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

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained 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 provide 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**.

ACM Reference Format:
Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3442188.3445922

1 INTRODUCTION

One of the biggest trends in natural language processing (NLP) has been the increasing 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 [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], 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 [57]. 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 are discussing (outlined in §2), the first consideration should be the environmental cost.

Just as environmental impact scales with model size, so does the difficulty of understanding what is in the training data. In §4, we discuss how large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations. In collecting ever larger datasets we risk incurring documentation debt. We recommend mitigating these risks by budgeting for curation and documentation at the start of a project and only creating datasets as large as can be sufficiently documented.

As argued by Bender and Koller [14], it is important to understand the limitations of LMs 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, LMs are not performing natural language understanding (NLU), and only have success in tasks that can be approached by manipulating linguistic form [14]. Focusing on state-of-the-art results on leaderboards without encouraging deeper understanding of the mechanism by which they are achieved can cause misleading results as shown

*Joint first authors

This work is licensed under a Creative Commons Attribution International 4.0 License.

FAccT '21, March 3–10, 2021, Virtual Event, Canada
ACM ISBN 978-1-4503-8309-7/21/03.
https://doi.org/10.1145/3442188.3445922

in [21, 93] and direct resources away from efforts that would facilitate long-term progress towards natural language understanding, without using unfathomable training data.

Furthermore, the tendency of human interlocutors to impute meaning where there is none can mislead both NLP researchers and the general public into taking 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 encountering derogatory language and experiencing discrimination at the hands of others who reproduce racist, sexist, ableist, extremist or other harmful ideologies reinforced through interactions with synthetic language. We explore these potential harms in §6 and potential paths forward in §7.

We hope that a critical overview of the risks of relying on ever-increasing size of LMs as the primary driver of increased performance of language technology 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

Similar to [14], we understand the term *language model* (LM) 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 bidirectional and masked LMs) its surrounding context. Such systems are unsupervised and when deployed, take a text as input, commonly outputting scores or string predictions. Initially proposed by Shannon in 1949 [117], some of the earliest implemented LMs date to the early 1980s and were used as components in systems for automatic speech recognition (ASR), machine translation (MT), document classification, and more [111]. In this section, we provide a brief overview of the general trend of language modeling in recent years. For a more in-depth survey of pretrained LMs, see [105].

Before neural models, n-gram models also used large amounts of data [20, 87]. 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, [20] developed an n-gram model for English with a total of 1.8T n-grams and noted steady improvements in BLEU score on the test set of 1797 Arabic translations as the training data was increased from 13M tokens.

The next big step was the move towards using pretrained representations of the distribution of words (called *word embeddings*) in other (supervised) NLP tasks. These word vectors came from systems such as word2vec [85] and GloVe [98] and later LSTM models such as context2vec [82] and ELMo [99] and supported state of the art performance on question answering, textual entailment, semantic role labeling (SRL), coreference resolution, named entity recognition (NER), 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 labeled data necessary for training on the various supervised tasks. For example, [99] showed that a model trained with ELMo reduced the necessary amount of training data needed to achieve similar results on SRL 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. This 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. 82].

Year	Model	# of Parameters	Dataset Size
2019	BERT [39]	3.4E+08	16GB
2019	DistilBERT [113]	6.60E+07	16GB
2019	ALBERT [70]	2.23E+08	16GB
2019	XLNet (Large) [150]	3.40E+08	126GB
2020	ERNIE-Gen (Large) [145]	3.40E+08	16GB
2019	RoBERTa (Large) [74]	3.55E+08	161GB
2019	MegatronLM [122]	8.30E+09	174GB
2020	T5-11B [107]	1.10E+10	745GB
2020	T-NLG [112]	1.70E+10	174GB
2020	GPT-3 [25]	1.75E+11	570GB
2020	GShard [73]	6.00E+11	–
2021	Switch-C [43]	1.57E+12	745GB

Table 1: Overview of recent large language models

Transformer models, on the other hand, have been able to continuously benefit from larger architectures and larger quantities of data. Devlin et al. [39] in particular noted that training on a large dataset and fine-tuning for specific tasks leads to strictly increasing results on the GLUE tasks [138] for English as the hyperparameters of the model were increased. Initially developed as Chinese LMs, the ERNIE family [130, 131, 145] produced ERNIE-Gen, which was also trained on the original (English) BERT dataset, joining the ranks of very large LMs. NVIDIA released the MegatronLM which has 8.3B parameters and was trained on 174GB of text from the English Wikipedia, OpenWebText, RealNews and CC-Stories datasets [122]. Trained on the same dataset, Microsoft released T-NLG,[1] an LM with 17B parameters. OpenAI's GPT-3 [25] and Google's GShard [73] and Switch-C [43] have increased the definition of large LM by orders of magnitude in terms of parameters at 175B, 600B, and 1.6T parameters, respectively. Table 1 summarizes a selection of these LMs in terms of training data size and parameters. As increasingly large amounts of text are collected from the web in datasets such as the Colossal Clean Crawled Corpus [107] and the Pile [51], this trend of increasingly large LMs can be expected to continue as long as they correlate with an increase in performance.

A number of these models also have multilingual variants such as mBERT [39] and mT5 [148] or are trained with some amount of multilingual data such as GPT-3 where 7% of the training data was not in English [25]. The performance of these multilingual models across languages is an active area of research. Wu and Drezde [144] found that while mBERT does not perform equally well across all 104 languages in its training data, it performed better at NER, POS tagging, and dependency parsing than monolingual models trained with comparable amounts of data for four low-resource languages. Conversely, [95] surveyed monolingual BERT models developed with more specific architecture considerations or additional monolingual data and found that they generally outperform

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

mBERT across 29 tasks. Either way, these models do not address the inclusion problems raised by [65], who note that over 90% of the world's languages used by more than a billion people currently have little to no support in terms of language technology.

