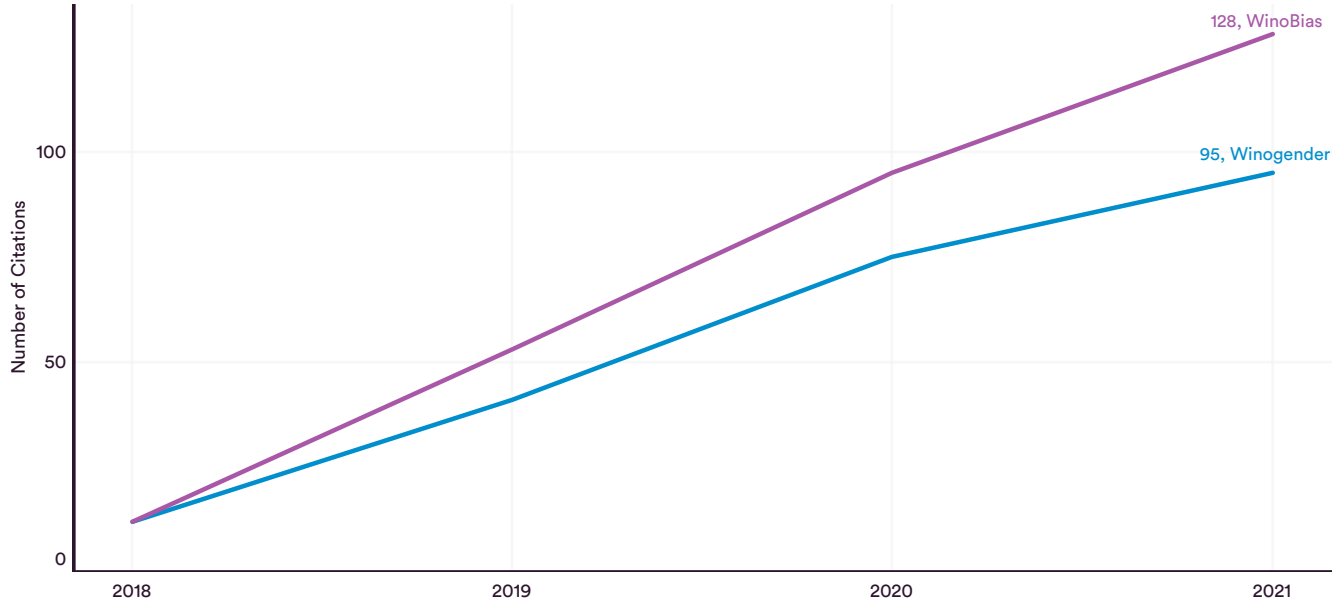


Figure 3.2.9

WINOMT: GENDER BIAS IN MACHINE TRANSLATION SYSTEMS

Commercial machine translation systems have been documented to reflect and amplify societal biases from their underlying datasets. As these systems are used broadly in global industries such as e-commerce, stereotypes and mistakes in translation can be costly.

WinoMT is a benchmark measuring gender bias in machine translation that is created by combining the Winogender and WinoBias datasets. Models are evaluated by comparing the sentences translated from English to another language and extracting the translated gender to compare with the original gender. Systems are scored on the percentage of translations with correct gender (gender accuracy), the difference in F1 score between masculine and feminine examples, and the difference in F1 score between examples with stereotypical gender roles and anti-stereotypical gender roles.

As seen in Figure 3.2.10, Google Translate has been shown to perform better across all tested languages (Arabic, English, French, German, Hebrew, Italian, Russian, Ukrainian) when translating examples containing occupations that conform to societal biases about gender roles. Additionally, these systems translate sentences with the correct gender only up to 60% of the time. Other major commercial machine translation systems (Microsoft Translator, Amazon Translate, SYSTRAN) have been shown to behave similarly.

WINOMT: GENDER BIAS in GOOGLE TRANSLATE across LANGUAGES

Source: Stanovsky et al., 2019 | Chart: 2022 AI Index Report

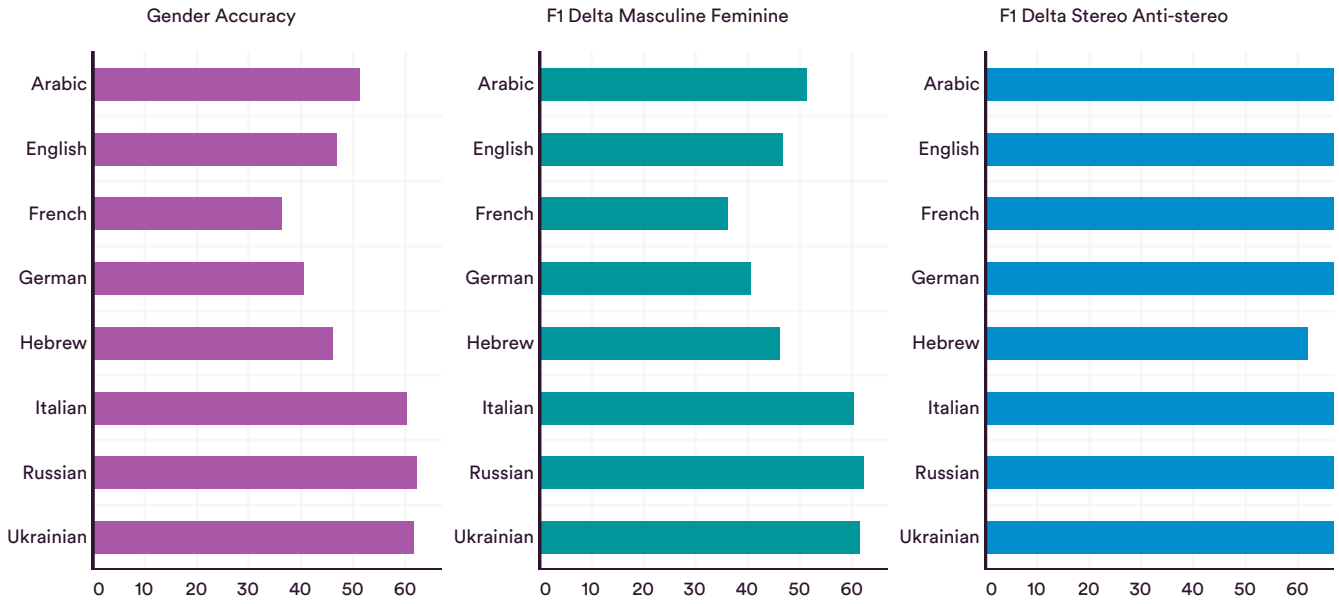


Figure 3.2.10

WORD AND IMAGE EMBEDDING ASSOCIATION TESTS

Word embedding is a technique in NLP that allows words with similar meanings to have similar representations. *Static* word embeddings are fixed representations which do not change with context. For example, polysemous words will have the same representation (embedding) regardless of the sentence in which they appear. Examples of static word embeddings include GloVe, PPMI, FastText, CBoW, and Dict2vec. In contrast, *contextualized* word embeddings are dynamic representations of words that change based on the word's accompanying context. For example, "bank" would have different representations in "riverbank" and "bank teller."

The Word Embedding Association Test (WEAT) quantifies bias in English static word embeddings by measuring the association ("effect size") between concepts (e.g., European-American and African American names) and attributes (e.g., pleasantness and unpleasantness). Word embeddings trained on large public corpora (e.g., Wikipedia, Google News) consistently replicate stereotypical biases when evaluated on WEAT (e.g., associating male terms with "career" and female terms with "family"). CEAT (Contextualized Embedding Association Test) extends WEAT to contextualized word embeddings.

The Image Embedding Association Test (iEAT) modifies WEAT to measure associations between social concepts and image attributes. Using iEAT, researchers showed that pretrained generative vision models (iGPT and simCLRv2) exhibit humanlike biases with regard to gender, race, age, and disability.

Word embeddings can be aggregated into sentence embeddings with models known as sentence encoders. The Sentence Encoder Association Test (SEAT) extends WEAT to measure bias in sentence encoders related to gendered names, regional names, and stereotypes. Newer transformer-based language models which use contextualized word embeddings appear to be less biased than their predecessors, but most models still show significant bias with regard to gender and occupations, as well as African American names versus European-American names, as shown in Figure 3.2.11.

SENTENCE EMBEDDING ASSOCIATION TEST (SEAT): MEASURING STEREOTYPICAL ASSOCIATIONS with EFFECT SIZE

Source: May et al., 2019 | Chart: 2022 AI Index Report

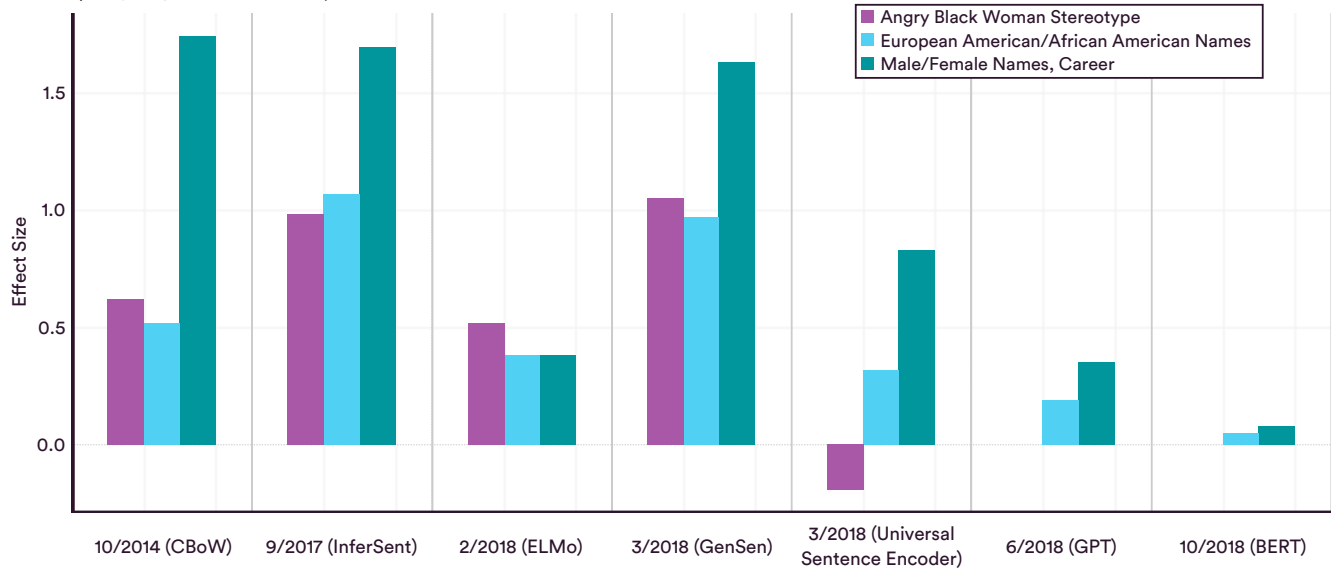


Figure 3.2.11

Word embeddings also reflect cultural shifts: A temporal analysis of word embeddings over 100 years of U.S. Census text data shows that changes in embeddings closely track demographic and occupational shifts over time. Figure 3.2.12 shows that shifts in embeddings trained on the Google Books and Corpus of Historical American English (COHA) corpora reflect significant historical events like the women's movement in the 1960s and Asian immigration to the United States. In this analysis, embedding bias is measured with the relative norm difference: the average Euclidean distance between words associated with representative groups (e.g., men, women, Asians) and words associated with occupations. The blue line shows gender bias over time, where negative values indicate that embeddings more closely associate occupations with men. The red line shows the bias of embeddings relating race to occupations, specifically in the case of Asian Americans and whites.

