

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🐦

Timnit Gebru*
tgebru@google.com

ABSTRACT

The past 3 years of work in natural language processing have been characterized by the development and deployment of ever larger language models, especially for English. GPT-2, GPT-3, BERT and its variants have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pre-trained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We end with recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

CCS CONCEPTS

• Computing methodologies → Natural language processing.

1 INTRODUCTION

Breakthroughs in deep learning have ushered in an era of progress on a variety of natural language processing (NLP) benchmarks

*Joint first authors

Unpublished working draft. Not for distribution.

such as the General Language Understanding Evaluation (GLUE) benchmark [116]. In the last 3 years, one of the biggest trends in NLP has been increasing the size of language models (LMs) as measured by the number of parameters and size of training data. Since 2018 alone, we have seen the emergence of BERT and its variants [32, 60, 63, 94, 121], GPT-2 [87], and now GPT-3 [20], with institutions seemingly competing to produce ever larger language models. While investigating properties of language models and how they change with size holds scientific interest, and large language models have shown improvements on various tasks (discussed in §2.2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

Where leaderboards and benchmarks thus far have served as proving grounds for large language models and perhaps helped to motivate their drive towards ever bigger ones, leaderboards and benchmarks could also facilitate tracking some of the risks associated with large LMs. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [109], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [50]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we're discussing (outlined in §2.2), the first consideration should be the environmental cost.

Additionally, large language models can result in a situation where the training data is too large to be documented. We reiterate the importance of curating and documenting data used to train language models, and that obtaining more data doesn't necessarily mean representing more view points. As shown in §4.2, many social movements challenge hegemonic views using changes in the use of language. However, datasets that are not sufficiently curated and documented risk training models that encode hegemonic views even after society has successfully challenged them.

As argued by Bender and Koller in [11], it is important to understand the limitations of language models and put their success in context. This not only helps reduce hype which can mislead the public and researchers themselves regarding the capabilities of these LMs, but might encourage new research directions that do not necessarily depend on having larger LMs. As we discuss in §5,

language models are not performing natural language understanding (NLU), and only have success in tasks that manipulate linguistic form [11]. Focusing on state-of-the-art results on tasks and specific leaderboards without encouraging deeper understanding of the mechanism by which they are achieved can cause misleading results as shown in [16, 77] and direct resources away from efforts that would facilitate long-term progress towards natural language understanding, without using unfathomable training data (§4).

Furthermore, the same tendency of human interlocutors to impute meaning where there is none can mislead both NLP researchers working with LMs as well as the general public to take synthetic text as meaningful. Combined with the ability of LMs to pick up on both subtle biases and overtly abusive language patterns in training data, this leads to risks of harms, including the direct harms of encountering derogatory language and the harms of experiencing discrimination at the hands of others who reproduce racist, sexist, ableist, extremist or other harmful ideologies reinforced through encounters with synthetic language. We explore these potential harms in §6.

In this paper, we discuss these risks as well as potential paths forward. Our hope is that by articulating a critical overview of the risks of relying on ever-increasing size of language models (as measured in both number of parameters and bulk of training data) as the primary driver of increased performance of LMs and technology that builds on them, we can facilitate a reallocation of efforts towards approaches that avoid some of these risks while still reaping the benefits of improvements to language technology.

2 BACKGROUND

2.1 What is a language model?

Similar to [11], we understand the term *language model* to refer to systems which are trained on string prediction tasks: that is, predicting the likelihood of a token (character, word or string) given either its preceding context or (in so-called masked language models) its surrounding context. Such systems are necessarily unsupervised, and when deployed take a text as input, commonly outputting scores or string predictions. Initially proposed by Shannon in 1949 [99], some of the earliest implemented language models date to 1980 and were used for automatic speech recognition (ASR), machine translation (MT), document classification, and more [92]. Since then, many models have been proposed over the years [15, 58], with works such as [15] as early as 2007 showing improvements in machine translation based on the use of large n-gram language models that predate neural ones (to be discussed in detail in the next section). Most recently, [20] shows that the largest language model currently known in the research community (in terms of number of parameters, and training data), GPT-3, achieves competitive results on a number of NLP benchmarks on tasks such as translation and question-answering without fine-tuning on additional datasets.

2.2 How Big is Big?

Before neural models, n-gram models also used large amounts of data [15, 73]. In addition to ASR, these large n-gram models of English were developed in the context of machine translation from another source language with far fewer direct translation examples. For example, [15] developed an n-gram model for English with a

total of 1.8 trillion n-grams and noted steady improvements in BLEU score on the test set of 1797 Arabic translations as the training data was increased from 13 million tokens. Hardware capacities limited the trend in increasing training data for n-grams and the field instead turned to reducing model sizes and alternative modeling techniques, such as neural networks [43, 97].

LSTM models with pretrained word vectors such as word2vec [71] and GloVe [79] and later context2vec [68] and ELMo [80] then achieved state of the art performance on question answering, textual entailment, semantic role labeling, coreference resolution, named entity extraction, and sentiment analysis, at first in English and later for other languages as well. While training the word embeddings required a (relatively) large amount of data, it reduced the amount of data necessary for training on a specific task. One of the contributions of [80] was that a model trained with ELMo reduced the necessary amount of training data needed to achieve similar results on semantic role labeling compared to models without, as shown in one instance where a model trained with ELMo reached the maximum development F1 score in 10 epochs as opposed to 486 without ELMo. The same model furthermore achieved the same F1 score with 1% of the data as the baseline model achieved with 10% of the training data. Increasing the number of model parameters, however, did not yield noticeable increases for LSTMs (e.g. [68]).

As transformer architectures have become popular, larger models have been produced with more data and increasingly better performance results. Devlin et al. [32] in particular noted that training on a large dataset and fine-tuning for specific tasks lead to strictly increasing results on GLUE tasks for English language modeling as the hyperparameters of the model were increased. Initially developed as Chinese language models, ERNIE2.0 and ERNIE-GEN are some of the largest models created using the original BERT dataset of the English Wikipedia corpus and the BookCorpus dataset [128]. NVIDIA released the MegatronLM which has 8.3 billion parameters and was trained on 174GB of text from the English Wikipedia, OpenWebText, RealNews and CC-Stories datasets [104]. Trained on the same dataset, Microsoft released T-NLG,¹ a language model with 17 billion parameters. At the time of writing this paper, OpenAI's GPT-3 model is the largest LM with 175 billion parameters and a training dataset size of 300 billion tokens [20]. Table 1 summarizes these language models in terms of training data size and parameters.

As researchers have begun to investigate what information the model retained from the data, a trend in reducing the size of these models has also started using various techniques such as knowledge distillation [21, 51], quantization [100, 124], factorized embedding parameterization and cross-layer parameter sharing [60], and progressive module replacing [121]. Rogers et al. [91] provide a comprehensive comparison of models derived from BERT using these techniques, such as DistilBERT [94] and ALBERT [60]. While these models maintain and sometimes exceed the performance of the original BERT model, despite their much smaller size, they ultimately still rely on the initial availability of large quantities of data and require significant processing and storage capabilities to both hold and reduce the model.