Alongside work investigating what information the models retain from the data, we see a trend in reducing the size of these models using various techniques such as knowledge distillation [26, 58], quantization [118, 153], factorized embedding parameterization and cross-layer parameter sharing [70], and progressive module replacing [146]. Rogers et al. [110] provide a comprehensive comparison of models derived from BERT using these techniques, such as DistilBERT [113] and ALBERT [70]. While these models maintain and sometimes exceed the performance of the original BERT model, despite their much smaller size, they ultimately still rely on large quantities of data and significant processing and storage capabilities to both hold and reduce the model.

We note that the change from n-gram LMs to word vectors distilled from neural LMs to pretrained Transformer LMs is paralleled by an expansion and change in the types of tasks they are useful for: n-gram LMs were initially typically deployed in selecting among the outputs of e.g. acoustical or translation models; 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 LMs in the sense we give above, 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 LMs 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 [129]. While the average human is responsible for an estimated 5t CO_2e per year,[2] the authors trained a Transformer (big) model [136] with neural architecture search and estimated that the training procedure emitted 284t of CO_2. Training a single BERT base model (without hyperparameter tuning) on GPUs was estimated to require as much energy as a trans-American flight.

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. In addition, renewable energy sources are still costly to the environment,[3] and data centers with increasing computation requirements take away from other potential uses of green energy,[4] underscoring the need for energy efficient model architectures and training paradigms.

Strubell et al. also examine the cost of these models vs. their accuracy gains. For the task of machine translation where large LMs have resulted in performance gains, they estimate that an increase in 0.1 BLEU score using neural architecture search for English to German translation results in an increase of $150,000 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. They also urge governments to invest in compute clouds to provide equitable access to researchers.

Initiatives such as the SustainNLP workshop[5] have since taken up the goal of prioritizing computationally efficient hardware and algorithms. Schwartz et al. [115] also call for the development of green AI, similar to other environmentally friendly scientific developments such as green chemistry or sustainable computing. As shown in [5], the amount of compute used to train the largest deep learning models (for NLP and other applications) has increased 300,000x in 6 years, increasing at a far higher pace than Moore's Law. To promote green AI, Schwartz et al. argue for promoting efficiency as an evaluation metric and show that most sampled papers from ACL 2018, NeurIPS 2018, and CVPR 2019 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 [57, 75] have released online tools to help researchers benchmark their energy usage. Among their recommendations are to run experiments in carbon friendly regions, consistently report energy and carbon metrics, and consider energy-performance trade-offs before deploying energy hungry models. In addition to these calls for documentation and technical fixes, Bietti and Vatanparast underscore the need for social and political engagement in shaping a future where data driven systems have minimal negative impact on the environment [16].

While [129] benchmarks the training process in a research setting, many LMs are deployed in industrial or other settings where the cost of inference might greatly outweigh that of training in the long run. In this scenario, it may be more appropriate to deploy models with lower energy costs during inference even if their training costs are high. In addition to benchmarking tools, works estimating the cost increase associated with the introduction of LMs for particular applications, and how they compare to alternative NLP methods, will be important for understanding the trade-offs.

When we perform risk/benefit analyses of language technology, we must keep 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 first [1, 27].[6] Is it fair or just to ask, for example, that the residents of the Maldives (likely to be underwater by 2100 [6]) or the 800,000 people in Sudan affected

[2] Data for 2017, from https://ourworldindata.org/co2-emissions, accessed Jan 21, 2021
[3] https://www.heraldscotland.com/news/18270734.14m-trees-cut-scotland-make-way-wind-farms/
[4] https://news.microsoft.com/2017/11/02/microsoft-announces-one-of-the-largest-wind-deals-in-the-netherlands-with-vattenfall/
[5] https://sites.google.com/view/sustainlp2020/organization
[6] https://www.un.org/sustainabledevelopment/blog/2016/10/report-inequalities-exacerbate-climate-impacts-on-poor/

by drastic floods[7] pay the environmental price of training and deploying ever larger English LMs, when similar large-scale models aren't being produced for Dhivehi or Sudanese Arabic?[8]

And, while some language technology is genuinely designed to benefit marginalized communities [17, 101], most language technology is built 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 LMs encode and reinforce hegemonic biases (see §§4 and 6), the harms that follow are most likely to fall on marginalized populations who, even in rich nations, are most likely to experience environmental racism [10, 104].

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 India,[9] to the worst fire season on record in Australia killing or displacing nearly three billion animals and at least 400 people,[10] 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 application areas, the training data has been shown to have problematic characteristics [18, 38, 42, 47, 61] resulting in models that encode stereotypical and derogatory associations along gender, race, ethnicity, and disability status [11, 12, 69, 69, 132, 132, 157]. 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 Size Doesn't Guarantee Diversity

The Internet is a large and diverse virtual space, and accordingly, it is easy to imagine that very large datasets, such as Common Crawl ("petabytes of data collected over 8 years of web crawling",[11] a filtered version of which is included in the GPT-3 training data) must therefore be broadly representative of the ways in which different people view the world. However, on closer examination, we find that there are several factors which narrow Internet participation, the discussions which will be included via the crawling methodology, and finally the texts likely to be contained after the crawled data are filtered. In all cases, the voices of people most likely to hew to a hegemonic viewpoint are also more likely to be retained. In the case of US and UK English, this means that white supremacist and misogynistic, ageist, etc. views are overrepresented in the training data, not only exceeding their prevalence in the general population but also setting up models trained on these datasets to further amplify biases and harms.

Starting with who is contributing to these Internet text collections, we see that Internet access itself is not evenly distributed, resulting in Internet data overrepresenting younger users and those from developed countries [100, 143].[12] However, it's not just the Internet as a whole that is in question, but rather specific subsamples of it. For instance, GPT-2's training data 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 29.[13] Similarly, recent surveys of Wikipedians find that only 8.8–15% are women or girls [9].