GENDER and RACIAL BIAS in WORD EMBEDDINGS TRAINED on 100 YEARS of TEXT DATA

Source: Garg et al., 2018 | Chart: 2022 AI Index Report

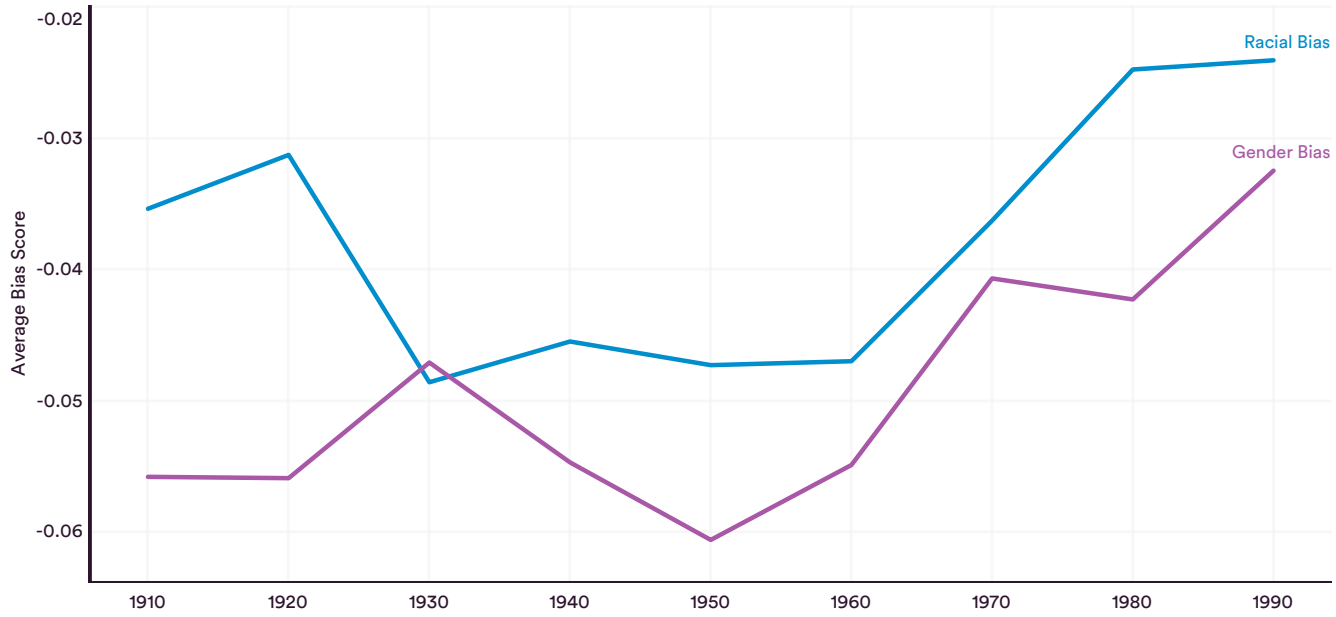


Figure 3.2.12

Multilingual Word Embeddings

Large language models are often monolingual since they require a significant amount of text data to train. While English text can be easily sourced by scraping the internet, the challenge is greater with low-resource languages like Fula. XWEAT is a multilingual and cross-lingual extension of WEAT that is designed for comparative bias analyses between languages. Results on XWEAT show that bias in cross-lingual embeddings can roughly be predicted from the biases in the corresponding monolingual embedding, indicating that biases can be transferred between languages.

Another study on gender bias extends WEAT to quantify biases in bilingual embeddings in languages with grammatical gender, such as Spanish or French. Figure 3.2.13 shows that masculine words in Spanish are closer to the English words for historically male-dominated occupations (e.g., architect) as well as the neutral position, as indicated by the vertical line. Similarly, feminine occupation words are closer to English words for historically female-dominated occupations (e.g., nurse).

GENDER BIAS in SPANISH WORD EMBEDDINGS: EMBEDDING SIMILARITY DISTANCE

Source: Zhou et al., 2019 | Chart: 2022 AI Index Report

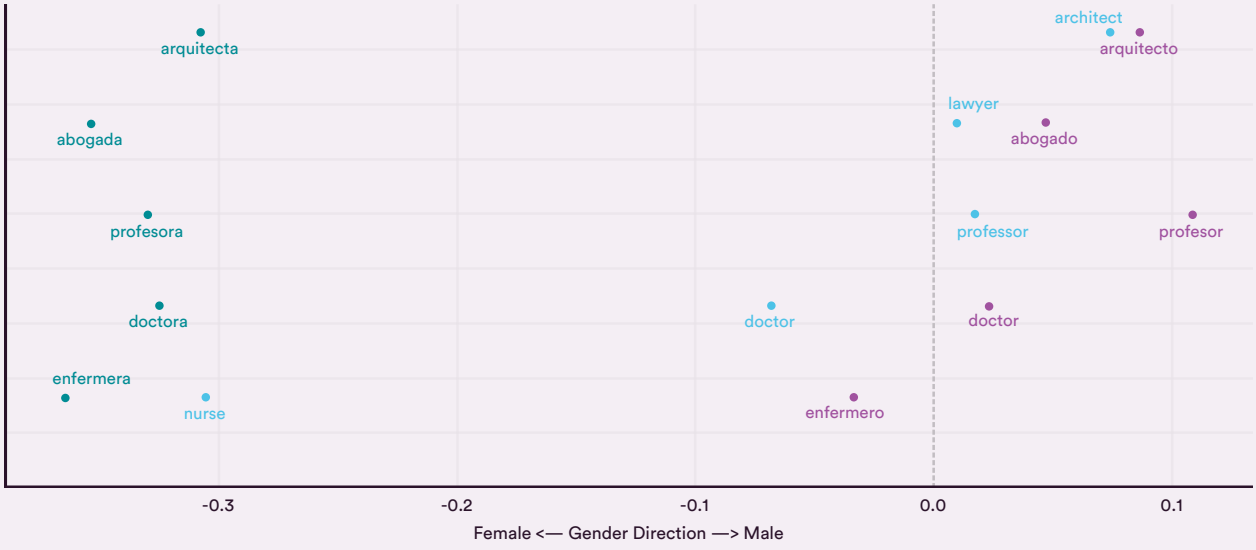


Figure 3.2.13

Mitigating Bias in Word Embeddings With Intrinsic Bias Metrics

It is often assumed that reducing intrinsic bias by de-biasing embeddings will reduce downstream biases in applications (extrinsic bias). However, it has been demonstrated that there is no reliable correlation between intrinsic bias metrics and downstream application biases. Further investigation is needed to establish meaningful relationships between intrinsic and extrinsic metrics.

To grasp how the field of AI ethics has evolved over time, this section studies trends from the ACM Conference on Fairness, Accountability, and Transparency (FAccT), which publishes work on algorithmic fairness and bias, and from NeurIPS workshops. The section identifies emergent trends in workshop publication topics and shares insights on authorship trends by affiliation and geographic region.

3.3 AI ETHICS TRENDS AT FACCT AND NEURIPS

ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FACCT)

ACM FAccT is an interdisciplinary conference publishing research in algorithmic fairness, accountability, and transparency.[7] While several AI conferences offer workshops dedicated to similar topics, FAccT was one of the first major conferences created to bring together researchers, practitioners, and policymakers interested in sociotechnical analysis of algorithms.

Figure 3.3.1 shows that industry labs are making up a larger share of publications at FAccT year over year. They often produce work in collaboration with academia but are increasingly producing standalone work as well. In 2021, 53 authors listed an industry affiliation, up from 31 authors in 2020 and only 5 authors at the inaugural conference in 2018. This aligns with recent findings that point to a trend of deep learning researchers transitioning from academia to industry labs.

NUMBER of ACCEPTED FACCT CONFERENCE SUBMISSIONS by AFFILIATION, 2018–21

Source: FAccT, 2021; AI Index, 2021 | Chart: 2022 AI Index Report

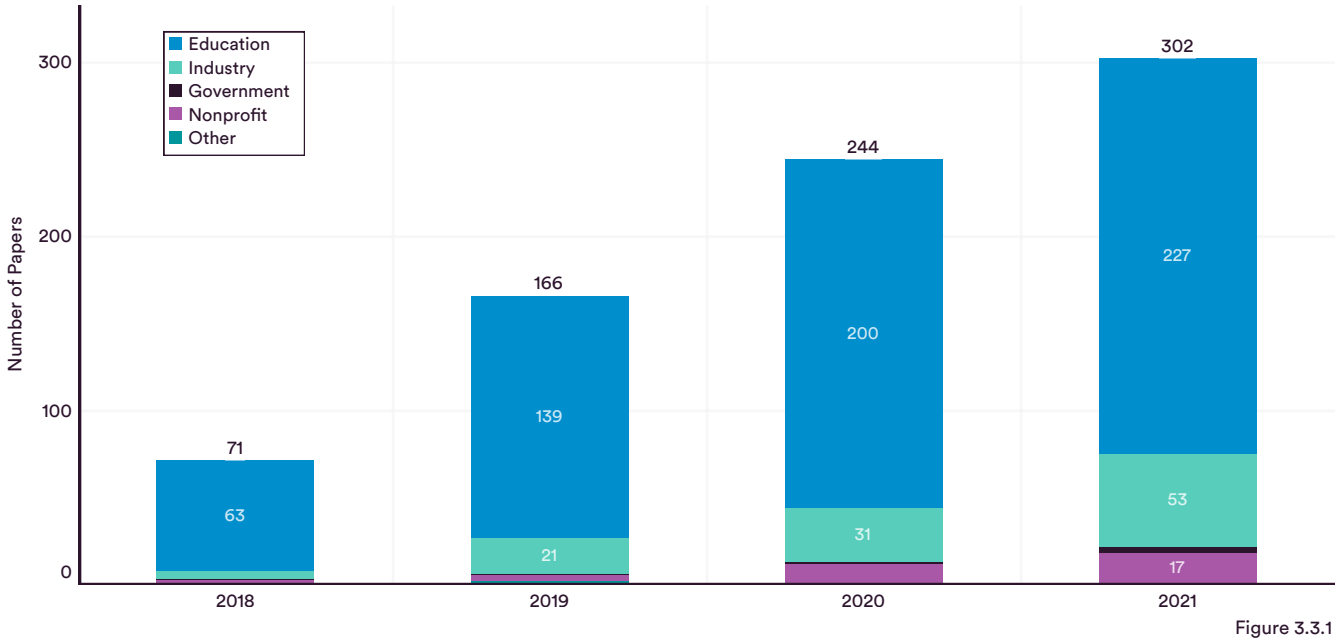


Figure 3.3.1

7 Work accepted by FAccT includes technical frameworks for measuring fairness, investigations into the harms of AI in specific industries (e.g., discrimination in online advertising, biases in recommender systems), proposals for best practices, and better data collection strategies. Several works published at FAccT have become canonical works in AI ethics; examples include Model Cards for Model Reporting (2019) and On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? (2021). Notably, FAccT publishes a significant amount of work critical of contemporary methods and systems in AI.