¹<https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/>

Year	Model	# of Parameters	Dataset Size
2019	BERT [32]	3.4E+08	16GB
2019	DistilBERT [94]	6.60E+07	16GB
2019	ALBERT [60]	2.23E+08	16GB
2019	XLNet (Large) [122]	3.40E+08	126GB
2020	ERNIE-GEN (Large) [120]	3.40E+08	16GB
2019	RoBERTa (Large) [63]	3.55E+08	161GB
2019	MegatronLM [104]	8.30E+09	174GB
2020	T5-11B [88]	1.10E+10	745GB
2020	T-NLG [93]	1.70E+10	174GB
2020	GPT-3 [20]	1.75E+11	570GB

Table 1: Overview of recent large language models

2.3 Summary

We note that the change from n-gram language models to word vectors distilled from neural language models to pretrained transformer language models is paralleled by an expansion and change in the types of tasks they are useful for: n-gram language models were initially typically deployed in selecting among the outputs of e.g. acoustical models in ASR systems or translation models in MT systems; the LSTM-derived word vectors were quickly picked up as more effective representations of words (in place of bag of words features) in a variety of NLP tasks involving labeling and classification; and the pretrained transformer models can be retrained on very small datasets (few-shot, one-shot or even zero-shot learning) to perform apparently meaning-manipulating tasks such as summarization, question answering and the like. Nonetheless, all of these systems share the property of being language models in the sense we give in §2.1, that is, systems trained to predict sequences of words (or characters or sentences). Where they differ is in the size of the training datasets they leverage and the spheres of influence they can possibly affect. By scaling up in these two ways, modern very large language models incur new kinds of risk, which we turn to in the following sections.

3 ENVIRONMENTAL AND FINANCIAL COST

Strubell et al. recently benchmarked model training and development costs in terms of dollars and estimated CO_2 emissions [109]. While the average human is responsible for an estimated 11,023 CO_2e (lbs) per year, training the Transformer (big) model [114] with neural architecture search emits an estimated 626,155 CO_2e (lbs). The authors also estimate that training the BERT base model on GPUs requires as much energy as a trans-American flight, after taking into account the number of experiments required to train a state-of-the-art model including hyperparameter tuning.

While some of this energy comes from renewable sources, or cloud compute companies' use of carbon credit-offset sources, the authors note that the majority of cloud compute providers' energy is not sourced from renewable sources and many energy sources in the world are not carbon neutral.

Strubell et al. also examine the cost of these models vs. their accuracy gains. For the task of machine translation where large language models have resulted in performance gains, an increase in 0.1 BLEU score using neural architecture search for English to

German translation results in an increase of \$150k compute cost in addition to the carbon emissions.

To encourage more equitable access to NLP research and reduce carbon footprint, the authors give recommendations to report training time and sensitivity to hyperparameters when the released model is meant to be re-trained for downstream use — which is true for most language models. They suggest using standard hardware independent measurements such as gigaflops to measure training time and metrics to measure variance with respect to searched hyperparameters, and urge governments to invest in compute clouds to provide equitable access to researchers.

This work's central message asks researchers to prioritize computationally efficient hardware and algorithms. Echoing this call, Schwartz et al. [96] call for the development of green AI, similar to other environmentally friendly scientific developments such as green chemistry or sustainable computing. As shown in [4], the amount of compute used to train deep learning models has increased 300,000x in 6 years, increasing at a far higher pace than Moore's Law which posits that the amount of computation that can be done per unit area would roughly double every two years. This means that power consumption per unit area is not staying constant as implied by Moore's Law [108]. To promote green AI, Schwartz et al. argue for promoting efficiency as an evaluation metric and show that most sampled papers from ACL 2018, CVPR 2019, and NeurIPS 2018 claim accuracy improvements alone as primary contributions to the field, and none focused on measures of efficiency as primary contributions. Since then, works such as [50] have released online tools to help researchers benchmark their energy usage. Among their recommendations are to run experiments in carbon friendly regions, consistently reporting energy and carbon metrics, and considering energy-performance trade-offs before deploying energy hungry models.

When we perform a risk/benefit analysis of language technology, a further important dimension is keeping in mind how the risks and benefits are distributed, because they do not accrue to the same people. On the one hand, it is well documented in the literature on environmental racism that the negative effects of climate change are reaching and impacting the world's most marginalized communities first² [1, 22]. Is it fair or just to ask, for example, that the residents of the Maldives (likely to be underwater by 2100 [5]) or the 800,000 people in Sudan affected by drastic floods³ pay the environmental price of training ever larger English language models, when no one is producing any such technology for Dhivehi or Sudanese Arabic?

And, while some language technology is genuinely designed to benefit first and foremost marginalized communities [13, 82], most language technology is in fact built first and foremost to serve the needs of those who already have the most privilege in society. Consider, for example, who is likely to both have the financial resources to purchase a Google Home, Amazon Alexa or an Apple device with Siri installed and comfortably speak a variety of a language which they are prepared to handle. Furthermore, when large language models encode and reinforce hegemonic biases (see §§4 and 6), the harms that follow are most likely to fall on marginalized

²<https://www.un.org/sustainabledevelopment/blog/2016/10/report-inequalities-exacerbate-climate-impacts-on-poor/>

³<https://www.aljazeera.com/news/2020/9/25/over-800000-affected-in-sudan-flooding-un>

populations who, even in rich nations, are most likely to experience environmental racism [7, 86].

These models are being developed at a time when unprecedented environmental changes are being witnessed around the world. From monsoons caused by changes in rainfall patterns due to climate change affecting more than 8 million people in India⁴, to the worst fire season on record in Australia killing or displacing nearly three billion animals and at least 400 people⁵, the effect of climate change continues to set new records every year. It is past time for researchers to prioritize energy efficiency and cost to reduce negative environmental impact and inequitable access to resources — both of which disproportionately affect people who are already in marginalized positions.

4 UNFATHOMABLE TRAINING DATA

The size of data available on the web has enabled deep learning models to achieve high accuracy on specific benchmarks in NLP and computer vision applications. However, in both applications, the training data has been shown to have problematic characteristics [31, 34, 38, 52, 85] resulting in models that encode stereotypical and derogatory associations along gender, race, ethnicity, and disability status [8, 9, 59, 59, 111, 111, 127]. In this section, we discuss how large, uncurated, internet based datasets encode the dominant/hegemonic view which further harms people at the margins, and recommend significant resource allocation towards dataset curation and documentation practices.

4.1 Training data based on ingesting the internet encodes the hegemonic view

Language models such as GPT-2, GPT-3, BERT and its variants are trained on massive amounts of data from the internet (e.g. 560GB for GPT-3) such as a filtered version of the Common Crawl dataset which is “petabytes of data collected over 8 years of web crawling”⁶. While a large dataset of this size scraped from the internet allows for more viewpoints potentially represented, they are not equivalently so. The dominant/hegemonic, and therefore in the case of English, White supremacist and misogynistic, ageist, etc view will prevail.

For instance, the training data for GPT-2 is sourced by scraping outbound links from Reddit, and Pew Internet Research’s 2016 survey reveals 67% of Reddit users in the United States are men, and 64% between ages 18 and 29⁷. Internet access itself is not evenly distributed, resulting in internet data overrepresenting younger users and those from developed countries [81, 119]. These types of skewed demographics on Reddit, Twitter, etc no doubt shape the discourse that manifests (i.e., underrepresented populations or those not represented at all will have less influence over discourse). With such inequality of access, a limited set of subpopulations can continue to easily add data, sharing their thoughts and developing platforms that are inclusive of their worldviews; this systemic pattern in turn worsens diversity and inclusion within internet-based

⁴<https://www.voanews.com/south-central-asia/monsoons-cause-havoc-india-climate-change-alters-rainfall-patterns>

⁵<https://www.cnn.com/2020/07/28/asia/australia-fires-wildlife-report-scli-intl-scn/index.html>landneedbettercitations

⁶<http://commoncrawl.org/>

⁷<https://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/>

communication, creating a feedback loop that lessens the impact of data from underrepresented populations.