Furthermore, while user-generated content sites like Reddit, Twitter, and Wikipedia present themselves as open and accessible to anyone, there are structural factors including moderation practices which make them less welcoming to marginalized populations. Jones [64] documents (using digital ethnography techniques [63]) multiple cases where people on the receiving end of death threats on Twitter have had their accounts suspended while the accounts issuing the death threats persist. She further reports that harassment on Twitter is experienced by "a wide range of overlapping groups including domestic abuse victims, sex workers, trans people, queer people, immigrants, medical patients (by their providers), neurodivergent people, and visibly or vocally disabled people." The net result is that 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 communication, creating a feedback loop that lessens the impact of data from underrepresented populations.

Even if populations who feel unwelcome in mainstream sites set up different fora for communication, these may be less likely to be included in training data for language models. 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 [71], which some older adults prefer over more popular social media sites for discussing sensitive topics [24]. These fora contain rich discussions about what constitutes age discrimination and the impacts thereof. However, 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.

Finally, the current practice of filtering datasets can further attenuate the voices of people from marginalized identities. The training set for GPT-3 was a filtered version of the Common Crawl dataset, developed by training a classifier to pick out those documents

[7] https://www.aljazeera.com/news/2020/9/25/over-800000-affected-in-sudan-flooding-un

[8] By this comment, we do not intend to erase existing work on low-resource languages. One particularly exciting example is the Masakhane project [91], which explores participatory research techniques for developing MT for African languages. These promising directions do not involve amassing terabytes of data.

[9] https://www.voanews.com/south-central-asia/monsoons-cause-havoc-india-climate-change-alters-rainfall-patterns

[10] https://www.cnn.com/2020/07/28/asia/australia-fires-wildlife-report-scli-intl-scn/index.html

[11] http://commoncrawl.org/

[12] This point is also mentioned in the model card for GPT-3: https://github.com/openai/gpt-3/blob/master/model-card.md

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

most similar to the ones used in GPT-2's training data, i.e. documents linked to from Reddit [25], plus Wikipedia and a collection of books. While this was reportedly effective at filtering out documents that previous work characterized as "unintelligible" [134], what is unmeasured (and thus unknown) is what else it filtered out. The Colossal Clean Crawled Corpus [107], used to train a trillion parameter LM in [43], is cleaned, *inter alia*, by discarding any page containing one of a list of about 400 "Dirty, Naughty, Obscene or Otherwise Bad Words" [p.6].[14] This list is overwhelmingly words related to sex, with a handful of racial slurs and words related to white supremacy (e.g. *swastika*, *white power*) included. While possibly effective at removing documents containing pornography (and the associated problematic stereotypes encoded in the language of such sites [125]) and certain kinds of hate speech, this approach will also undoubtedly attenuate, by suppressing such words as *twink*, the influence of online spaces built by and for LGBTQ people.[15] If we filter out the discourse of marginalized populations, we fail to provide training data that reclaims slurs and otherwise describes marginalized identities in a positive light.

Thus at each step, from initial participation in Internet fora, to continued presence there, to the collection and finally the filtering of training data, current practice privileges the hegemonic viewpoint. In accepting large amounts of web text as 'representative' of 'all' of humanity we risk perpetuating dominant viewpoints, increasing power imbalances, and further reifying inequality. We instead propose practices that actively seek to include communities underrepresented on the Internet. For instance, one can take inspiration from movements to decolonize education by moving towards oral histories due to the overrepresentation of colonial views in text [35, 76, 127], and curate training datasets through a thoughtful process of deciding what to put in, rather than aiming solely for scale and trying haphazardly to weed out, post-hoc, flotsam deemed 'dangerous', 'unintelligible', or 'otherwise bad'.

4.2 Static Data/Changing Social Views

A central aspect of social movement formation involves using language strategically to destabilize dominant narratives and call attention to underrepresented social perspectives. Social movements produce new norms, language, and ways of communicating. This adds challenges to the deployment of LMs, as methodologies reliant on LMs run the risk of 'value-lock', where the LM-reliant technology reifies older, less-inclusive understandings.

For instance, the Black Lives Matter movement (BLM) influenced Wikipedia article generation and editing such that, as the BLM movement grew, articles covering shootings of Black people increased in coverage and were generated with reduced latency [135]. 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 [102]. More generally, Twyman et al. [135] highlight how social movements actively influence framings and reframings of minority narratives in the type of online discourse that potentially forms the data that underpins LMs.

An important caveat is that social movements which are poorly documented and which do not receive significant media attention will not be captured at all. Media coverage can fail to cover protest events and social movements [41, 96] and can distort events that challenge state power [36]. This is exemplified by media outlets that tend to ignore peaceful protest activity and instead focus on dramatic or violent events that make for good television but nearly always result in critical coverage [81]. As a result, the data underpinning LMs stands to misrepresent social movements and disproportionately align with existing regimes of power.

Developing and shifting frames stand to be learned in incomplete ways or lost in the big-ness of data used to train large LMs — particularly if the training data isn't continually updated. Given the compute costs alone of training large LMs, it likely isn't feasible for even large corporations to fully retrain them frequently enough to keep up with the kind of language change discussed here. Perhaps fine-tuning approaches could be used to retrain LMs, but here again, what would be required is thoughtful curation practices to find appropriate data to capture reframings and techniques for evaluating whether such fine-tuning appropriately captures the ways in which new framings contest hegemonic representations.

4.3 Encoding Bias

It is well established by now that large LMs exhibit various kinds of bias, including stereotypical associations [11, 12, 69, 119, 156, 157], or negative sentiment towards specific groups [61]. Furthermore, we see the effects of intersectionality [34], where BERT, ELMo, GPT and GPT-2 encode more bias against identities marginalized along more than one dimension than would be expected based on just the combination of the bias along each of the axes [54, 132]. Many of these works conclude that these issues are a reflection of training data characteristics. For instance, Hutchinson et al. find that BERT associates phrases referencing persons with disabilities with more negative sentiment words, and that gun violence, homelessness, and drug addiction are overrepresented in texts discussing mental illness [61]. Similarly, Gehman et al. show that models like GPT-3 trained with at least 570GB of data derived mostly from Common Crawl[16] can generate sentences with high toxicity scores even when prompted with non-toxic sentences [53]. Their investigation of GPT-2's training data[17] also finds 272K documents from unreliable news sites and 63K from banned subreddits.