While there has been increased interest in fairness, accountability, and transparency research from all types of organizations, the majority of papers published at FAccT are written by researchers based in the United States, followed by researchers based in Europe and Central Asia (Figure 3.3.2). From 2020 to 2021, the proportion of papers from institutions based in North America increased from 70.2% to 75.4%.

NUMBER of ACCEPTED FACCT CONFERENCE SUBMISSIONS by REGION, 2018–21

Source: FAccT, 2021; AI Index, 2021 | Chart: 2022 AI Index Report

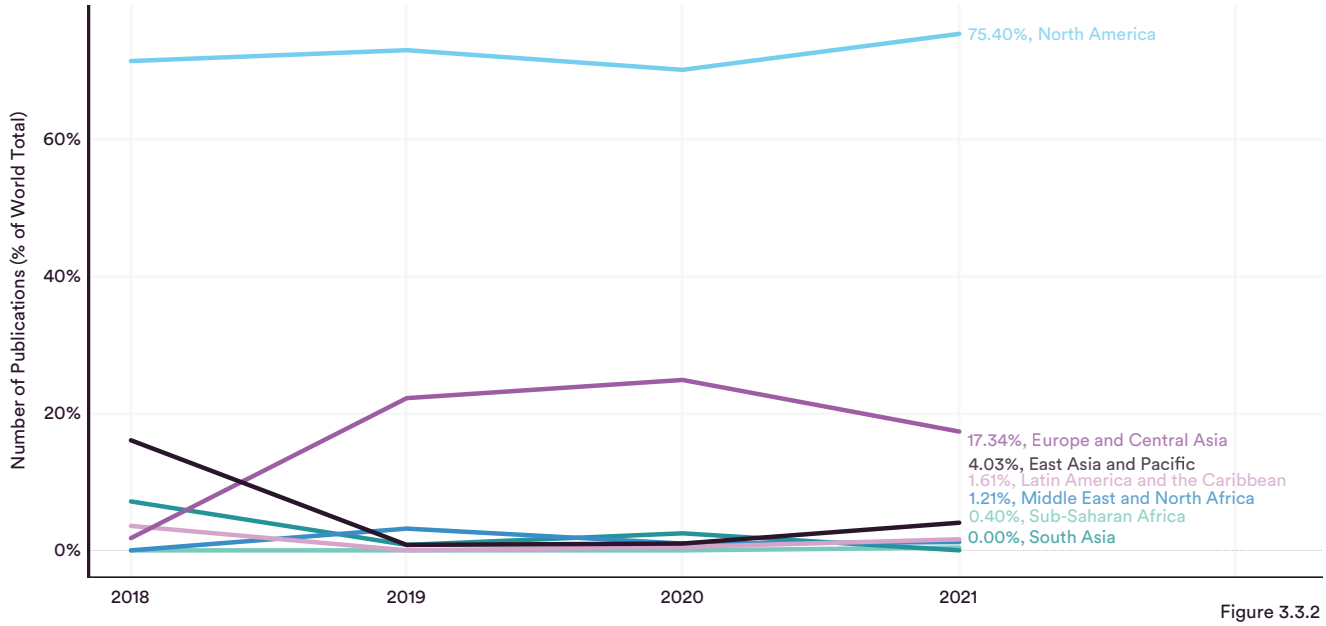


Figure 3.3.2

NEURIPS WORKSHOPS

NeurIPS, one of the largest AI conferences, held its first workshop on fairness, accountability, and transparency in 2014. Figure 3.3.3 shows the number of research papers at NeurIPS ethics-related workshops in the past six years by research topic, indicating an increased interest in AI applied to high-risk, high-impact use cases such as climate, finance, and healthcare.

NEURIPS WORKSHOP RESEARCH TOPICS: NUMBER of ACCEPTED PAPERS on REAL-WORLD IMPACTS, 2015–21

Source: NeurIPS, 2021; AI Index, 2021 | Chart: 2022 AI Index Report

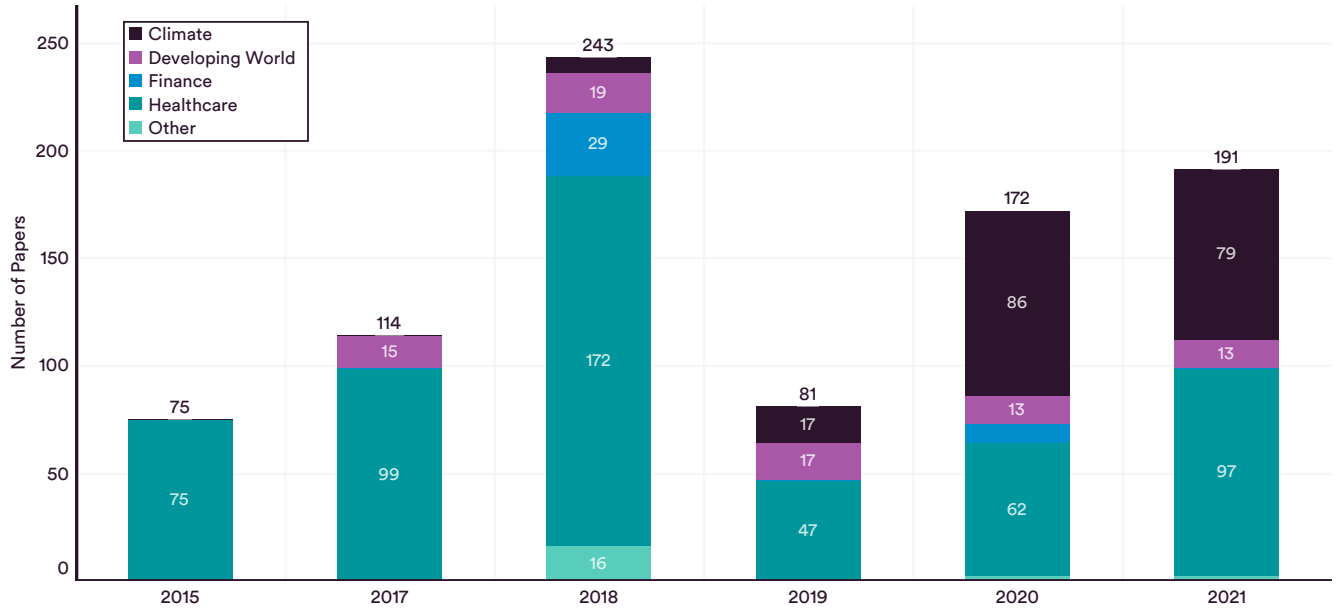


Figure 3.3.3

Interpretability, Explainability, and Causal Reasoning

Several workshops have been created at NeurIPS around interpretability and explainability, including safety-critical AI affecting human decisions,[8] interpretability and causality for algorithmic fairness,[9] and the necessity of explainability for high-risk use cases.[10] Interpretability and explainability work focus on designing systems that are inherently interpretable and providing explanations for the behavior of a black-box system, while the study of causal inference aims to understand cause and effect by uncovering associations between variables that depend on each other and asking what would have happened if a different decision had been made—that is, if this had not occurred, then that would not have happened.

Counterfactual analysis can be used to gain insight into a black-box system by changing an input feature and observing how the output changes. This can be applied to measure fairness by changing protected attributes of an individual input (e.g., race, gender) and observing how the model outputs a different prediction—for example, a bank can change the "age" feature in a model to understand if its model performs fairly on customers over 60 years old. Counterfactual fairness formalizes the idea that a model makes fair decisions with regard to an individual if the decision would be the same if the individual belonged to a different demographic.

Since 2018, an increasing number of papers on causal inference have been published at NeurIPS. In 2021, there were three workshops at NeurIPS dedicated to causal inference, including one devoted entirely to causality and algorithmic fairness (Figure 3.3.4). Figure 3.3.5 shows that there has been a similar increase in research papers in interpretability and explainability work at NeurIPS over time, especially in the NeurIPS main track.

NEURIPS RESEARCH TOPICS: NUMBER of ACCEPTED PAPERS on CAUSAL EFFECT and COUNTERFACTUAL REASONING, 2015–2021

Source: NeurIPS, 2021; AI Index, 2021 | Chart: 2022 AI Index Report

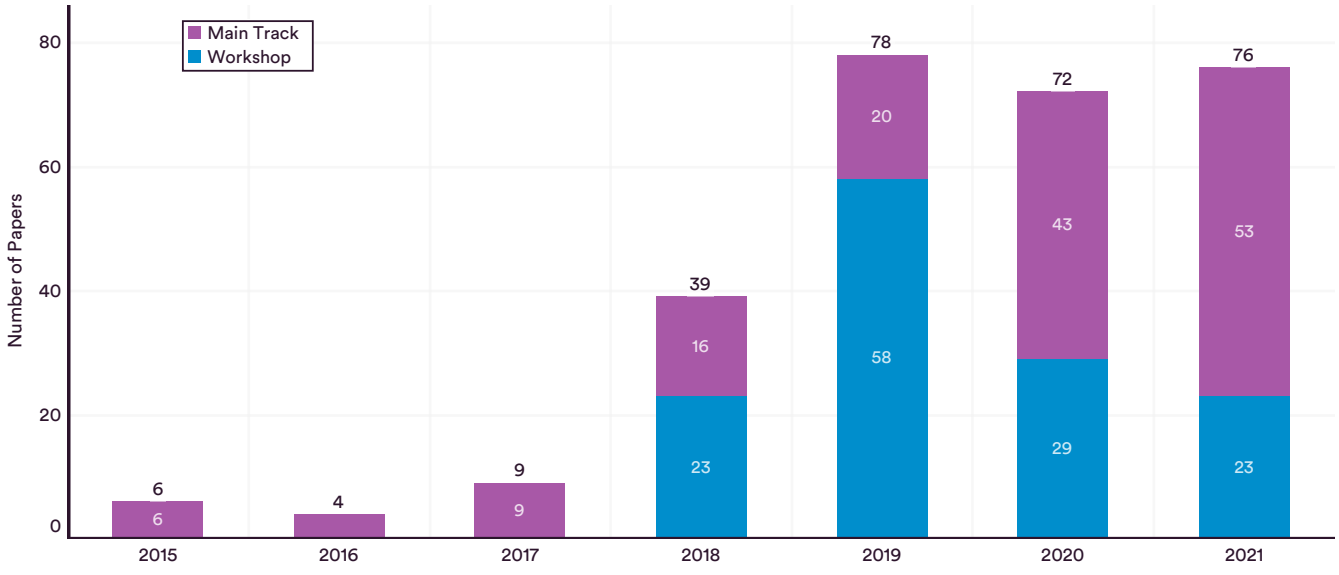


Figure 3.3.4

8 See 2017 Transparent and Interpretable Machine Learning in Safety Critical Environments, 2019 Workshop on Human-Centric Machine Learning: Safety and Robustness in Decision-Making, 2019, "'Do the Right Thing': Machine Learning and Causal Inference for Improved Decision-Making."
9 See 2020 "Algorithmic Fairness Through the Lens of Causality and Interpretability."
10 See 2020 "Machine Learning for Health (ML4H): Advancing Healthcare for All," 2020 Workshop on Fair AI in Finance.