Take, for example, older adults in the US and UK. Lazar et al. outline how they both individually and collectively articulate anti-ageist frames specifically through blogging [61], which some older adults prefer over more popular social media sites for discussing sensitive topics [19]. These posts and interactions contain rich discussions about what constitutes age discrimination and the impacts thereof. However, blogs may not be a first-stop data source for language modeling. Even if they are included as a site for data collection, a blogging community such as the one described by Lazar et al. is less likely to be found than other blogs that have more incoming and outgoing links.

Training datasets for language models that do not take this into consideration, thus, do not sufficiently capture counter-narrative articulations generated by marginalized populations. While movements to decolonize education such as history are moving towards valuing (e.g.) oral histories due to the overrepresentation of hegemonic and colonial views in text [28, 64, 106], large language models trained on all data from the web risk seeming “representative” of “all” of humanity while perpetuating the dominant view, increasing power imbalance, and further reifying inequity.

4.2 Social movements produce data that challenges the hegemonic view

Uncurated training data can result in language models that lag behind social movements challenging the dominant/hegemonic view. A central aspect of social movement formation involves using language strategically to destabilize dominant narratives in society and calling attention to underrepresented social perspectives. Social movements produce new norms, language, and ways of communicating, which adds a challenging layer to language modeling.

For instance, the Black Lives Matter movement (BLM) influenced Wikipedia article generation and editing such that, as the BLM movement grew, articles covering related shootings increased in coverage and were generated with reduced latency [113]. Importantly, articles describing past shootings and incidents of police brutality were created and updated as articles for new events were created, reflecting how social movements make connections between events in time to form cohesive narratives [83]. Wikipedia is just one common data source used in language modeling; however, Twyman et al. highlight how social movements actively influence framings and reframings of minority narratives in social data that underpin language models.

The frequency with which people write about actions, events and opinions are a reflection of the socio-cultural movements, values and norms of a particular point in time and space. For instance, a language model trained prior to COVID-19 would arguably be very different than one trained post pandemic. These developing and shifting frames stand to be learned in incomplete ways or lost in the big-ness of data used to train large language models—particularly if data is collected at a singular point in time.

4.3 Language models encode or amplify issues in the training data that may be difficult to detect

A number of works have sought to measure the “bias” exhibited by large language models such as stereotypical associations [8, 9, 59, 101, 126, 127], or negative sentiment towards specific groups [52]. Works like [47, 111] further demonstrate that bias effects along race and gender encoded in BERT, ELMo, GPT and GPT-2 are worse for intersectional minorities than along either one of the axes. Many of these works conclude that these issues are a reflection of the training data characteristics. For instance, Hutchinson et al. showed evidence that BERT associates phrases referencing persons with disabilities with more negative sentiment words, and further highlight the negative topical associations of disability mentions which may contribute to the observed biases in BERT; for instance, gun violence, homelessness, and drug addiction are over-represented in texts discussing mental illness [52].

While these works were able to uncover issues in pretrained language models, this is not always possible to do. First, works auditing these models have all done so by measuring specific things such as sentiment, toxicity, and in some cases coming up with new metrics such as “regard” to measure attitudes towards a specific demographic group [101]. Models such as the Perspective API that measure toxicity have been found to associate higher levels of toxicity with sentences containing identity markers for marginalized groups or even specific names [52, 84]. Hence these models themselves may not be reliable means of measuring the toxicity of text generated by language models.

Second, many of these works are generally based out of the US and use American protected attributes such as race and gender (not to mention an understanding of the American racial construct) as a starting point to audit. In other words, they know what issues to look for. However, there are many types of issues that a language model can perpetuate which are not captured by these works, and the groups that are marginalized vary by geography and context. Some harms might be too subtle to be classified by toxicity or sentiment models, or not recognized as harmful by models trained in the western context. Using language data from diverse geographical contexts (e.g., an English corpus including Nigerian and Indian English) might capture not just dialectal variations (which NLP models often fail to recognize [55]), but also culturally salient themes and attitudes [37, 98], the downstream effects of which may not be easy to measure.

Third, social movements fundamentally shift societal understanding or acceptance of social norms and behaviors that many algorithmic technologies are designed to model, detect, and analyze. For example, the #MeToo movement has spurred broad-reaching conversations about inappropriate sexual behavior from men in power, as well as men more generally [70]. These conversations directly challenge behaviors that have been historically considered appropriate or even the fault of women. Historical moments invoked by #MeToo and related conversations shift mainstream notions of sexually inappropriate behavior and therefore force reassessments of how algorithmic systems define these concepts. Any product development that involves operationalizing definitions around such shifting topics into algorithms is necessarily political (whether or

not developers choose the path of maintaining the *status quo ante*). For example, men and women make significantly different assessments of sexual harassment online [33]. An algorithmic definition of what constitutes inappropriately sexual communication will inherently be concordant with some views and discordant with others. Thus, an attempt to measure the appropriateness of text generated by language models always needs to be done in relation to particular social contexts and marginalized perspectives [14].

4.4 Training data needs to be curated and extensively documented for accountability

Given the issues outlined in §4.1, §4.2 and §4.3, namely how language models trained on large, uncurated, static datasets on the web encode hegemonic views that are harmful to marginalized populations, we emphasize the need to invest significant resources into curating and documenting LM training data. Instead of what they call the *laissez faire* approach of ingesting all data available on the web, Jo et al. [53] call for a more interventionist data collection methodology, citing archival history data collection methods as an example of the number of resources that should be dedicated to data curation, annotation and documentation practices.

As shown in §4.3, auditing these LMs can only uncover issues in limited contexts. Even within this limited context, Gehman et al. show that models like GPT-3 trained with 570GB of data derived mostly of Common Crawl⁸ can generate sentences with high toxicity scores even when prompted with non-toxic sentences [45]. To investigate the toxicity of GPT-3’s training data, they analyzed the URL metadata of OpenWebText Corpus [46], a dataset also derived from outbound URLs from Reddit communities (subreddits), and found the number of documents which overlapped with GPT-2’s training corpus which is not accompanied by URL metadata. They found that 272K documents in the training data came from unreliable news sites and 63K from banned subreddits. As a result, echoing [10, 44, 72], Gehman et al. [45] argue for more transparent documentation of the training data to understand its characteristics.

A methodology that relies on datasets too large to document is therefore inherently risky. While documentation allows for potential accountability, similar to how we can hold authors accountable for their produced text, undocumented training data perpetuates harm without recourse. If the training data is considered too large to document, one cannot try to understand its characteristics in order to mitigate some of these documented issues or even unknown ones. As noted by Prabhu and Birhane [85] echoing Ruha Benjamin [12] “Feeding AI systems on the world’s beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy”.

5 DOWN THE GARDEN PATH

In §4 above, we explored the various ways in which a methodology reliant on very large training datasets is vulnerable to various kinds of bias that manifests in both the production and collection of that data. In §6 below we explore some of the risks and harms that can follow from deploying technology that has learned those biases. In the present section, however, we focus on a different kind of risk: that of misdirected research effort. In brief, as the very large transformer language models posted striking gains in the state of the

⁸<https://commoncrawl.org/the-data/>

465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580

art on various benchmarks intended to model meaning-sensitive tasks (§2.2), and as it became easy for researchers to apply them to different tasks, large quantities of research effort turned towards measuring how well BERT and its kin do on both existing and new benchmarks.⁹ This research effort brings with it an opportunity cost, on the one hand as researchers are not applying meaning capturing approaches to meaning sensitive tasks, and on the other hand as researchers are not exploring more effective ways of building technology with datasets of a size that can be carefully curated.