These demonstrations of biases learned by LMs are extremely valuable in pointing out the potential for harm when such models are deployed, either in generating text or as components of classification systems, as explored further in §6. However, they do not represent a methodology that can be used to exhaustively discover all such risks, for several reasons.

First, model auditing techniques typically rely on automated systems for measuring sentiment, toxicity, or novel metrics such as 'regard' to measure attitudes towards a specific demographic group [119]. But these systems themselves may not be reliable

[14] Available at https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en, accessed Jan 18, 2021
[15] This observation is due to William Agnew.

[16] https://commoncrawl.org/the-data/
[17] GPT-3's training data is not openly available, but GPT-2's training data was used indirectly to construct GPT-3's [53].

means of measuring the toxicity of text generated by LMs. For example, the Perspective API model has been found to associate higher levels of toxicity with sentences containing identity markers for marginalized groups or even specific names [61, 103].

Second, auditing an LM for biases requires an *a priori* understanding of what social categories might be salient. The works cited above generally start from US protected attributes such as race and gender (as understood within the US). But, of course, protected attributes aren't the only identity characteristics that can be subject to bias or discrimination, and the salient identity characteristics and expressions of bias are also culture-bound [46, 116]. Thus, components like toxicity classifiers would need culturally appropriate training data for each context of audit, and even still we may miss marginalized identities if we don't know what to audit for.

Finally, we note that moving beyond demonstrating the existence of bias to building systems that verify the 'safety' of some LM (even for a given protected class) requires engaging with the systems of power that lead to the harmful outcomes such a system would seek to prevent [19]. For example, the #MeToo movement has spurred broad-reaching conversations about inappropriate sexual behavior from men in power, as well as men more generally [84]. These conversations challenge behaviors that have been historically considered appropriate or even the fault of women, shifting notions of sexually inappropriate behavior. 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 [40]. 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 LMs, or the biases encoded by a system, always needs to be done in relation to particular social contexts and marginalized perspectives [19].

4.4 Curation, Documentation & Accountability

In summary, LMs trained on large, uncurated, static datasets from the Web encode hegemonic views that are harmful to marginalized populations. We thus emphasize the need to invest significant resources into curating and documenting LM training data. In this, we follow Jo et al. [62], who cite archival history data collection methods as an example of the amount of resources that should be dedicated to this process, and Birhane and Prabhu [18], who call for a more justice-oriented data collection methodology. Birhane and Prabhu note, echoing Ruha Benjamin [15], "Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy." [p.1541]

When we rely on ever larger datasets we risk incurring *documentation debt*,[18] i.e. putting ourselves in a situation where the datasets are both undocumented and too large to document post hoc. While documentation allows for potential accountability [13, 52, 86], undocumented training data perpetuates harm without recourse. Without documentation, one cannot try to understand training data characteristics in order to mitigate some of these attested issues or even unknown ones. The solution, we propose, is to budget for documentation as part of the planned costs of dataset creation, and only collect as much data as can be thoroughly documented within that budget.

[18] On the notion of documentation debt as applied to code, rather than data, see [154].

5 DOWN THE GARDEN PATH

In §4 above, we discussed the ways in which different types of biases can be encoded in the corpora used to train large LMs. 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, specifically around the application of LMs to tasks intended to test for natural language understanding (NLU). As the very large Transformer LMs posted striking gains in the state of the art on various benchmarks intended to model meaning-sensitive tasks, and as initiatives like [142] made the models broadly accessible to researchers seeking to apply them, large quantities of research effort turned towards measuring how well BERT and its kin do on both existing and new benchmarks.[19] This allocation of research effort brings with it an opportunity cost, on the one hand in terms of time not spent applying meaning capturing approaches to meaning sensitive tasks, and on the other hand in terms of time not spent exploring more effective ways of building technology with datasets of a size that can be carefully curated and available for a broader set of languages [65, 91].

The original BERT paper [39] showed the effectiveness of the architecture and the pretraining technique by evaluating on the General Language Understanding Evaluation (GLUE) benchmark [138], the Stanford Question Answering Datasets (SQuAD 1.1 and 2.0) [108], and the Situations With Adversarial Generations benchmark (SWAG) [155], 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." [39, p.4179]. Even before [39] was published, BERT was picked up by the NLP community and applied with great success to a wide variety of tasks [e.g. 2, 149].

However, no actual language understanding is taking place in LM-driven approaches to these tasks, as can be shown by careful manipulation of the test data to remove spurious cues the systems are leveraging [21, 93]. Furthermore, as Bender and Koller [14] argue from a theoretical perspective, languages are systems of signs [37], i.e. pairings of form and meaning. But the training data for LMs is only form; they do not have access to meaning. Therefore, claims about model abilities must be carefully characterized.

As the late Karen Spärck Jones pointed out: the use of LMs ties us to certain (usually unstated) epistemological and methodological commitments [124]. 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 LMs 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.[20] Although

[19] ~26% of papers sampled from ACL, NAACL and EMNLP since 2018 cite [39].

[20] Specifically, that the mutual information between the input and the meaning given the output is zero — what Spärck Jones calls "the model of ignorance".

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-3 and their kin are learning about linguistic structure from the unsupervised language modeling task, as studied in the emerging field of 'BERTology' [e.g. 110, 133]. 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 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 LMs to amplify biases and other issues in the training data, and the tendency of researchers and other people to mistake LM-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. We focus primarily on cases where LMs are used in generating text, but we will also touch on risks that arise when LMs or word embeddings derived from them are components of systems for classification, query expansion, or other tasks, or when users can query LMs for information memorized from their training data.