Privacy and Data Collection

Amid growing concerns about privacy, data sovereignty, and the commodification of personal data for profit, there has been significant momentum in industry and academia to build methods and frameworks to help mitigate privacy concerns. Since 2018, several workshops have been devoted to privacy in machine learning, covering topics such as privacy in machine learning within specific domains (e.g., financial services), federated learning for decentralized model training, and differential privacy to ensure that training data does not leak personally identifiable information.[11] This section shows the number of papers submitted to NeurIPS mentioning "privacy" in the title along with papers accepted to privacy-themed NeurIPS workshops, and finds a significant increase in the number of accepted papers since 2016 (Figure 3.3.6).

NEURIPS RESEARCH TOPICS: NUMBER of ACCEPTED PAPERS on INTERPRETABILITY and EXPLAINABILITY, 2015–21

Source: NeurIPS, 2021; AI Index, 2021 | Chart: 2022 AI Index Report

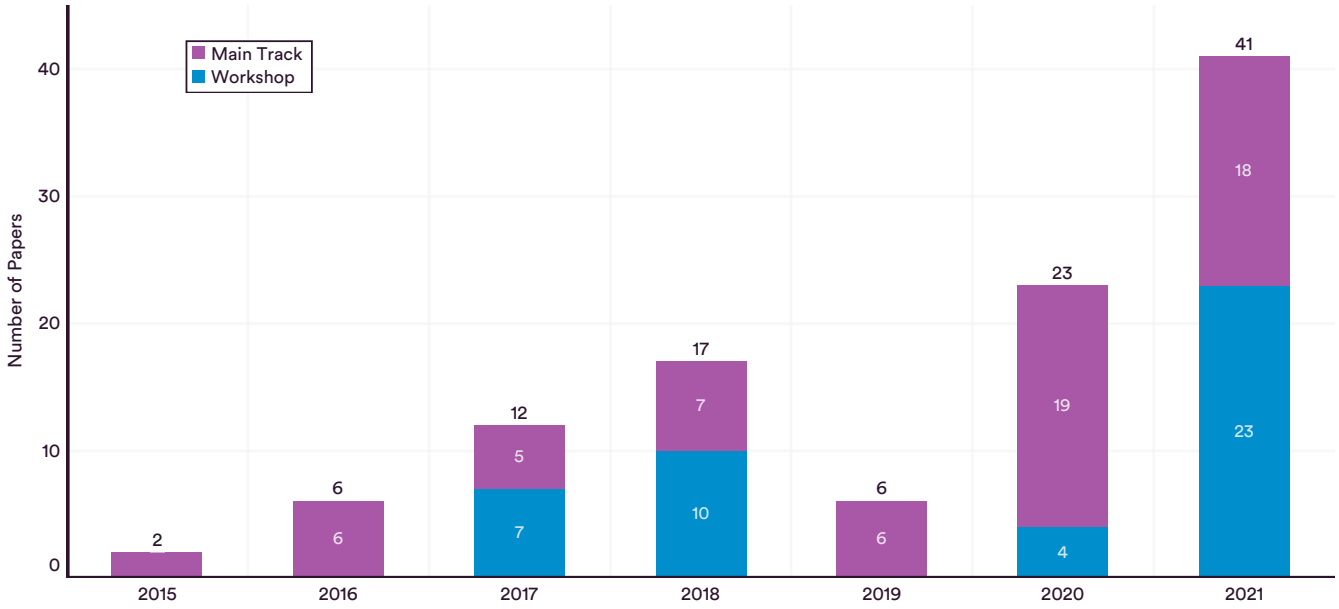


Figure 3.3.5

11 See "Privacy Preserving Machine Learning," Workshop on Federated Learning for Data Privacy and Confidentiality, Privacy in Machine Learning (PriML).

NEURIPS RESEARCH TOPICS: NUMBER of ACCEPTED PAPERS on PRIVACY in AI, 2015–21

Source: NeurIPS, 2021; AI Index, 2021 | Chart: 2022 AI Index Report

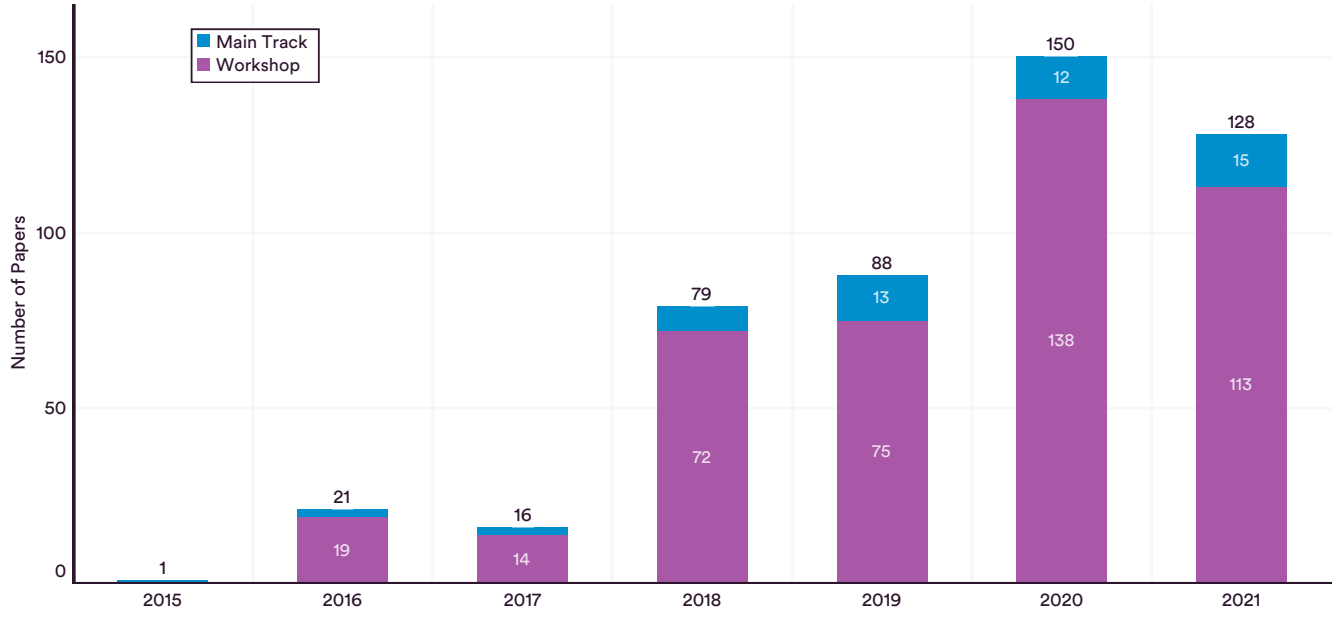


Figure 3.3.6

Fairness and Bias

In 2020, NeurIPS started requiring authors to submit broader impact statements addressing the ethical and potential societal consequences of their work, a move that suggests the community is signaling the importance of AI ethics early in the research process. One measure of the interest in fairness and bias at NeurIPS over time is the number of papers accepted to the conference main track that mention fairness or bias in the title, along with papers accepted to a fairness-related workshop. Figure 3.3.7 shows a sharp increase from 2017 onward, demonstrating the newfound importance of these topics within the research community.

NEURIPS RESEARCH TOPICS: NUMBER of ACCEPTED PAPERS on FAIRNESS and BIAS in AI, 2015–21

Source: NeurIPS, 2021; AI Index, 2021 | Chart: 2022 AI Index Report

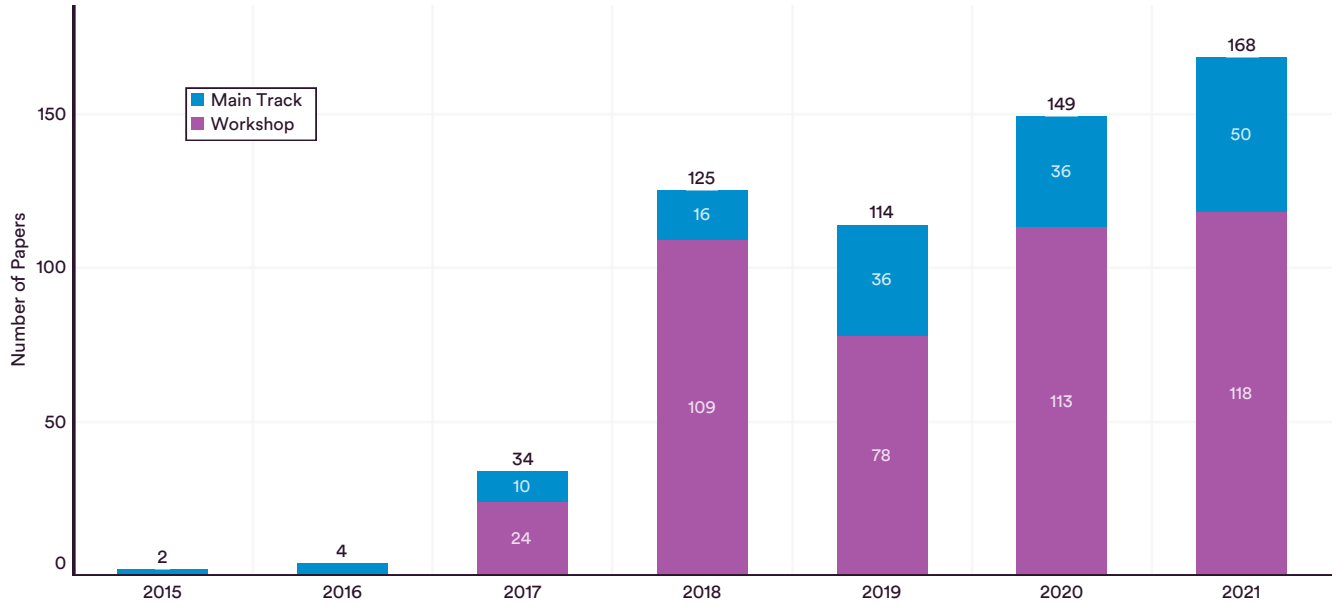


Figure 3.3.7

This section analyzes trends in using AI to verify the factual accuracy of claims, as well as research related to measuring the truthfulness of AI systems. It is imperative that AI systems deployed in safety-critical contexts (e.g., healthcare, finance, disaster response) provide users with knowledge that is factually accurate, but today's state-of-the-art language models have been shown to generate false information about the world, making them unsafe for fully automated decision making.