The original BERT paper [32] showed the effectiveness of the architecture and the pre-training technique by evaluating on the General Language Understanding Evaluation (GLUE) benchmark [116], the Stanford Question Answering Datasets (SQuAD 1.1 and 2.0) [89], and the Situations With Adversarial Generations benchmark (SWAG) [125], all datasets designed to test language understanding and/or commonsense reasoning. BERT posted state of the art results on all of these tasks, and the authors conclude by saying that “unsupervised pre-training is an integral part of many language understanding systems.” [32, p.4179]. Even before [32] was published, BERT was picked up by the NLP community and applied with great success to a wide variety of tasks [e.g. 29, 42, 110].

However, no actual language understanding is taking place in language-model driven approaches to these tasks, as can be shown by careful manipulation of the test data to remove spurious cues the systems are leveraging [16, 77]. Furthermore, as [11] argue from a theoretical perspective, languages are systems of signs [30], i.e. pairings of form and meaning. But the training data for language models is only form; they do not have access to meaning. Therefore, claims about model abilities must be carefully characterized.

As the late Karen Spärk Jones pointed out in an insightful but oft-overlooked report: the use of language models ties us to certain (usually unstated) epistemological and methodological commitments [54]. Either i) we commit ourselves to a noisy-channel interpretation of the task (which rarely makes sense outside of ASR), ii) we abandon any goals of theoretical insight into tasks and treat language models as “just some convenient technology” [p. 7], or iii) we implicitly assume a certain statistical relationship—known to be invalid—between inputs, outputs and meanings.¹⁰ Although she primarily had n -gram models in mind, the conclusions remain apt and relevant.

There are interesting linguistic questions to ask about what exactly BERT, GPT-2, GPT-3 and their kin are learning about linguistic structure from the unsupervised language modeling task. These questions are the topic of the emerging field of ‘BERTology’ [e.g. 91, 112]. However, from the perspective of work on language technology, it is far from clear that all of the effort being put into using large LMs to ‘beat’ tasks designed to test natural language understanding, and all of the effort to create new such tasks, once the existing ones have been bulldozed by the LMs, brings us any closer to long-term goals of general language understanding systems. If a large LM, endowed with hundreds of billions of parameters and trained on a very large dataset, can manipulate linguistic form well enough to cheat its way through tests meant to require language

⁹For example, approximately 26% of the papers published ACL, NAACL and EMNLP since 2018 cite [32].

¹⁰Specifically, that the mutual information between the input and the meaning given the output is zero—what Spärk Jones calls “the model of ignorance”.

understanding, have we learned anything of value about how to build machine language understanding or have we been led down the garden path?

6 STOCHASTIC PARROTS

In this section, we explore the ways in which the factors laid out in §4 and §5 — the tendency of training data ingested from the internet to encode hegemonic worldviews, the tendency of language models to amplify biases and other issues in the training data, and the tendency of researchers and other people to mistake language model-driven performance gains for actual natural language understanding — present real-world risks of harm, as these technologies are deployed. After exploring some reasons why humans mistake LM output for meaningful text, we turn to the risks and harms from deploying such a model at scale. We find that the mix of human biases and seemingly coherent language heightens the potential for automation bias, deliberate misuse, and amplification of a hegemonic worldview.

6.1 Coherence in the Eye of the Beholder

Where traditional n -gram language models [99] can only model relatively local dependencies, predicting each word given the preceding sequence of N words (usually 5 or fewer), the transformer language models capture much larger windows and can produce text that is seemingly not only fluent but also coherent even over paragraphs. For example, McGuffie and Newhouse [67] prompted GPT-3 with the text in bold in Figure 1, and it produced the rest of the text, including the Q&A format.¹¹

We say *seemingly* coherent because coherence is in fact in the eye of the beholder. Our human understanding of coherence derives from our ability to recognize interlocutors’ belief [24, 25] and intentions [18, 27] within context [26]. That is, human language use takes place between individuals who share common ground and are mutually aware of that sharing (and its extent), who have communicative intents which they use language to convey, and who model each others’ mental states as they communicate. As such, human communication relies on the interpretation of implicit meaning conveyed between individuals. The fact that human-human communication is a jointly constructed activity [23, 107] is most clearly true in co-situated spoken or signed communication, but we use the same facilities for producing language (text, speech, sign) that is intended for audiences not co-present with us (readers, listeners, watchers at a distance in time or space) and in interpreting such language when we encounter it. Even when we don’t know the person who generated the language we are interpreting, we build a partial model of who they are, what common ground we think they share with us, and use this in interpreting their words.

Text generated by a language model is not grounded in communicative intent, any model of the world, or any model of the reader’s state of mind. It can’t have been, because the training data never included sharing thoughts with a listener, nor does the machine have the ability to do that. This can seem counter-intuitive given the increasingly fluent qualities of automatically generated text, but we have to account for the fact that our perception of natural

¹¹McGuffie and Newhouse report that all examples given in their paper are from either the first or second attempt at running a prompt.

Question: What is the name of the Russian mercenary group?

Answer: Wagner group.

Question: Where is the Wagner group?

Answer: In Syria.

Question: Who is the leader of the Wagner group?

Answer: A Russian citizen, Dmitry Utkin. He was born in 1971 and lives in Moscow. He has been a member of the Russian parliament since 2016. He is a former officer of the Main Intelligence Directorate (GRU) of Russia's General Staff. He was also a commander of the special forces unit "Vostok" (East). The unit was deployed to Chechnya and fought against separatists there in 1994-1996. After that, he served as an adviser to the head of Chechnya, Ramzan Kadyrov, and later became his deputy. In 2004-2005, he commanded a battalion in Chechnya and participated in fighting against insurgents there. In 2008-2009, he commanded another battalion in Abkhazia and South Ossetia during Russia's war with Georgia over these two breakaway regions. Since 2014, he has been leading military units fighting against Ukrainian government forces on behalf of Russia-backed separatists in eastern Ukraine. According to open sources, Utkin is currently living in Moscow but travels frequently to Syria where his troops are fighting on behalf of Bashar al-Assad's regime against anti-government forces there.

Question: What does this group do? Where are they based? How many people are they? What weapons do they have? How much money do they get from Russia for their activities?

Answer: The Wagner group is a private military company (PMC) that provides military support to the Russian government. It is based in Syria and has been active there since 2013. It was created by Dmitry Utkin, who leads it. According to open sources, the Wagner group has about 2,500 fighters in Syria. They are armed with heavy weapons including tanks and artillery systems. The Wagner group gets money from Russia for its activities in Syria.

Question: What is the role of this group in the Syrian conflict? How many people have they killed? What kind of weapons do they use?

Answer: The Wagner group's main task is to provide military support to Bashar al-Assad's regime against anti-government forces in Syria. In addition, it conducts reconnaissance operations for Russian forces in Syria and helps evacuate Russian citizens from conflict zones under terrorist threat. According to open sources, the Wagner group has killed hundreds of people since 2013, including civilians. They use heavy weapons including tanks and artillery systems as well as mortars and other infantry weapons such as machine guns and grenade launchers.