6.1 Coherence in the Eye of the Beholder

Where traditional n-gram LMs [117] can only model relatively local dependencies, predicting each word given the preceding sequence of N words (usually 5 or fewer), the Transformer LMs 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 [80] prompted GPT-3 with the text in bold in Figure 1, and it produced the rest of the text, including the Q&A format.[21] This example illustrates GPT-3's ability to produce coherent and on-topic text; the topic is connected to McGuffie and Newhouse's study of GPT-3 in the context of extremism, discussed below.

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' beliefs [30, 31] and intentions [23, 33] within context [32]. That is, human language use

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.

Figure 1: GPT-3's response to the prompt (in bold), from [80]

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 [29, 128] is most clearly true in co-situated spoken or signed communication, but we use the same facilities for producing language 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. It must follow that even when we don't know the person who generated the language we are interpreting, we build a partial model of who they are and what common ground we think they share with us, and use this in interpreting their words.

Text generated by an LM 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 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 [89, 140]. 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).[22] Contrary

[21] This is just the first part of the response that McGuffie and Newhouse show. GPT-3 continues for two more question answer pairs with similar coherence. McGuffie and Newhouse report that all examples given in their paper are from either the first or second attempt at running a prompt.

[22] Controlled generation, where an LM is deployed within a larger system that guides its generation of output to certain styles or topics [e.g. 147, 151, 158], is not the same thing as communicative intent. One clear way to distinguish the two is to ask whether

to how it may seem when we observe its output, an LM 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 LMs 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, laying out the potential follow-on harms.

The first risks we consider are the risks that follow from the LMs absorbing the hegemonic worldview from their training data. When humans produce language, our utterances reflect our worldviews, including our biases [78, 79]. As people in positions of privilege with respect to a society's racism, misogyny, ableism, etc., tend to be overrepresented 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 [45, 109]. Furthermore, in addition to abusive language [139] and hate speech [67], there are subtler forms of negativity such as gender bias [137], microaggressions [22], dehumanization [83], and various socio-political framing biases [44, 114] that are prevalent in language data. For example, describing a woman's account of her experience of sexism with the word *tantrum* both reflects a worldview where the sexist actions are normative and foregrounds a stereotype of women as childish and not in control of their emotions.

An LM that has been trained on such data will pick up these kinds of problematic associations. If such an LM 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 LMs producing text will reproduce and even amplify the biases in their input [53]. Thus the risk is that people disseminate text generated by LMs, meaning more text in the world that reinforces and propagates stereotypes and problematic associations, both to humans who encounter the text and to future LMs 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 [88, 141] and stereotype threat [97, 126]. Other readers may be introduced to stereotypes or have ones they already carry reinforced, leading them to engage in discrimination (consciously or not) [55], which in turn leads to harms of subjugation, denigration, belittlement, loss of opportunity [3, 4, 56] and others on the part of those discriminated against.

If the LM outputs overtly abusive language (as Gehman et al. [53] 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.

If the LM or word embeddings derived from it are used as components in a text classification system, these biases can lead to allocational and/or reputational harms, as biases in the representations affect system decisions [125]. This case is especially pernicious for being largely invisible to both the direct user of the system and any indirect stakeholders about whom decisions are being made. Similarly, biases in an LM used in query expansion could influence search results, further exacerbating the risk of harms of the type documented by Noble in [94], where the juxtaposition of search queries and search results, when connected by negative stereotypes, reinforce those stereotypes and cause psychological harm.

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. These include prosaic cases, such as services set up to 'automatically' write term papers or interact on social media,[23] as well as use cases connected to promoting extremism. For example, McGuffie and Newhouse [80] 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, in the first instance, befall the people so recruited and (likely more severely) to others as a result of violence carried out by the extremists.

Yet another risk connected to seeming coherence and fluency involves machine translation (MT) and the way that increased fluency of MT output changes the perceived adequacy of that output [77]. This differs somewhat from the cases above in that there was an initial human communicative intent, by the author of the source language text. However, MT systems can (and frequently do) produce output that is inaccurate yet both fluent and (again, seemingly)

the system (or the organization deploying the system) has accountability for the truth of the utterances produced.

[23] Such as the GPT-3 powered bot let loose on Reddit; see https://thenextweb.com/neural/2020/10/07/someone-let-a-gpt-3-bot-loose-on-reddit-it-didnt-end-well/amp/.

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).[24] 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 [77].

Finally, we note that there are risks associated with the fact that LMs with extremely large numbers of parameters model their training data very closely and can be prompted to output specific information from that training data. For example, [28] demonstrate a methodology for extracting personally identifiable information (PII) from an LM and find that larger LMs are more susceptible to this style of attack than smaller ones. Building training data out of publicly available documents doesn't fully mitigate this risk: just because the PII was already available in the open on the Internet doesn't mean there isn't additional harm in collecting it and providing another avenue to its discovery. This type of risk differs from those noted above because it doesn't hinge on seeming coherence of synthetic text, but the possibility of a sufficiently motivated user gaining access to training data via the LM. In a similar vein, users might query LMs for 'dangerous knowledge' (e.g. tax avoidance advice), knowing that what they were getting was synthetic and therefore not credible but nonetheless representing clues to what is in the training data in order to refine their own search queries.

6.3 Summary

In this section, we have discussed how the human tendency to attribute meaning to text, in combination with large LMs' ability 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. We have also reviewed risks connected to using LMs as components in classification systems and the risks of LMs memorizing training data. We note that the risks associated with synthetic but seemingly coherent text are deeply connected to the fact that such synthetic text can enter into conversations without any person or entity being accountable for it. This accountability both involves responsibility for truthfulness and is important in situating meaning. As Maggie Nelson [92] writes: "Words change depending on who speaks them; there is no cure."