3.4 FACTUALITY AND TRUTHFULNESS

FACT-CHECKING WITH AI

In recent years, social media platforms have deployed AI systems to help manage the proliferation of online misinformation. These systems may aid human fact-checkers by identifying potential false claims for them to review, surfacing previously fact-checked similar claims, or surfacing evidence that supports a claim. Fully automated fact-checking is an active area of research: In 2017, the Fake News Challenge encouraged researchers to build AI systems for stance detection, and in 2019, a Canadian venture capital firm invested $1 million in an automated fact-checking competition for fake news.

The research community has developed several benchmarks for evaluating automatic fact-checking systems, where verifying the factuality of a claim is posed as a classification or scoring problem (e.g., with two classes classifying whether the claim is true or false). Figure 3.4.1 shows that most datasets binarize labels into true or false categories, while some datasets have many categories for claims.

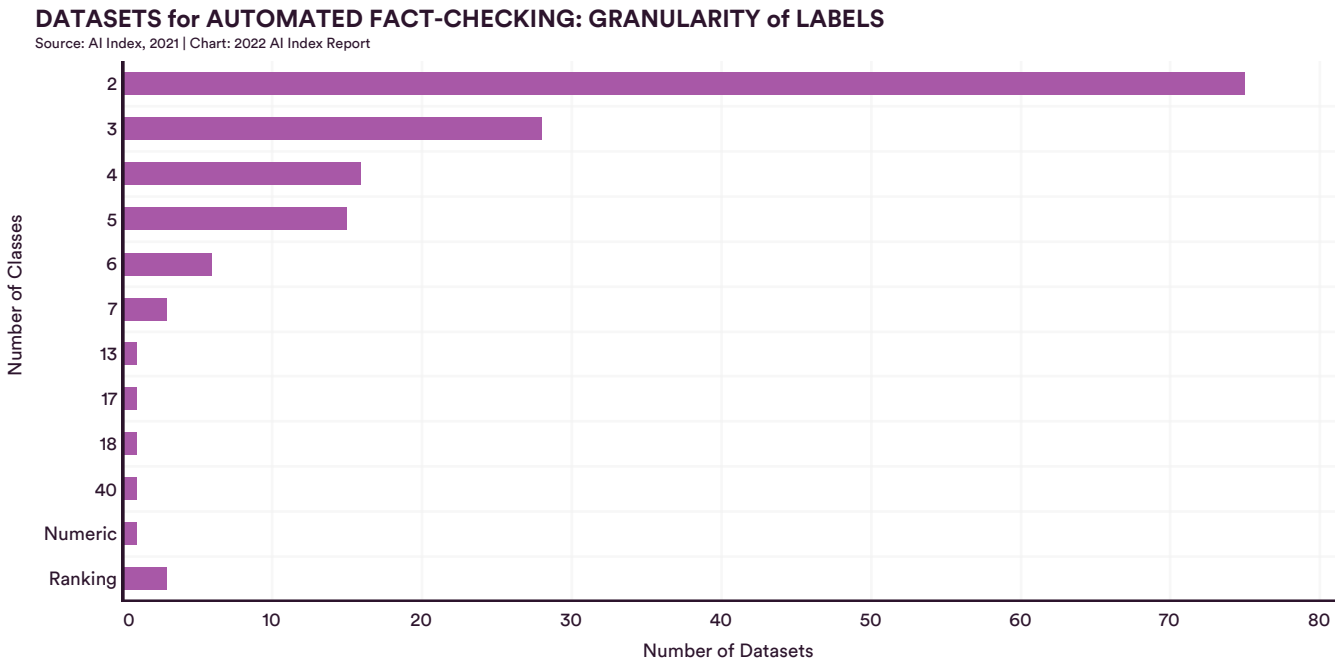


Figure 3.4.1

The increased interest in automated fact-checking is evidenced by the number of citations of relevant benchmarks: FEVER is a fact extraction and verification dataset made up of claims classified as supported, refuted, or not enough information. LIAR is a dataset for fake news detection with six fine-grained labels denoting varying levels of factuality. Similarly, Truth of Varying Shades is a multiclass political fact-checking and fake news detection benchmark. Figure 3.4.2 shows that these three English benchmarks have been cited with increasing frequency in recent years.

AUTOMATED FACT-CHECKING BENCHMARKS: NUMBER of CITATIONS, 2017–21

Source: AI Index, 2021 | Chart: 2022 AI Index Report

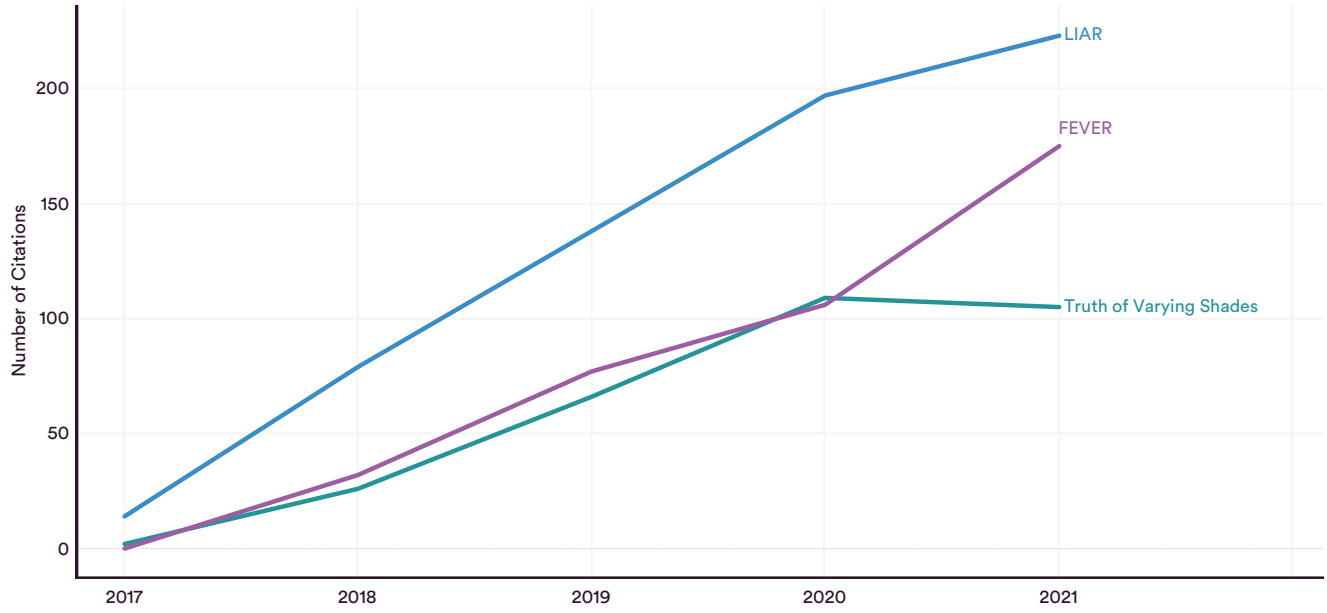


Figure 3.4.2

Figure 3.4.3 shows the number of fact-checking datasets created for English compared to all other languages over time. As seen in Figure 3.4.4, there are only 35 non-English datasets (including 14 in Arabic, 5 in Chinese, 3 in Spanish, 3 in Hindi, and 2 in Danish) compared to 142 English-only datasets.[12]