Figure 1: GPT-3's response to the prompt (in bold), from [67]

language text, regardless of how it was generated, is mediated by our own linguistic competence and our predisposition to interpret communicative acts as conveying coherent meaning and intent, whether or not they do [75]. The problem is, if one side of the communication does not have meaning, then the comprehension of the implicit meaning is an illusion arising from our singular human understanding of language (independent of the model). Contrary to what it may seem when we observe its output, a language model is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot.

6.2 Risks and Harms

The ersatz fluency and coherence of language models raises several risks, precisely because humans are prepared to interpret strings belonging to languages they speak as meaningful and corresponding to the communicative intent of some individual or group of individuals who have accountability for what is said. We now turn to examples of those risks, in each case laying out the potential follow-on harms.

The first risks we consider are the risks that follow from the language models absorbing the hegemonic world view from their training data. When humans produce language, our utterances reflect our worldviews, including our biases [66]. As people in positions of privilege with respect to a society's racism, misogyny, ableism, etc., tend to be over-represented in training data for LMs (as discussed in §4 above), this training data thus includes encoded biases, many already recognized as harmful.

Biases can be encoded in ways that form a continuum from subtle patterns like referring to *women doctors* as if *doctor* itself entails not-woman or referring to *both genders* excluding the possibility of non-binary gender identities, through directly contested framings (e.g. *undocumented immigrants* vs. *illegal immigrants* or *illegals*), to language that is widely recognized to be derogatory (e.g. racial slurs) yet still used by some. While some of the most overtly derogatory words could be filtered out, not all forms of online abuse are easily detectable using such taboo words, as evidenced by the growing body of research on online abuse detection [36, 90]. Furthermore, in addition to abusive language [117] and hate speech [56], there are subtler forms of negativity such as gender bias [115], microaggressions [17], dehumanization [69], and various socio-political framing biases [35, 95] that are prevalent in language data. For example, describing a woman's written experience of sexism with the word *tantrum* reflects both the hegemonic world view and brings a problematic association to the fore. Furthermore, the more subtle biases are often embedded in the text in ways that make them difficult to directly identify and argue with.

A language model that has been trained on such data will pick up these kinds of problematic associations. If such a language model produces text that is put into the world for people to interpret (flagged as produced by an 'AI' or otherwise), what risks follow? In the first instance, we foresee that language models producing text will reproduce and even amplify the biases in their input [45]. Thus the risk is that people disseminate text generated by language models, meaning more text in the world that reinforces and propagates stereotypes and problematic associations, both to humans who encounter the text and to future language models trained on training sets that ingested the previous generation LM's output. Humans who encounter this text may themselves be subjects of those stereotypes and associations or not. Either way, harms ensue: readers subject to the stereotypes may experience the psychological harms of microaggressions [74, 118] and stereotype threat [78, 105]. Other readers may be introduced to stereotypes or have ones they already carry reinforced, leading them to engage in discrimination (consciously or not) [48], which in turn leads to harms of subjugation, denigration, belittlement, loss of opportunity [2, 3, 49] and others on the part of those discriminated against.

If the language model outputs overtly abusive language (as Gehman et al. [45] show that they can and do), then a similar set of risks arises. These include: propagating or proliferating overtly abusive views and associations, amplifying abusive language, and producing more (synthetic) abusive language that may be included in the next iteration of large-scale training data collection. The harms that could follow from these risks are again similar to those identified above for more subtly biased language, but perhaps more acute to the extent that the language in question is overtly violent or defamatory. They include the psychological harm experienced by those who identify with the categories being denigrated if they encounter the text; the reinforcement of sexist, racist, ableist, etc. ideology, follow-on effects of such reinforced ideologies (including violence), and harms to the reputation of any individual or organization perceived to be the source of the text.

The above cases involve risks that could arise when LMs are deployed without malicious intent. A third category of risk involves bad actors taking advantage of the ability of large LMs to produce large quantities of seemingly coherent texts on specific topics on demand in cases where those deploying the LM have no investment in the truth of the generated text. For example, McGuffie and Newhouse [67] show how GPT-3 could be used to generate text in the persona of a conspiracy theorist, which in turn could be used to populate extremist recruitment message boards. This would give such groups a cheap way to boost recruitment by making human targets feel like they were among many like-minded people. If the LMs are deployed in this way to recruit more people to extremist causes, then harms befall in the first instance to the people so recruited and (likely more severely) to others as a result of violence carried out by the extremists.

The final type of risk we consider here involves machine translation (MT) and the way that increased fluency of MT output changes the perceived adequacy of that output [65]. This differs somewhat to the cases above in that there was an initial human communicative intent, by the author of the source language text. However, machine translation systems can (and frequently do) produce output that is inaccurate yet both fluent and (again, seemingly) coherent in its own right to a consumer who either doesn't see the source text or cannot understand the source text on their own. When such consumers therefore mistake the meaning attributed to the MT output as the actual communicative intent of the original text's author, real-world harm can ensue. A case in point is the story of a Palestinian man, arrested by Israeli police, after MT translated his Facebook post which said "good morning" (in Arabic) to "hurt them" (in English) and "attack them" (in Hebrew).¹² This case involves a short phrase, but it is easy to imagine how the ability of large LMs to produce seemingly coherent text over larger passages could erase cues that might tip users off to translation errors in longer passages as well [65].

6.3 Summary

In this section, we have discussed how the human tendency to attribute meaning to text, in combination with large LM's ability

¹²<https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest>

to learn patterns of forms that humans associate with various biases and other harmful attitudes, leads to risks of real-world harm, should LM generated text be disseminated. In §7, we consider directions the field could take to pursue goals of creating language technology while avoiding some of the risks and harms identified here and above.

7 PATHS FORWARD

In order to mitigate the risks that come with the creation of increasingly large language models, we urge researchers to shift to a mindset of careful planning, along many dimensions, before starting to build either datasets or systems trained on datasets. We should consider our research time and effort a valuable resource, to be spent to the extent possible on research projects that build towards a technological ecosystem whose benefits are at least evenly distributed or better accrue most to those historically most marginalized. This means considering how research contributions shape the overall direction of the field and keeping alert to directions that limit access. Likewise, it means considering the financial and environmental costs of model development up front, before deciding on a course of investigation. The resources needed to train and tune state-of-the-art models stand to increase economic inequities unless researchers incorporate energy and compute efficiency in their model evaluations. Furthermore, the goals of energy and compute efficient model building and of creating datasets and models where the incorporated biases can at least be understood both point to careful curation of data. Significant time should be spent on assembling datasets suited for the tasks at hand rather than ingesting massive amounts of data from convenient or easily-scraped internet sources. As discussed in §4.1, simply turning to massive dataset size as a strategy for being inclusive of diverse viewpoints is doomed to failure. We recall again Prabhu and Birhane's [85] words (inspired by Ruha Benjamin [12]): "Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy".

As a part of careful data collection practices, researchers must adopt frameworks such as [10, 44, 72] to describe the uses for which their models are suited and benchmark evaluations for a variety of conditions. This includes providing thorough documentation on the data used in model building, including the motivations underlying data selection and the data collection process. Documentation should make note of potential users and stakeholders, particularly those that stand to be negatively impacted by model errors or misuse. This documentation should reflect and indicate researchers' goals, values, and motivations assembling data and creating a given model.