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 LMs, 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 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 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 Birhane and Prabhu's [18] words (inspired by Ruha Benjamin [15]): "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 [13, 52, 86] to describe the uses for which their models are suited and benchmark evaluations for a variety of conditions. This involves providing thorough documentation on the data used in model building, including the motivations underlying data selection and collection processes. This documentation should reflect and indicate researchers' goals, values, and motivations in assembling data and creating a given model. It should also make note of potential users and stakeholders, particularly those that stand to be negatively impacted by model errors or misuse. We note that just because a model might have many different applications doesn't mean that its developers don't need to consider stakeholders. An exploration of stakeholders for likely use cases can still be informative around potential risks, even when there is no way to guarantee that all use cases can be explored.

We also advocate for a re-alignment of research goals: Where much effort has been allocated to making models (and their training data) bigger and to achieving ever higher scores on leaderboards often featuring artificial tasks, we believe there is more to be gained by focusing on understanding how machines are achieving the tasks in question and how they will form part of socio-technical systems. To that end, LM development may benefit from guided evaluation exercises such as pre-mortems [68]. 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.[25] 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 LMs in a guided way while also considering fixes to current designs or alternative

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

[25] This would be one way to build a evaluation culture that considers not only average-case performance (as measured by metrics) and best-case performance (cherry-picked examples), but also worst-case performance.

methods of achieving a task-oriented goal in relation to specific pitfalls.

Value sensitive design [49, 50] 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 designing systems that support those values. These include such techniques as envisioning cards [48], the development of value scenarios [90], and working with panels of experiential experts [152]. These approaches help surface not only stakeholder values, but also values expressed by systems and enacted through interactions between systems and society [120]. For researchers working with LMs, value sensitive design is poised to help 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 [72]. 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 LMs that have specifically served marginalized populations. If, as we advocate, the field backs off from the path of ever larger LMs, are we thus sacrificing benefits that would accrue to these populations? As a case in point, consider automatic speech recognition, which has seen some improvements thanks to advances in LMs, including both in size and in architecture [e.g. 8, 59, 121], though the largest LMs typically are too large and too slow for the near real-time needs of ASR systems [60]. 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.[26] We see two beneficial paths forward here: The first is a broader search for means of improving ASR systems, as indeed is underway, since the contexts of application of the technology aren't conducive to using ever larger LMs [60]. But even if larger LMs could be used, just because we've seen that large LMs 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 LMs are critical (when available), is to recognize this as an instance of a dual use problem and consider how to mitigate the harms of LMs used as stochastic parrots while still preserving them for use in ASR systems. Could LMs be built in such a way that synthetic text generated with them would be watermarked and thus detectable [7, 66, 123]? Are there policy approaches that could effectively regulate their use?

In summary, we advocate for research that centers the people who stand to be adversely 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, for exploring multiple possible paths towards long-term goals, for keeping alert to dual-use scenarios, and finally for 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 NLP. Neural approaches in general, and large, Transformer LMs 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 LMs 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 LMs? 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 LMs, including: environmental costs (borne typically by those not benefiting from the resulting technology); financial costs, which in turn erect barriers to entry, limiting who can contribute to this research area and which languages can benefit from the most advanced techniques; opportunity cost, as researchers pour effort away from directions requiring less resources; and the risk of substantial harms, including stereotyping, denigration, increases in extremist ideology, and wrongful arrest, should humans encounter seemingly coherent LM output and take it for the words of some person or organization who has accountability for what is said.

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

[26] Note 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.

REFERENCES

[1] Hussein M Adam, Robert D Bullard, and Elizabeth Bell. 2001. *Faces of environmental racism: Confronting issues of global justice.* Rowman & Littlefield.

[2] Chris Alberti, Kenton Lee, and Michael Collins. 2019. A BERT Baseline for the Natural Questions. arXiv:1901.08634 [cs.CL]

[3] Larry Alexander. 1992. What makes wrongful discrimination wrong? Biases, preferences, stereotypes, and proxies. *University of Pennsylvania Law Review* 141, 1 (1992), 149–219.

[4] American Psychological Association. 2019. Discrimination: What it is, and how to cope. *https://www.apa.org/topics/discrimination* (2019).

[5] Dario Amodei and Daniel Hernandez. 2018. AI and Compute. https://openai.com/blog/ai-and-compute/

[6] 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.

[7] Mikhail J Atallah, Victor Raskin, Christian F Hempelmann, Mercan Karahan, Radu Sion, Umut Topkara, and Katrina E Triezenberg. 2002. Natural Language Watermarking and Tamperproofing. In *International Workshop on Information Hiding*. Springer, 196–212.

[8] 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.

[9] Michael Barera. 2020. Mind the Gap: Addressing Structural Equity and Inclusion on Wikipedia. (2020). Accessible at http://hdl.handle.net/10106/29572.

[10] Russel Barsh. 1990. Indigenous peoples, racism and the environment. *Meanjin* 49, 4 (1990), 723.

[11] 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.

[12] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 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, 3615–3620. https://doi.org/10.18653/v1/D19-1371

[13] 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.

[14] 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

[15] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press, Cambridge, UK.

[16] Elettra Bietti and Roxana Vatanparast. 2020. Data Waste. *Harvard International Law Journal* 61 (2020).

[17] 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

[18] Abeba Birhane and Vinay Uday Prabhu. 2021. Large Image Datasets: A Pyrrhic Win for Computer Vision?. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1537–1547.

[19] 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

[20] 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

[21] 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*.

[22] 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

[23] 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.

[24] 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.

[25] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[26] Cristian Buciluǎ, 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

[27] Robert D Bullard. 1993. *Confronting environmental racism: Voices from the grassroots*. South End Press.

[28] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting Training Data from Large Language Models. arXiv:2012.07805 [cs.CR]

[29] Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

[30] Herbert H. Clark and Adrian Bangerter. 2004. Changing ideas about reference. In *Experimental Pragmatics*. Springer, 25–49.

[31] Herbert H. Clark and Meredyth A Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language* 50, 1 (2004), 62–81.

[32] 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

[33] 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

[34] Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum* (1989), 139.