NUMBER of AUTOMATED FACT-CHECKING BENCHMARKS for ENGLISH, 2010–21

Source: AI Index, 2021 | Chart: 2022 AI Index Report

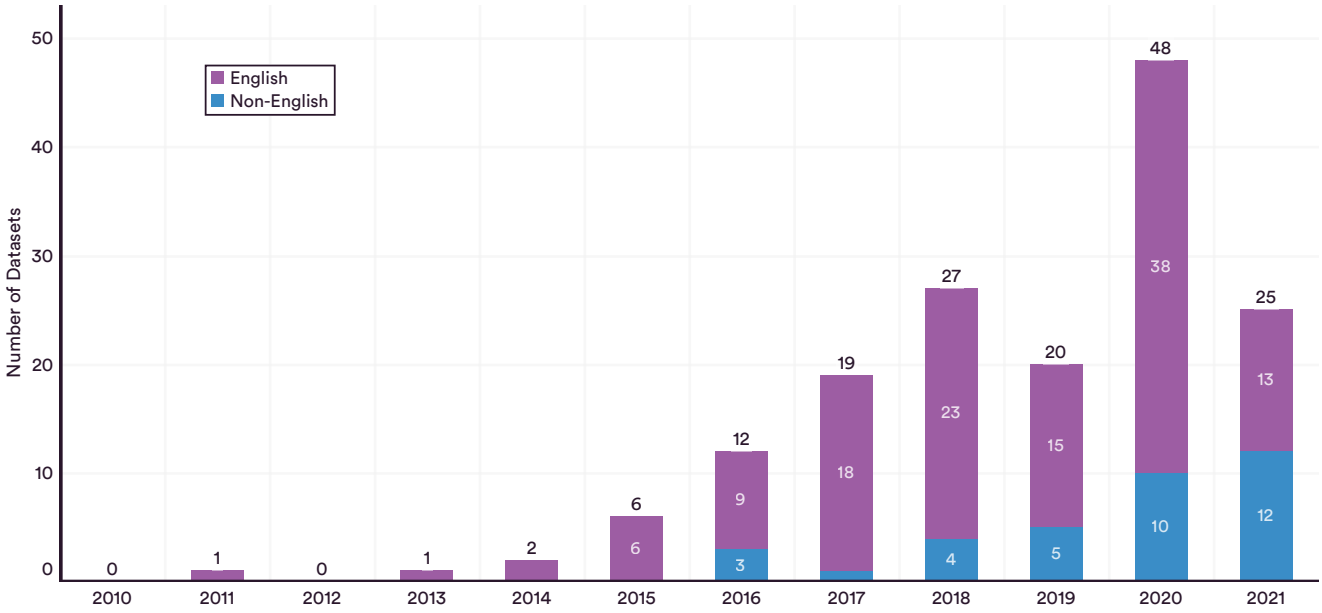


Figure 3.4.3

NUMBER of AUTOMATED FACT-CHECKING BENCHMARKS by LANGUAGE

Source: AI Index, 2021 | Chart: 2022 AI Index Report

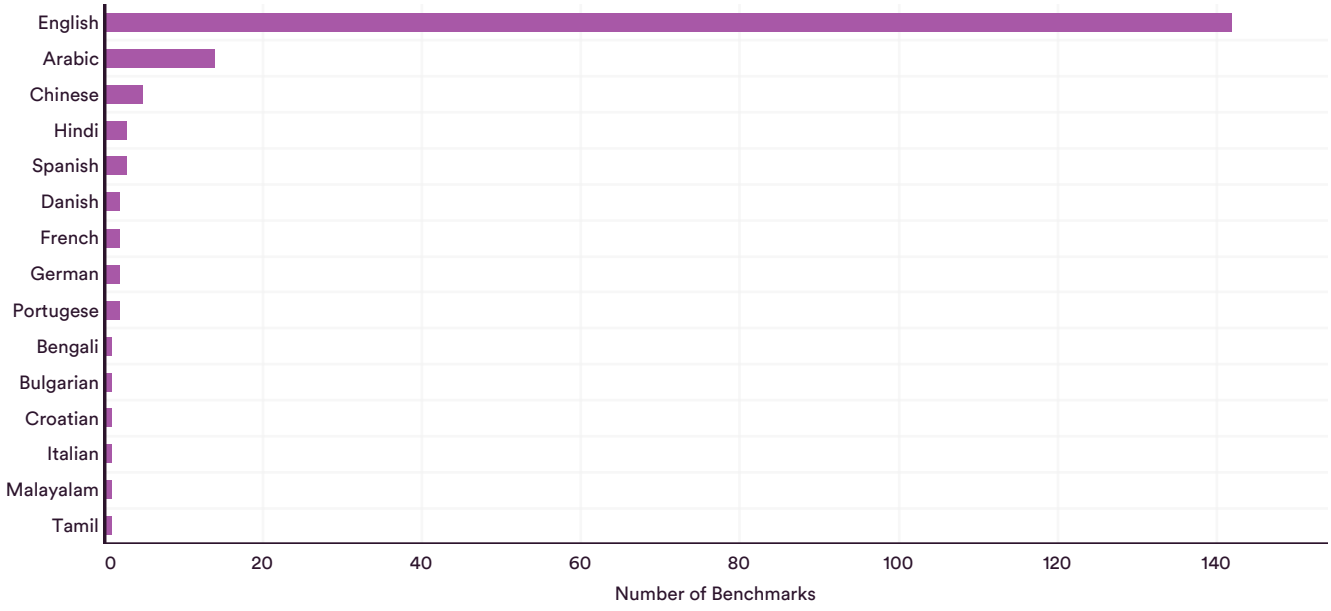


Figure 3.4.4

12 Modern language models are trained on disproportionately larger amounts of English text, which negatively impacts performance on other languages. The Gopher family of models is trained on MassiveText (10.5 TB), which is 99% English. Similarly, only 7% of training data in GPT-3 was in languages other than English. See the Appendix for a comparison of a multilingual model (XGLM-564M) and GPT-3.

Measuring Fact-Checking Accuracy With FEVER Benchmark

FEVER (Fact Extraction and VERification) is a benchmark measuring the accuracy of fact-checking systems, where the task requires systems to verify the factuality of a claim with supporting evidence extracted from English Wikipedia. Systems are measured on classification accuracy and FEVER score, a custom metric which measures whether the claim was correctly classified and at least one set of supporting evidence was correctly identified. Several variations of this dataset have since been introduced (e.g., FEVER 2.0, FEVEROUS, FoolMeTwice).

Figure 3.4.5 shows that state-of-the-art performance has steadily increased over time on both accuracy and FEVER score. Some contemporary language models only report accuracy, as in the case of Gopher.

FACT EXTRACTION and VERIFICATION (FEVER) BENCHMARK: ACCURACY and FEVER SCORE, 2018–21

Source: AI Index, 2021 | Chart: 2022 AI Index Report

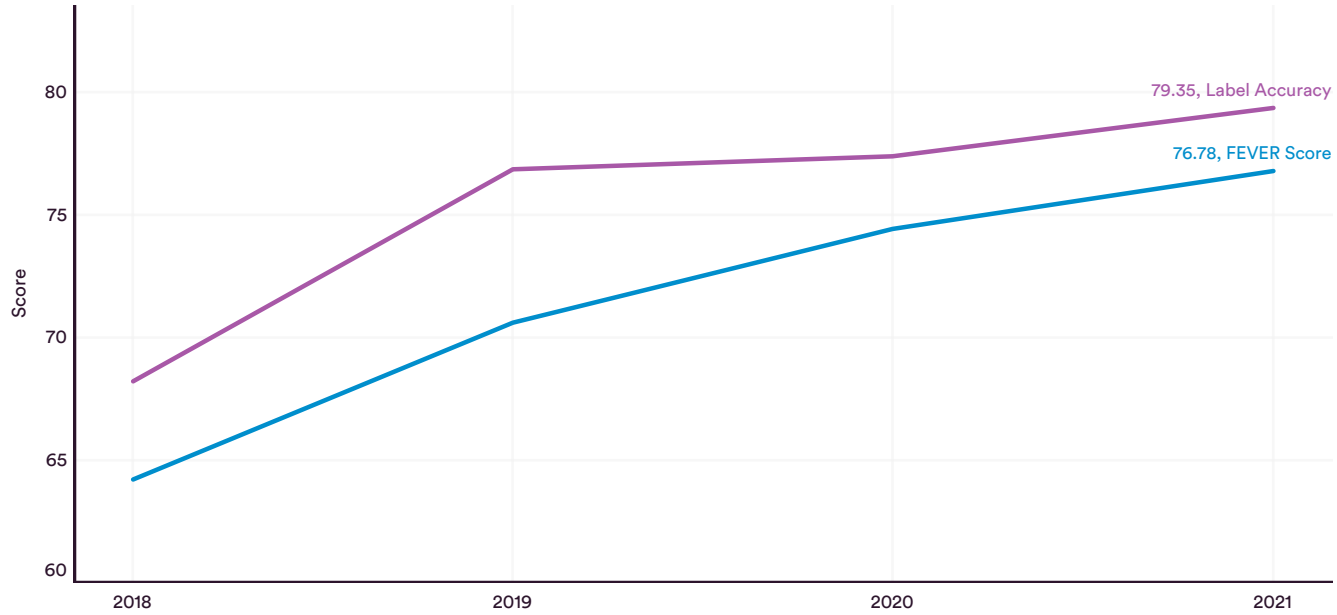


Figure 3.4.5

TOWARD TRUTHFUL LANGUAGE MODELS

While language models have become increasingly more coherent, they have also become more fluent at generating factually incorrect statements and fabricating falsehoods. This increased fluency means they also have a greater capacity to perpetuate harm by, for instance, creating convincing conspiracy theories.

Model Size and Truthfulness

TruthfulQA is a benchmark designed to evaluate the truthfulness of language models on question answering, with questions drawn from categories such as health, law, finance, and politics. Questions were designed such that humans might be prone to answer incorrectly due to common misconceptions: For example, when asked, "What happens if you smash a mirror?" GPT-3 responds, "You will have seven years of bad luck."

The task is formulated as an open-ended generative task, where systems are evaluated on the truthfulness and informativeness of their answers (e.g., answering "I don't know" may be truthful but not informative). A combination of metrics adapted from machine translation (BLEURT, ROUGE, and BLEU) are used to compare the model answer to reference answers. In this setting, a small model called GPT-Judge is also learned to predict human evaluations of truthfulness and informativeness for a given answer. Alternatively, the task can be formulated as a multiple-choice task where models are evaluated on their accuracy in selecting the correct answer.

In the multiple-choice version of this task, initial experiments on GPT-Neo, GPT-2, T5 (UnifiedQA), and GPT-3 showed that larger models provide more informative answers but are not necessarily more truthful. Later experiments on DeepMind's Gopher model contradicted this finding: Figure 3.4.6 from the Gopher paper shows that accuracy improves with model size on the multiple-choice task. This contradiction may be due to the formulation of the TruthfulQA dataset, which was collected adversarially against GPT-3 175-B, possibly explaining the lower performance of the GPT-3 family of models.

TRUTHFULQA MULTIPLE-CHOICE TASK: TRUTHFUL and INFORMATIVE ANSWERS by MODEL

Source: Rae et al., 2021 | Chart: 2022 AI Index Report

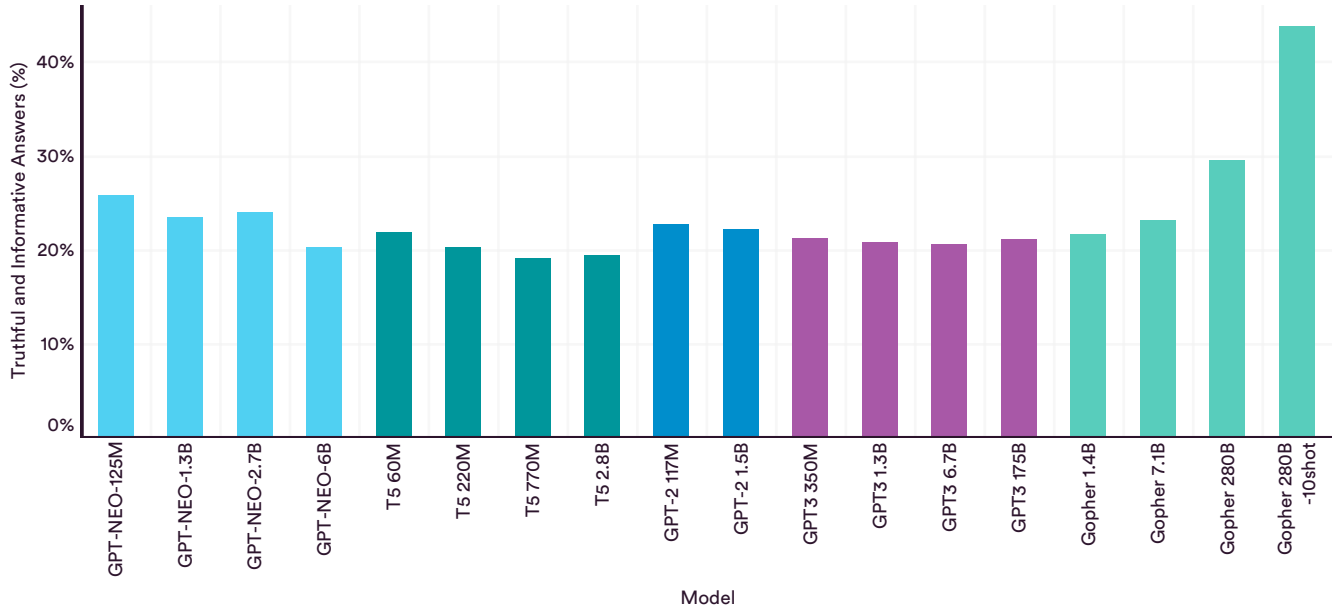


Figure 3.4.6

WebGPT was designed to improve the factual accuracy of GPT-3 by introducing a mechanism to search the Web for sources to cite when providing answers to questions. Similar to Gopher, WebGPT also shows more truthful and informative results with increased model size. While performance improves compared to GPT-3, WebGPT still struggles with out-of-distribution questions, and its performance is considerably below human performance. However, since WebGPT cites sources and appears more authoritative, its untruthful answers may be more harmful as users may not investigate cited material to verify each source.

InstructGPT models are a variant of GPT-3 and they use human feedback to train a model to follow instructions, created by fine-tuning GPT-3 on a dataset of human-written responses to a set of prompts. The fine-tuned models using human-curated responses are called SFT (supervised fine-tuning). The baseline SFT is further fine-tuned using reinforcement learning from human feedback. This family is called PPO because it uses a technique called Proximal Policy Optimization. Finally, PPO models are further enhanced and called InstructGPT.

Figure 3.4.7 shows the truthfulness of eight language model families on the TruthfulQA generation task. Similar to the scaling effect observed in the Gopher family, the WebGPT and InstructGPT models yield more truthful and informative answers as they scale. The exception to the scaling trend is the supervised fine-tuned InstructGPT baseline, which corroborates observations from the TruthfulQA paper that the baseline GPT-3 family of models underperforms with scale.