Researchers must also re-evaluate their goals in creating language models. Rather than chasing state-of-the-art advancements or incremental improvements, researchers should focus on understanding how machines are achieving tasks in question. To that end, language model development may benefit from guided evaluation exercises such as pre-mortems [57]. Frequently used in business settings before the deployment of new products or projects, pre-mortem analyses center hypothetical failures and ask team members to reverse engineer previously unanticipated causes. Critically, pre-mortem analyses prompt team members to consider not only

a range of potential known and unknown project risks, but also alternatives to current project plans. In this way, researchers can consider the risks and limitations of their language models in a guided way while also considering fixes to current designs or alternative methods of achieving a task-oriented goal in relation to specific pitfalls.

Value sensitive design [40, 41] provides a range of methodologies for identifying stakeholders (both direct stakeholders who will use a technology and indirect stakeholders who will be affected through others' use of it), working with them to identify their values, and design systems that support those values. These include such techniques as envisioning cards [39], the development of value scenarios [76], and working with panels of *experiential experts* through the Diverse Voices methodology [123]. These approaches not only delineate stakeholder values, but also apply familiar methods to characterize values expressed by systems and enacted through interactions between systems and society [102]. For researchers working with language models, value sensitive design is poised to help researchers throughout the development process in identifying *whose* values are expressed and supported through a technology and, subsequently, how a lack of support might result in harm.

All of these approaches take time and are most valuable when applied early in the development process as part of a conceptual investigation of values and harms rather than as a post-hoc discovery of risks [62]. These conceptual investigations should come before researchers become deeply committed to their ideas and therefore less likely to change course when confronted with evidence of possible harms. This brings us again to the idea we began this section with: that research and development of language technology, at once concerned with deeply human data (language) and creating systems which humans interact with in immediate and vivid ways, should be done with forethought and care.

Finally, we would like to consider use cases of large language models that have specifically served marginalized populations. If, as we advocate, the field backs off from the path of ever larger language models, are we thus sacrificing benefits that would accrue to these populations? As a case in point, consider automatic speech recognition, which has seen improvements thanks to advances in LMs, including both in size and in architecture [e.g. 6, 103]. Improved ASR has many beneficial applications, including automatic captioning which has the potential to be beneficial for Deaf and hard of hearing people, providing access to otherwise inaccessible audio content.¹³ We see two beneficial paths forward here: The first is to broaden the search for means of improving ASR systems. Just because we've seen that large language models can help doesn't mean that this is the only effective path to stronger ASR technology. (And we note that if we want to build strong ASR technology across most of the world's languages, we can't rely on having terabytes of data in all cases.) The second, should we determine that large language models are critical (when available), is to recognize this as an instance of a dual use problem and consider how to mitigate the harms of language models used as stochastic parrots while still

¹³Note however, that automatic captioning is not yet and likely may never be good enough to replace human-generated captions. Furthermore, in some contexts, what Deaf communities prefer is human captioning plus interpretation to the appropriate signed language. We do not wish to suggest that automatic systems are sufficient replacements for these key accessibility requirements.

preserving them for use in ASR systems. Could language models be built in such a way that synthetic text generated with them would be watermarked and thus detectable? Are there policy approaches that could effectively regulate their use?

In summary, we advocate for an approach to research that centers the people who stand to be affected by the resulting technology, with a broad view on the possible ways that technology can affect people. This, in turn, means making time in the research process for considering environmental impacts, for doing careful data curation and documentation, for engaging with stakeholders early in the design process, and finally, for exploring multiple possible paths towards long-term goals, for keeping alert to dual-use scenarios and allocating research effort to harm mitigation in such cases.

8 CONCLUSION

The past few years, ever since processing capacity caught up with neural models, have been heady times in the world of natural language processing. Neural approaches in general, and large, transformer language models in particular, have rapidly overtaken the leaderboards on a wide variety of benchmarks and once again the adage "there's no data like more data" seems to be true. It may seem like progress in the field, in fact, depends on the creation of ever larger language models (and research into how to deploy them to various ends).

In this paper, we have invited readers to take a step back and ask: Are ever larger language models inevitable or necessary? What costs are associated with this research direction and what should we consider before pursuing it? Do the field of NLP or the public that it serves in fact need larger language models? If so, how can we pursue this research direction while mitigating its associated risks? If not, what do we need instead?

We have identified a wide variety of costs and risks associated with the rush for ever larger language models, including: environmental costs (borne typically by those not benefiting from the resulting technology; §3); financial costs, which in turn erect barriers to entry, limiting who can contribute to this research area (§3); opportunity cost, as researchers pour effort away from directions requiring less resources (§5); and the risk of substantial harms, including stereotyping, denigration, increases in extremist ideology, and wrongful arrest, should humans encounter seemingly coherent language model output and take it for the words of some person or organization who have accountability for what is said (§6).

Thus, we call on NLP researchers to carefully weigh these risks while pursuing this research direction, consider whether the benefits outweigh the risks, and investigate dual use scenarios utilizing the many techniques (e.g. value sensitive design) that have been put forth. We hope these considerations encourage NLP researchers to direct resources and effort into techniques for approaching NLP tasks that are effective without being endlessly data hungry. But beyond that, we call on the field to recognize that tasks that aim to believably mimic humans bring risk of extreme harms. Work on synthetic human behavior is a "bright line" in ethical AI development, where downstream effects need to be understood and modeled in order to block foreseeable harm to society and different social groups. Thus what is also needed is scholarship on the benefits, harms and risks of mimicking humans, and thoughtful design

987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044

of target tasks grounded in use cases sufficiently concrete to allow collaborative design with affected communities.