[35] Benjamin Dangl. 2019. *The Five Hundred Year Rebellion: Indigenous Movements and the Decolonization of History in Bolivia*. AK Press.

[36] Christian Davenport. 2009. *Media bias, perspective, and state repression: The Black Panther Party*. Cambridge University Press.

[37] Ferdinand de Saussure. 1959. *Course in General Linguistics*. The Philosophical Society, New York. Translated by Wade Baskin.

[38] 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.

[39] 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

[40] Maeve Duggan. 2017. *Online Harassment 2017*. Pew Research Center.

[41] Jennifer Earl, Andrew Martin, John D. McCarthy, and Sarah A. Soule. 2004. The use of newspaper data in the study of collective action. *Annual Review of Sociology* 30 (2004), 65–80.

[42] Ethan Fast, Tina Vachovsky, and Michael Bernstein. 2016. Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 10.

[43] William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv:2101.03961 [cs.LG]

[44] Anjalie Field, Doron Kliger, 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

[45] 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

[46] 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.

[47] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[48] 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

[49] Batya Friedman and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.

[50] 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.

[51] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. arXiv:2101.00027 [cs.CL]

[52] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. arXiv:1803.09010 [cs.DB]

[53] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

[54] 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* (2020).

[55] Melissa Hart. 2004. Subjective decisionmaking and unconscious discrimination. *Alabama Law Review* 56 (2004), 741.

[56] Deborah Hellman. 2008. *When is Discrimination Wrong?* Harvard University Press.

[57] 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. *Journal of Machine Learning Research* 21, 248 (2020), 1–43. http://jmlr.org/papers/v21/20-312.html

[58] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[59] Chao-Wei Huang and Yun-Nung Chen. 2019. Adapting Pretrained Transformer to Lattices for Spoken Language Understanding. In *Proceedings of 2019 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2019)*. Sentosa, Singapore, 845–852.

[60] Hongzhao Huang and Fuchun Peng. 2019. An Empirical Study of Efficient ASR Rescoring with Transformers. arXiv:1910.11450 [cs.CL]

[61] 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

[62] 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.

[63] Leslie Kay Jones. 2020. #BlackLivesMatter: An Analysis of the Movement as Social Drama. *Humanity & Society* 44, 1 (2020), 92–110.

[64] Leslie Kay Jones. 2020. Twitter wants you to know that you're still SOL if you get a death threat — unless you're President Donald Trump. (2020). https://medium.com/@agua.carbonica/twitter-wants-you-to-know-that-youre-still-sol-if-you-get-a-death-threat-unless-you-re-a5cce316b706.

[65] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

[66] Nurul Shamimi Kamaruddin, Amirrudin Kamsin, Lip Yee Por, and Hameedur Rahman. 2018. A Review of Text Watermarking: Theory, Methods, and Applications. *IEEE Access* 6 (2018), 8011–8028. https://doi.org/10.1109/ACCESS.2018.2796585

[67] 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).

[68] Gary Klein. 2007. Performing a project premortem. *Harvard business review* 85, 9 (2007), 18–19.

[69] 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.

[70] 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).

[71] 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.

[72] 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.

[73] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. arXiv:2006.16668 [cs.CL]

[74] Yinhan 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).

[75] Kadan Lottick, Silvia Susai, Sorelle A. Friedler, and Jonathan P. Wilson. 2019. Energy Usage Reports: Environmental awareness as part of algorithmic accountability. arXiv:1911.08354 [cs.LG]

[76] Mette Edith Lundsfryd. 2017. Speaking Back to a World of Checkpoints: Oral History as a Decolonizing Tool in the Study of Palestinian Refugees from Syria in Lebanon. *Middle East Journal of Refugee Studies* 2, 1 (2017), 73–95.

[77] Marianna Martindale and Marine Carpuat. 2018. Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*. Association for Machine Translation in the Americas, Boston, MA, 13–25. https://www.aclweb.org/anthology/W18-1803

[78] Sally McConnell-Ginet. 1984. The Origins of Sexist Language in Discourse. *Annals of the New York Academy of Sciences* 433, 1 (1984), 123–135.

[79] Sally McConnell-Ginet. 2020. *Words Matter: Meaning and Power*. Cambridge University Press.

[80] 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.

[81] Douglas M McLeod. 2007. News coverage and social protest: How the media's protect paradigm exacerbates social conflict. *Journal of Dispute Resolution* (2007), 185.

[82] 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

[83] Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A Framework for the Computational Linguistic Analysis of Dehumanization. *Frontiers in Artificial Intelligence* 3 (2020), 55. https://doi.org/10.3389/frai.2020.00055

[84] 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.

[85] 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.

[86] 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.

[87] 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

[88] Kevin L. Nadal. 2018. *Microaggressions and Traumatic Stress: Theory, Research, and Clinical Treatment*. American Psychological Association. https://books.google.com/books?id=ogzhswEACAAJ

[89] 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.

[90] 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.

[91] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2144–2160. https://doi.org/10.18653/v1/2020.findings-emnlp.195

[92] Maggie Nelson. 2015. *The Argonauts*. Graywolf Press, Minneapolis.

[93] 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

[94] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

[95] Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [MASK]? Making Sense of Language-Specific BERT Models. arXiv:2003.02912 [cs.CL]

[96] David Ortiz, Daniel Myers, Eugene Walls, and Maria-Elena Diaz. 2005. Where do we stand with newspaper data? *Mobilization: An International Quarterly* 10, 3 (2005), 397–419.

[97] 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

[98] 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

[99] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237. https://doi.org/10.18653/v1/N18-1202

[100] Pew. 2018. Internet/Broadband Fact Sheet. (2 2018). https://www.pewinternet.org/fact-sheet/internet-broadband/

[101] Aidan Pine and Mark Turin. 2017. *Language Revitalization*. Oxford Research Encyclopedia of Linguistics.

[102] Francesca Polletta. 1998. Contending stories: Narrative in social movements. *Qualitative sociology* 21, 4 (1998), 419–446.