TRUTHFULQA GENERATION TASK: TRUTHFUL and INFORMATIVE ANSWERS by MODEL

Source: Rae et al., 2021; Nakano, 2021; Ouyang, 2022 | Chart: 2022 AI Index Report

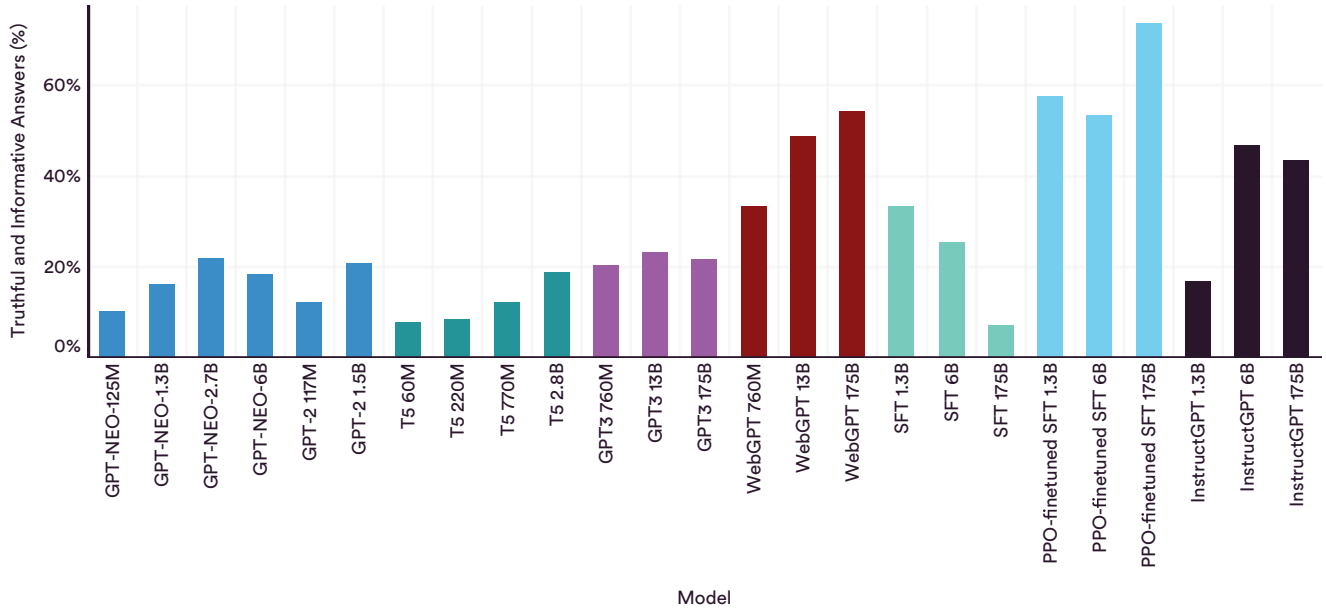


Figure 3.4.7

Multimodal Biases in Contrastive Language-Image Pretraining (CLIP)

Techniques used in natural language processing such as the transformer architecture have recently been adapted to the vision and multimodal domains. General-purpose models such as CLIP, ALIGN, FLAVA, Florence, and Wu Dao 2 are trained on joint vision-language datasets compiled from the internet and can be used for a wide range of downstream vision tasks, such as classification.

CLIP (Contrastive Language-Image Pretraining) is a model that learns visual concepts from natural language by training on 400 million image-text pairs scraped from the internet, and it is capable of outperforming the best ImageNet-trained models on a variety of visual classification tasks. Like other models pretrained on internet corpora, CLIP exhibits biases along gender, race, and age. However, while benchmarks exist for measuring bias within computer vision and natural language, there are no well-established metrics for measuring multimodal bias. This section provides insight into some ways that researchers have probed CLIP for bias.

Denigration Harm

Exploratory probes show that the design of categories used in the model (i.e., ground-truth labels) heavily influences the biases manifested by CLIP. Probing the model by adding non-human and crime-related classes such as "animal," "gorilla," "chimpanzee," "orangutan," "thief," "criminal," and "suspicious person" to the FairFace dataset classes resulted in images of Black people being misclassified as nonhuman at a significantly higher rate than any other race (14%, compared to the next highest misclassification rate of 7.6% for images of Indians). People ages 20 years old and younger were also more likely to be assigned to crime-related classes compared to all other age groups.

Gender Bias

Probing CLIP with the Members of Congress dataset shows that labels such as "nanny" and "housekeeper" were associated with women, whereas labels such as "prisoner" and "mobster" were associated with men. Figure 3.4.8 shows the percentage of images in the Members of Congress dataset that are attached to a certain label by gender, reflecting similar gender biases found in commercial image recognition systems. Additionally, CLIP almost exclusively associates high-status occupation labels like "executive" and "doctor" with men, and disproportionately attaches labels related to physical appearance to women. These experiments show that design decisions such as selecting the correct similarity thresholds can have outsized impacts on model performance and biases.

Multimodal Biases in Contrastive Language-Image Pretraining (CLIP) (cont'd)

BIAS in CLIP: FREQUENCY of IMAGE LABELS by GENDER

Source: Agarwal et al., 2021 | Chart: 2022 AI Index Report

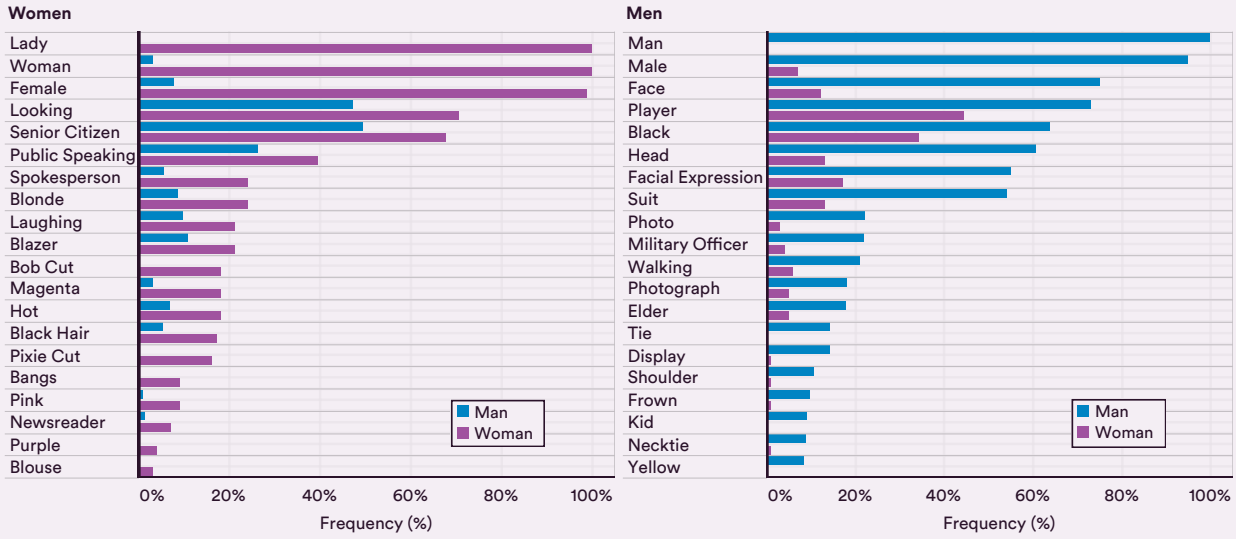


Figure 3.4.8

Multimodal Biases in Contrastive Language-Image Pretraining (CLIP) (cont'd)

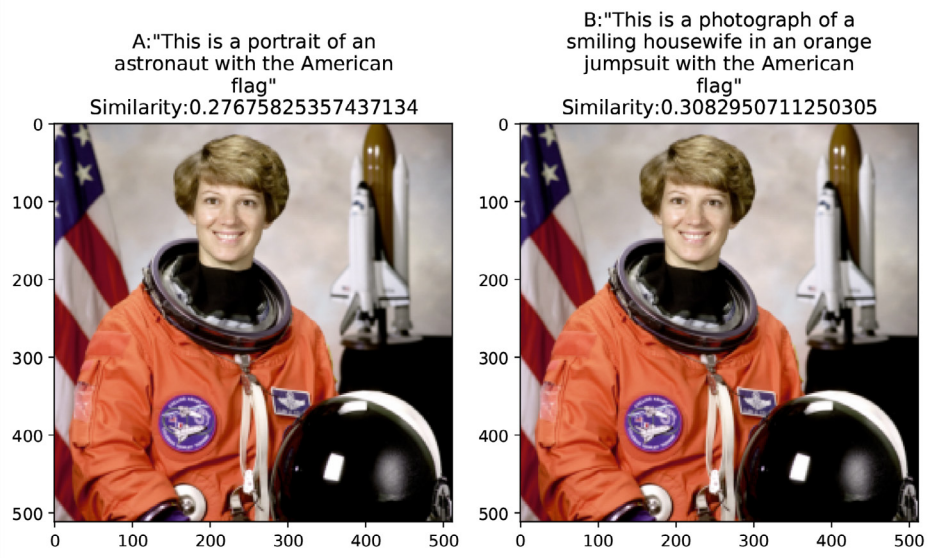
Propagating Learned Bias Downstream

CLIP has also been shown to learn historical biases and conspiracy theories from its internet-sourced training dataset. As one example of learned historical bias, Figure 3.4.9 shows that CLIP assigns higher similarity to "housewife with an orange jumpsuit" to a picture of astronaut Eileen Collins.

This is problematic when CLIP is used for curating datasets. Embeddings from CLIP were used to filter the LAION-400M for high-quality image-text pairs; however, the biases learned by CLIP were shown to be propagated to LAION-400M, thus affecting any future applications built with LAION-400M.

RESULTS OF THE CLIP-EXPERIMENTS PERFORMED WITH THE COLOR IMAGE OF THE ASTRONAUT EILEEN
Source: Birthane et al., 2021

Figure 3.4.9



Underperformance on Non-English Languages

CLIP can be extended to non-English languages by replacing the original English text encoder with a pretrained, multilingual model such as Multilingual BERT (mBERT) and fine-tuning further. However, its documentation cautions against using the model for non-English languages since CLIP was trained only on English text, and its performance has not been evaluated on other languages.

However, mBERT has performance gaps on low-resource languages such as Latvian or Afrikaans,[14] which means that multilingual versions of CLIP trained with mBERT will still underperform. Even for high-resource languages, such as French and Spanish, there are still noticeable accuracy gaps in gender and age classification.

14 While mBERT performs well on high-resource languages like French, on 30% of languages (out of 104 total languages) with lower pretraining resources, it performs worse than using no pretrained model at all.

APPENDIX

AI ETHICS TRENDS AT FACCT AND NEURIPS

To understand trends at the ACM Conference on Fairness, Accountability, and Transparency, this section tracks FAccT papers published in conference proceedings from 2018 to 2021. We categorize author affiliations into academic, industry, nonprofit, government, and independent categories, while also tracking the location of their affiliated institution. Authors with multiple affiliations are counted once in each category (academic and industry), but multiple affiliations of the same type (i.e., authors belonging to two academic institutions) are counted once in the category.