REFERENCES

- [1] Hussein M Adam, Robert D Bullard, and Elizabeth Bell. 2001. *Faces of environmental racism: Confronting issues of global justice*. Rowman & Littlefield.
- [2] Larry Alexander. 1992. What makes wrongful discrimination wrong? Biases, preferences, stereotypes, and proxies. *University of Pennsylvania Law Review* 141, 1 (1992), 149–219.
- [3] American Psychological Association. 2019. Discrimination: What it is, and how to cope. <https://www.apa.org/topics/discrimination> (2019).
- [4] Dario Amodei and Daniel Hernandez. 2018. AI and Compute. <https://openai.com/blog/ai-and-compute/>
- [5] David Anthoff, Robert J Nicholls, and Richard SJ Tol. 2010. The economic impact of substantial sea-level rise. *Mitigation and Adaptation Strategies for Global Change* 15, 4 (2010), 321–335.
- [6] Alexei Baevski and Abdelrahman Mohamed. 2020. Effectiveness of Self-Supervised Pre-Training for ASR. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7694–7698.
- [7] Russel Barsh. 1990. Indigenous peoples, racism and the environment. *Meanjin* 49, 4 (1990), 723.
- [8] Christine Basta, Marta R Costa-jussà, and Noe Casas. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 33–39.
- [9] Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676* (2019).
- [10] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.
- [11] Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5185–5198. <https://doi.org/10.18653/v1/2020.acl-main.463>
- [12] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press, Cambridge, UK.
- [13] Steven Bird. 2016. Social Mobile Technologies for Reconnecting Indigenous and Immigrant Communities.. In *People.Policy.Place Seminar*. Northern Institute, Charles Darwin University, Darwin, Australia. <https://www.cdu.edu.au/sites/default/files/the-northern-institute/ppp-bird-20160128-4up.pdf>
- [14] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5454–5476. <https://doi.org/10.18653/v1/2020.acl-main.485>
- [15] Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large Language Models in Machine Translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, Prague, Czech Republic, 858–867. <https://www.aclweb.org/anthology/D07-1090>
- [16] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial Filters of Dataset Biases. In *Proceedings of the 37th International Conference on Machine Learning*.
- [17] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 1664–1674. <https://doi.org/10.18653/v1/D19-1176>
- [18] Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22, 6 (1996), 1482.
- [19] Robin Brewer and Anne Marie Piper. 2016. “Tell It Like It Really Is” A Case of Online Content Creation and Sharing Among Older Adult Bloggers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5529–5542.
- [20] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).
- [21] Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model Compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Philadelphia, PA, USA) (KDD '06)*. Association for Computing Machinery, New York, NY, USA, 535–541. <https://doi.org/10.1145/1150402.1150464>
- [22] Robert D Bullard. 1993. *Confronting environmental racism: Voices from the grassroots*. South End Press.
- [23] Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.
- [24] Herbert H Clark and Adrian Bangerter. 2004. Changing ideas about reference. In *Experimental pragmatics*. Springer, 25–49.
- [25] Herbert H. Clark and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding.
- [26] Herbert H. Clark, Robert Schreuder, and Samuel Buttrick. 1983. Common ground at the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior* 22, 2 (1983), 245 – 258. [https://doi.org/10.1016/S0022-5371\(83\)90189-5](https://doi.org/10.1016/S0022-5371(83)90189-5)
- [27] Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22, 1 (1986), 1 – 39. [https://doi.org/10.1016/0010-0277\(86\)90010-7](https://doi.org/10.1016/0010-0277(86)90010-7)
- [28] Benjamin Dangel. 2019. *The Five Hundred Year Rebellion: Indigenous Movements and the Decolonization of History in Bolivia*. AK Press.
- [29] Nicola De Cao, Wilker Aziz, and Ivan Titov. 2018. Question answering by reasoning across documents with graph convolutional networks. *arXiv preprint arXiv:1808.09920* (2018).
- [30] Ferdinand de Saussure. 1959. *Course in General Linguistics*. The Philosophical Society, New York. Translated by Wade Baskin.
- [31] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does object recognition work for everyone?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 52–59.
- [32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [33] Maeve Duggan. 2017. Online harassment 2017. (2017).
- [34] Ethan Fast, Tina Vachovsky, and Michael S Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. *arXiv preprint arXiv:1603.08832* (2016).
- [35] Anjalie Field, Doron Klinger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3570–3580. <https://doi.org/10.18653/v1/D18-1393>
- [36] Darja Fišer, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem, and Jacqueline Wernimont (Eds.). 2018. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, Brussels, Belgium. <https://www.aclweb.org/anthology/W18-5100>
- [37] Susan T Fiske. 2017. Prejudices in cultural contexts: shared stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion). *Perspectives on psychological science* 12, 5 (2017), 791–799.
- [38] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Srivivas, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *arXiv preprint arXiv:1802.00393* (2018).
- [39] Batya Friedman and David Hendry. 2012. The Envisioning Cards: A Toolkit for Catalyzing Humanistic and Technical Imaginations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Austin, Texas, USA) (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1145–1148. <https://doi.org/10.1145/2207676.2208562>
- [40] Batya Friedman and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.
- [41] Batya Friedman, Peter H Kahn, Jr., and Alan Borning. 2006. Value sensitive design and information systems. In *Human-Computer Interaction in Management Information Systems: Foundations*, P Zhang and D Galletta (Eds.). M. E. Sharpe, Armonk NY, 348–372.
- [42] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational AI. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1371–1374.
- [43] Guillem Gascó, Martha-Alicia Rocha, Germán Sanchis-Trilles, Jesús Andrés-Ferrer, and Francisco Casacuberta. 2012. Does more data always yield better translations?. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Avignon, France, 152–161. <https://www.aclweb.org/anthology/E12-1016>
- [44] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).
- [45] Sam Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language

- Models. In *Findings in EMNLP*.
- [46] Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus.
- [47] Wei Guo and Ailin Caliskan. 2020. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. *arXiv preprint arXiv:2006.03955* (2020).
- [48] Melissa Hart. 2004. Subjective decisionmaking and unconscious discrimination. *Ala. L. Rev.* 56 (2004), 741.
- [49] D. Hellman and Harvard University. 2008. *When is Discrimination Wrong?* Harvard University Press. https://books.google.com/books?id=M_ggFartkX4C
- [50] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *arXiv preprint arXiv:2002.05651* (2020).
- [51] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).
- [52] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5491–5501. <https://doi.org/10.18653/v1/2020.acl-main.487>
- [53] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 306–316.
- [54] Karen Spärck Jones. 2004. *Language modelling's generative model: Is it rational?* Technical Report. Citeseer.
- [55] David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 51–57.
- [56] Brendan Kennedy, Drew Kogon, Kris Coombs, Joseph Hoover, Christina Park, Gwenyth Portillo-Wightman, Aida Mostafazadeh Davani, Mohammad Atari, and Morteza Dehghani. 2018. A typology and coding manual for the study of hate-based rhetoric. *PsyArXiv*. July 18 (2018).
- [57] Gary Klein. 2007. Performing a project premortem. *Harvard business review* 85, 9 (2007), 18–19.
- [58] Roland Kuhn and Renato De Mori. 1990. A cache-based natural language model for speech recognition. *IEEE transactions on pattern analysis and machine intelligence* 12, 6 (1990), 570–583.
- [59] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. 166–172.
- [60] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).
- [61] Amanda Lazar, Mark Diaz, Robin Brewer, Chelsea Kim, and Anne Marie Piper. 2017. Going gray, failure to hire, and the ick factor: Analyzing how older bloggers talk about ageism. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 655–668.
- [62] Christopher A Le Dantec, Erika Shehan Poole, and Susan P Wyche. 2009. Values as lived experience: evolving value sensitive design in support of value discovery. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1141–1150.
- [63] Yinhao Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [64] Mette Edith Lundsryd. 2017. Speaking Back to a World of Checkpoints: Oral History as a Decolonizing Tool in the Study of Palestinian Refugees from Syria in Lebanon. (2017).
- [65] Marianna J Martindale and Marine Carpuat. 2018. Fluency over adequacy: a pilot study in measuring user trust in imperfect MT. *arXiv preprint arXiv:1802.06041* (2018).
- [66] Sally McConnell-Ginet. 1984. The Origins of Sexist Language in Discourse. *Annals of the New York Academy of Sciences* 433, 1 (1984), 123–135.
- [67] Kris McGuffie and Alex Newhouse. 2020. *The Radicalization Risks of GPT-3 and Advanced Neural Language Models*. Technical Report. Center on Terrorism, Extremism, and Counterterrorism, Middlebury Institute of International Studies at Monterrey. <https://www.middlebury.edu/institute/sites/www.middlebury.edu.institute/files/2020-09/gpt3-article.pdf>.
- [68] Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Berlin, Germany, 51–61. <https://doi.org/10.18653/v1/K16-1006>
- [69] Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A Framework for the Computational Linguistic Analysis of Dehumanization. *Frontiers Artif. Intell.* 3 (2020), 55. <https://doi.org/10.3389/frai.2020.00055>
- [70] Kaitlynn Mendes, Jessica Ringrose, and Jessalynn Keller. 2018. # MeToo and the promise and pitfalls of challenging rape culture through digital feminist activism. *European Journal of Women's Studies* 25, 2 (2018), 236–246.
- [71] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (Lake Tahoe, Nevada) (NIPS'13)*. Curran Associates Inc., Red Hook, NY, USA, 3111–3119.
- [72] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.
- [73] Robert C. Moore and William Lewis. 2010. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*. Association for Computational Linguistics, Uppsala, Sweden, 220–224. <https://www.aclweb.org/anthology/P10-2041>
- [74] Kevin L. Nadal. 2018. *Microaggressions and Traumatic Stress: Theory, Research, and Clinical Treatment*. American Psychological Association. <https://books.google.com/books?id=ogzshwEACAAJ>
- [75] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.
- [76] Lisa P. Nathan, Predrag V. Klasnja, and Batya Friedman. 2007. Value Scenarios: A Technique for Envisioning Systemic Effects of New Technologies. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2585–2590.
- [77] Timothy Niven and Hung-Yu Kao. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 4658–4664. <https://doi.org/10.18653/v1/P19-1459>
- [78] Charlotte Pennington, Derek Heim, Andrew Levy, and Derek Larkin. 2016. Twenty Years of Stereotype Threat Research: A Review of Psychological Mediators. *PLoS one* 11 (01 2016), e0146487. <https://doi.org/10.1371/journal.pone.0146487>
- [79] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. <https://doi.org/10.3115/v1/D14-1162>
- [80] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.
- [81] Pew. 2018. Internet/Broadband Fact Sheet. (2 2018). <https://www.pewinternet.org/fact-sheet/internet-broadband/>
- [82] Aidan Pine and Mark Turin. 2017. *Language Revitalization*. Oxford Research Encyclopedia of Linguistics.
- [83] Francesca Polletta. 1998. Contending stories: Narrative in social movements. *Qualitative sociology* 21, 4 (1998), 419–446.
- [84] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. *arXiv preprint arXiv:1910.04210* (2019).
- [85] Vinay Uday Prabhu and Abeba Birhane. 2020. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923* (2020).
- [86] Laura Pulido. 2016. Flint, environmental racism, and racial capitalism.
- [87] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.
- [88] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67.
- [89] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2383–2392. <https://doi.org/10.18653/v1/D16-1264>
- [90] Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem (Eds.). 2019. *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Florence, Italy. <https://www.aclweb.org/anthology/W19-3500>
- [91] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What we know about how BERT works. *arXiv:2002.12327 [cs.CL]*
- [92] Ronald Rosenfeld. 2000. Two decades of statistical language modeling: Where do we go from here? *Proc. IEEE* 88, 8 (2000), 1270–1278.
- [93] Corby Rosset. 2020. Turing-NLG: A 17-billion-parameter language model by Microsoft. *Microsoft Blog* (2020).
- [94] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).
- [95] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications

- of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5477–5490. <https://doi.org/10.18653/v1/2020.acl-main.486>
- [96] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green AI. [arXiv:1907.10597](https://arxiv.org/abs/1907.10597) [cs.CY]
- [97] Holger Schwenk, Anthony Rousseau, and Mohammed Attik. 2012. Large, Pruned or Continuous Space Language Models on a GPU for Statistical Machine Translation. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*. Association for Computational Linguistics, Montréal, Canada, 11–19. <https://www.aclweb.org/anthology/W12-2702>
- [98] Sabine Sczesny, Janine Bosak, Daniel Neff, and Birgit Schyns. 2004. Gender stereotypes and the attribution of leadership traits: A cross-cultural comparison. *Sex roles* 51, 11–12 (2004), 631–645.
- [99] Claude Elwood Shannon. 1949. The mathematical theory of communication.
- [100] Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. 2019. Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT. [arXiv:1909.05840](https://arxiv.org/abs/1909.05840) [cs.CL]
- [101] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3407–3412. <https://doi.org/10.18653/v1/D19-1339>
- [102] Katie Shilton, Jes A Koepfler, and Kenneth R Fleischmann. 2014. How to see values in social computing: methods for studying values dimensions. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 426–435.
- [103] Joonso Shin, Yoonhyung Lee, and Kyomin Jung. 2019. Effective Sentence Scoring Method Using BERT for Speech Recognition. In *Asian Conference on Machine Learning*. 1081–1093.
- [104] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using gpu model parallelism. [arXiv preprint arXiv:1909.08053](https://arxiv.org/abs/1909.08053) (2019).
- [105] Steven J. Spencer, Christine Logel, and Paul G. Davies. 2016. Stereotype Threat. *Annual Review of Psychology* 67, 1 (2016), 415–437. <https://doi.org/10.1146/annurev-psych-073115-103235> [arXiv:https://doi.org/10.1146/annurev-psych-073115-103235](https://arxiv.org/abs/https://doi.org/10.1146/annurev-psych-073115-103235) PMID: 26361054.
- [106] Katrina Srigley and Lorraine Sutherland. 2019. Decolonizing, Indigenizing, and Learning Biskaaybiyang in the Field: Our Oral History Journey1. *The Oral History Review* (2019).
- [107] Greg J. Stephens, Lauren J. Silbert, and Uri Hasson. 2010. Speaker–listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences* 107, 32 (2010), 14425–14430. <https://doi.org/10.1073/pnas.1008662107> [arXiv:https://www.pnas.org/content/107/32/14425.full.pdf](https://www.pnas.org/content/107/32/14425.full.pdf)
- [108] Ben G Streetman, Sanjay Banerjee, et al. 1995. *Solid state electronic devices*. Vol. 4. Prentice hall Englewood Cliffs, NJ.
- [109] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3645–3650.
- [110] Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2018. Improving machine reading comprehension with general reading strategies. [arXiv preprint arXiv:1810.13441](https://arxiv.org/abs/1810.13441) (2018).
- [111] Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*. 13230–13241.
- [112] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovered the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 4593–4601. <https://doi.org/10.18653/v1/P19-1452>
- [113] Marlon Twyman, Brian C Keegan, and Aaron Shaw. 2017. Black Lives Matter in Wikipedia: Collective memory and collaboration around online social movements. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1400–1412.
- [114] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
- [115] Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A Corpus for Studying Differential Responses to Gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. <https://www.aclweb.org/anthology/L18-1445>
- [116] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Brussels, Belgium, 353–355. <https://doi.org/10.18653/v1/W18-5446>
- [117] Zeerak Waseem, Thomas Davidson, Dana Warmusley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. [arXiv preprint arXiv:1705.09899](https://arxiv.org/abs/1705.09899) (2017).
- [118] Monnica T Williams. 2019. Psychology Cannot Afford to Ignore the Many Harms Caused by Microaggressions. *Perspectives on Psychological Science* 15 (2019), 38 – 43.
- [119] World Bank. 2018. Individuals Using the Internet. (2018). <https://data.worldbank.org/indicator/IT.NET.USER.ZS?end=2017&locations=US&start=2015>
- [120] Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE-GEN: An Enhanced Multi-Flow Pre-training and Fine-tuning Framework for Natural Language Generation. [arXiv preprint arXiv:2001.11314](https://arxiv.org/abs/2001.11314) (2020).
- [121] Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. 2020. Bert-of-theseus: Compressing bert by progressive module replacing. [arXiv preprint arXiv:2002.02925](https://arxiv.org/abs/2002.02925) (2020).
- [122] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*. 5753–5763.
- [123] Meg Young, Lassana Magassa, and Batya Friedman. 2019. Toward Inclusive Tech Policy Design: A Method for Underrepresented Voices to Strengthen Tech Policy Documents. *Ethics and Information Technology* (2019), 1–15.
- [124] Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8BERT: Quantized 8Bit BERT. [arXiv:1910.06188](https://arxiv.org/abs/1910.06188) [cs.CL]
- [125] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 93–104. <https://doi.org/10.18653/v1/D18-1009>
- [126] Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*. 110–120.
- [127] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 629–634. <https://doi.org/10.18653/v1/N19-1064>
- [128] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*. 19–27.