[103] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation Sensitivity Analysis to Detect Unintended Model Biases. 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, 5740–5745. https://doi.org/10.18653/v1/D19-1578

[104] Laura Pulido. 2016. Flint, environmental racism, and racial capitalism. *Capitalism Nature Socialism* 27, 3 (2016), 1–16.

[105] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained Models for Natural Language Processing: A Survey. arXiv:2003.08271 [cs.CL]

[106] 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.

[107] 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. http://jmlr.org/papers/v21/20-074.html

[108] 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

[109] 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

[110] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics* 8 (2021), 842–866.

[111] Ronald Rosenfeld. 2000. Two decades of statistical language modeling: Where do we go from here? *Proc. IEEE* 88, 8 (2000), 1270–1278.

[112] Corby Rosset. 2020. Turing-NLG: A 17-billion-parameter language model by Microsoft. *Microsoft Blog* (2020).

[113] 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).

[114] 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

[115] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *Commun. ACM* 63, 12 (Nov. 2020), 54–63. https://doi.org/10.1145/3381831

[116] 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.

[117] Claude Elwood Shannon. 1949. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

[118] 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 [cs.CL]

[119] 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

[120] 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.

[121] Joonbo Shin, Yoonhyung Lee, and Kyomin Jung. 2019. Effective Sentence Scoring Method Using BERT for Speech Recognition. In *Asian Conference on Machine Learning*. 1081–1093.

[122] 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* (2019).

[123] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203* (2019).

[124] Karen Spärck Jones. 2004. *Language modelling's generative model: Is it rational?* Technical Report. Computer Laboratory, University of Cambridge.

[125] Robyn Speer. 2017. ConceptNet Numberbatch 17.04: better, less-stereotyped word vectors. (2017). Blog post, https://blog.conceptnet.io/2017/04/24/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/.

[126] 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 PMID: 26361054.

[127] Katrina Srigley and Lorraine Sutherland. 2019. Decolonizing, Indigenizing, and Learning Biskaaybiiyang in the Field: Our Oral History Journey1. *The Oral History Review* (2019).

[128] 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

[129] 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.

[130] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: Enhanced Representation through Knowledge Integration. arXiv:1904.09223 [cs.CL]

[131] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational*

Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. AAAI Press, 8968–8975. https://aaai.org/ojs/index.php/AAAI/article/view/6428

[132] 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.

[133] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers 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

[134] Trieu H. Trinh and Quoc V. Le. 2019. A Simple Method for Commonsense Reasoning. arXiv:1806.02847 [cs.AI]

[135] 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.

[136] 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.

[137] 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

[138] 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

[139] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, Vancouver, BC, Canada, 78–84. https://doi.org/10.18653/v1/W17-3012

[140] Joseph Weizenbaum. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. WH Freeman & Co.

[141] Monnica T Williams. 2019. Psychology Cannot Afford to Ignore the Many Harms Caused by Microaggressions. *Perspectives on Psychological Science* 15 (2019), 38 – 43.

[142] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[143] World Bank. 2018. Indiviuals Using the Internet. (2018). https://data.worldbank.org/indicator/IT.NET.USER.ZS?end=2017&locations=US&start=2015

[144] Shijie Wu and Mark Dredze. 2020. Are All Languages Created Equal in Multilingual BERT?. In *Proceedings of the 5th Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Online, 120–130. https://doi.org/10.18653/v1/2020.repl4nlp-1.16

[145] 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* (2020).

[146] Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. 2020. BERT-of-Theseus: Compressing BERT by Progressive Module Replacing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7859–7869. https://doi.org/10.18653/v1/2020.emnlp-main.633

[147] Peng Xu, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung. 2019. Clickbait? Sensational Headline Generation with Auto-tuned Reinforcement Learning. 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, 3065–3075. https://doi.org/10.18653/v1/D19-1303

[148] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. arXiv:2010.11934 [cs.CL]

[149] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-End Open-Domain Question Answering with BERTserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Association for Computational Linguistics, Minneapolis, Minnesota, 72–77. https://doi.org/10.18653/v1/N19-4013

[150] 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.

[151] Ze Yang, Can Xu, Wei Wu, and Zhoujun Li. 2019. Read, Attend and Comment: A Deep Architecture for Automatic News Comment 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, 5077–5089. https://doi.org/10.18653/v1/D19-1512

[152] 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.

[153] Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8BERT: Quantized 8Bit BERT. arXiv:1910.06188 [cs.CL]

[154] Nico Zazworka, Rodrigo O. Spínola, Antonio Vetro', Forrest Shull, and Carolyn Seaman. 2013. A Case Study on Effectively Identifying Technical Debt. In *Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering* (Porto de Galinhas, Brazil) *(EASE '13)*. Association for Computing Machinery, New York, NY, USA, 42–47. https://doi.org/10.1145/2460999.2461005

[155] 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

[156] 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.

[157] 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

[158] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Computational Linguistics* 46, 1 (March 2020), 53–93. https://doi.org/10.1162/coli_a_00368

ACKNOWLEDGMENTS

This paper represents the work of seven authors, but some were required by their employer to remove their names. The remaining listed authors are extremely grateful to our colleagues for the effort and wisdom they contributed to this paper.

In addition, in drafting and revising this paper, we benefited from thoughtful comments and discussion from many people: Alex Hanna, Amandalynne Paullada, Ben Hutchinson, Ben Packer, Brendan O'Connor, Dan Jurafsky, Ehud Reiter, Emma Strubell, Emily Denton, Gina-Anne Levow, Iason Gabriel, Jack Clark, Kristen Howell, Lucy Vasserman, Maarten Sap, Mark Díaz, Miles Brundage, Nick Doiron, Rob Munro, Roel Dobbe, Samy Bengio, Suchin Gururangan, Vinodkumar Prabhakaran, William Agnew, William Isaac, and Yejin Choi and our anonymous reviewers.