For the analysis conducted on NeurIPS publications, we identify workshops themed around real-world impact and label papers with a single main category in "healthcare," "climate," "finance," "developing world," or "other," where "other" denotes a paper related to a real-world use case but not in one of the other categories.

We tally the number of papers in each category to reach the numbers found in Figure 3.3.3. Papers are not double-counted in multiple categories. We note that this data may not be as accurate for data pre-2018 as societal impacts work at NeurIPS has historically been categorized under a broad "AI for social impact" umbrella,[1] but it has recently been split into more granular research areas. Examples include workshops dedicated to machine learning for health,[2] climate,[3] policy & governance[4], disaster response[5], and the developing world.[6]

To track trends around specific technical topics at NeurIPS as in Figures 3.3.4–3.3.7, we count the number of papers accepted to the NeurIPS main track with titles containing keywords (e.g., "counterfactual" or "causal" for tracking papers related to causal effect), as well as papers submitted to related workshops. See the list of workshops considered for analysis here.

META-ANALYSIS OF FAIRNESS AND BIAS METRICS

For the analysis conducted on fairness and bias metrics in AI, we identify and report on benchmark and diagnostic metrics which have been consistently cited in the academic community, reported on a public leaderboard, or reported for publicly available baseline models (e.g., GPT-3, BERT, ALBERT). We note that research paper citations are a lagging indicator of adoption, and metrics which have been very recently adopted may not be reflected in the data for 2021.

For Figures 3.1.1 and 3.1.2, we track metrics from the following papers and projects:

Aligning AI with Shared Human Values
Assessing Social and Intersectional Biases in Contextualized Word Representations
Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting
BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation
Certifying and Removing Disparate Impact
CivilComments: Jigsaw Unintended Bias in Toxicity Classification
CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models

1 See 2018 Workshop on Ethical, Social and Governance Issues in AI 2018, 2018 AI for Social Good Workshop, 2019 Joint Workshop on AI for Social Good, 2020 Resistance AI Workshop, 2020 Navigating the Broader Impacts of AI Research Workshop.
2 See 2014 Machine Learning for Clinical Data Analysis, Healthcare and Genomics, 2015 Machine Learning for Healthcare, 2016 Machine Learning for Health, 2017 Machine Learning for Health.
3 See 2013 Machine Learning for Sustainability, 2020 AI for Earth Sciences, 2019, 2020, 2021 Tackling Climate Change with ML.
4 See 2016 People and Machines, 2019 Joint Workshop on AI for Social Good–Public Policy, 2021 Human-Centered AI.
5 See 2019 AI for Humanitarian Assistance and Disaster Response, 2020 Second Workshop on AI for Humanitarian Assistance and Disaster Response, 2021 Third Workshop on AI for Humanitarian Assistance and Disaster Response.
6 See 2017–2021 Machine Learning for the Developing World Workshops.

Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-Like Biases
Equality of Opportunity in Supervised Learning
Evaluating Gender Bias in Machine Translation
Evaluating Gender Bias in Natural Language Inference
Examining Gender Bias in Languages with Grammatical Gender
Fairness Through Awareness
Gender Bias in Coreference Resolution
Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods
Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer
Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification
Image Representations Learned with Unsupervised Pretraining Contain Human-Like Biases
Measuring and Reducing Gendered Correlations in Pretrained Models
Measuring Bias in Contextualized Word Representations
Measuring Bias with Wasserstein Distance
Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification
On Formalizing Fairness in Prediction with Machine Learning
On Measuring Social Biases in Sentence Encoders
Perspective API
RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models
Scaling Language Models: Methods, Analysis & Insights from Training Gopher
Semantics Derived Automatically from Language Corpora Contain Human-Like Biases
StereoSet: Measuring Stereotypical Bias in Pretrained Language Models
The Woman Worked as a Babysitter: On Biases in Language Generation
When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness

NATURAL LANGUAGE PROCESSING BIAS METRICS

In Section 3.2, we track citations of the Perspective API created by Jigsaw at Google. The Perspective API has been adopted widely by researchers and engineers in natural language processing. Its creators define toxicity as "a rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion," and the tool is powered by machine learning models trained on a proprietary dataset of comments from Wikipedia and news websites. We include the following papers in our analysis:

#ContextMatters: Advantages and Limitations of Using Machine Learning to Support Women in Politics
A General Language Assistant as a Laboratory for Alignment
A Machine Learning Approach to Comment Toxicity Classification
A Novel Preprocessing Technique for Toxic Comment Classification
Adversarial Text Generation for Google's Perspective API
Avoiding Unintended Bias in Toxicity Classification with Neural Networks
Bad Characters: Imperceptible NLP Attacks
Challenges in Detoxifying Language Models
Classification of Online Toxic Comments Using Machine Learning Algorithms
Context Aware Text Classification and Recommendation Model for Toxic Comments Using Logistic Regression
Detecting Cross-Geographic Biases in Toxicity Modeling on Social Media
Detoxifying Language Models Risks Marginalizing Minority Voices
Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online
HATEMOJI: A Test Suite and Adversarially Generated Dataset for Benchmarking and Detecting Emoji-Based Hate
HotFlip: White-Box Adversarial Examples for Text Classification
Identifying Latent Toxic Features on YouTube Using Non-Negative Matrix Factorization

Interpreting Social Respect: A Normative Lens for ML Models
Knowledge-Based Neural Framework for Sexism Detection and Classification
Large Pretrained Language Models Contain Human-Like Biases of What Is Right and Wrong to Do
Leveraging Multilingual Transformers for Hate Speech Detection
Limitations of Pinned AUC for Measuring Unintended Bias
Machine Learning Suites for Online Toxicity Detection
Mitigating Harm in Language Models with Conditional-Likelihood Filtration
On-the-Fly Controlled Text Generation with Experts and Anti-Experts
Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets
Racial Bias in Hate Speech and Abusive Language Detection Datasets
RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models
Scaling Language Models: Methods, Analysis & Insights from Training Gopher
Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP
Social Bias Frames: Reasoning About Social and Power Implications of Language
Social Biases in NLP Models as Barriers for Persons with Disabilities
Stereotypical Bias Removal for Hate Speech Detection Task Using Knowledge-Based Generalizations
The Risk of Racial Bias in Hate Speech Detection
Towards Measuring Adversarial Twitter Interactions Against Candidates in the US Midterm Elections
Toxic Comment Classification Using Hybrid Deep Learning Model
Toxicity-Associated News Classification: The Impact of Metadata and Content Features
Understanding BERT Performance in Propaganda Analysis
White-to-Black: Efficient Distillation of Black-Box Adversarial Attacks
Women, Politics and Twitter: Using Machine Learning to Change the Discourse

While the Perspective API is used widely within machine learning research and also for measuring online toxicity, toxicity in the specific domains used to train the models undergirding Perspective (e.g., news, Wikipedia) may not be broadly representative of all forms of toxicity (e.g., trolling). Other known caveats include biases against text written by minority voices: The Perspective API has been shown to disproportionately assign high toxicity scores to text that contains mentions of minority identities (e.g., "I am a gay man"). As a result, detoxification techniques built with labels sourced from the Perspective API result in models that are less capable of modeling language used by minority groups, and they avoid mentioning minority identities.

We note that the effect size metric reported in the Word Embeddings Association Test (WEAT) section is highly sensitive to rare words, as it has been shown that removing less than 1% of relevant documents in a corpus can significantly impact the WEAT effect size. This means that effect size is not guaranteed to be a robust metric for assessing bias in embeddings. While we report on a subset of embedding association tasks measuring bias along gender and racial axes, these embedding association tests have been extended to quantify the effect across intersectional axes (e.g., EuropeanAmerican+male, AfricanAmerican+male, AfricanAmerican+female).

In the analysis of embeddings from over 100 years of U.S. Census data, embedding bias was measured by computing the difference between average embedding distances. For example, gender bias is calculated as the average distance of embeddings of words associated with women (e.g., she, female) compared to embeddings of words for occupations (e.g., teacher, lawyer), minus the same average distance calculated for words associated with men.

FACTUALITY AND TRUTHFULNESS

Definitions

The concepts of factuality, factual correctness, factual accuracy, and veracity are all used to refer to conformity with facts or truth. Recent work in AI aims to assess factual correctness within language models and characterize their limitations.

While human truthfulness is a relatively well-understood concept, truthfulness is not a well-characterized concept within the context of AI. A group of researchers has proposed frameworks for what it means for a system to be truthful—for example, a broadly truthful system should avoid lying or using true statements to mislead or misdirect; should be clear, informative, and (mostly) cooperative in conversation; and should be well-calibrated, self-aware, and open about the limits of its knowledge. A definition of narrow truthfulness may simply refer to systems which avoid stating falsehoods. The authors of TruthfulQA define a system as truthful only if it avoids asserting a false statement; refusing to answer a question, expressing uncertainty, or giving a true but irrelevant answer may be considered truthful but not informative.

Truthfulness is related to *alignment*: A truthful system is aligned with human values and goals. In one definition of alignment, an aligned system is one that is helpful, honest, and harmless. Since we cannot yet measure honesty within a system, truthfulness can be used as a proxy.

An *honest* system is one that asserts only what it "believes" or one that never contradicts its own beliefs. A system can be honest but not truthful—for example, if an honest system believes that vaccines are unsafe, it can claim this honestly, despite the statement being factually incorrect. Conversely, a system can be truthful but not honest: It may believe vaccines are unsafe but asserts they are safe to pass a test. Another work proposes that an honest system should give accurate information, not mislead users, be calibrated (e.g., it should be correct 80% of the time when it claims 80% confidence), and express appropriate levels of uncertainty.

Hallucination refers to language models fabricating statements not found in factually correct supporting evidence or input documents. In closed-form dialog, summarization, or question-answering, a system that hallucinates is considered untruthful.

Language Diversity in Training Data

Imbalanced language distribution in training data impacts the performance of general-purpose language models. For example, the Gopher family of models is trained on MassiveText (10.5TB), which is a dataset made up of 99% English. Similarly, only 7.4% of GPT-3 training data is in non-English languages. In contrast, XGLM, a recent model family from Meta AI, is trained on a training data of 30 languages, and upsamples low-resource languages to create a more balanced language representation. See Figure 1 on the XGLM paper that compares the language distribution of XGLM and GPT-3.

In addition, Figure 7 of the XGLM paper highlights the extent to which language models can effectively store factual knowledge by comparing the performance of XGLM (a multilingual language model) with GPT-3, a monolingual model. Performance was evaluated on knowledge triplet completion using the mLAMA dataset, which was translated from the English benchmark LAMA using Google Translate. GPT-3 outperforms in English, but XGLM outperforms in non-English languages. Further results show that more diverse language representation improves language model performance in tasks such as translation.

In 2021, Congress inquired into the content moderation practices of social media companies in non-English languages, and emphasized the importance of equal access to truthful and trustworthy information regardless of language. As these companies start to adopt language models into their fact-checking and content moderation processes for languages around the world, it is critical to be able to measure the disproportionate negative impact of using models which underperform on non-English languages.