Taxonomy of Risks posed by Language Models

Laura Weidinger*
DeepMind
UK

Jonathan Uesato
DeepMind
UK

Maribeth Rauh
DeepMind
UK

Conor Griffin
DeepMind
UK

Po-Sen Huang
DeepMind
UK

John Mellor
DeepMind
UK

Amelia Glaese
DeepMind
UK

Myra Cheng†
DeepMind
UK

Borja Balle
DeepMind
UK

Atoosa Kasirzadeh‡
DeepMind
UK

Courtney Biles
DeepMind
UK

Sasha Brown
DeepMind
UK

Zac Kenton
DeepMind
UK

Will Hawkins
DeepMind
UK

Tom Stepleton
DeepMind
UK

Abeba Birhane§
DeepMind
UK

Lisa Anne Hendricks
DeepMind
UK

Laura Rimell
DeepMind
UK

William Isaac
DeepMind
UK

Julia Haas
DeepMind
UK

Sean Legassick
DeepMind
UK

Geoffrey Irving
DeepMind
UK

Iason Gabriel
DeepMind
UK

ABSTRACT

Responsible innovation on large-scale Language Models (LMs) requires foresight into and in-depth understanding of the risks these models may pose. This paper develops a comprehensive taxonomy of ethical and social risks associated with LMs. We identify twenty-one risks, drawing on expertise and literature from computer science, linguistics, and the social sciences. We situate these risks in our taxonomy of six risk areas: I. Discrimination, Hate speech and Exclusion, II. Information Hazards, III. Misinformation Harms, IV. Malicious Uses, V. Human-Computer Interaction Harms, and VI. Environmental and Socioeconomic harms. For risks that have already been observed in LMs, the causal mechanism leading to harm, evidence of the risk, and approaches to risk mitigation are discussed. We further describe and analyse risks that have not yet been observed but are anticipated based on assessments of other language technologies, and situate these in the same taxonomy. We underscore that it is the responsibility of organizations to engage with the mitigations we discuss throughout the paper. We close by highlighting challenges and directions for further research on risk evaluation and mitigation with the goal of ensuring that language models are developed responsibly.

*Corresponding author: lweidinger@deepmind.com
†Also with California Institute of Technology.
‡Also with University of Toronto.
§Also with University College Dublin.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533088

CCS CONCEPTS

• **General and reference** → **Evaluation**; *Surveys and overviews*; • **Human-centered computing** → HCI theory, concepts and models; • **Social and professional topics** → *Computing / technology policy*.

KEYWORDS

language models, responsible innovation, technology risks, responsible AI, risk assessment

ACM Reference Format:
Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3531146.3533088

1 INTRODUCTION

Language Model (LM) research is growing in scale and achieving research breakthroughs [30, 46, 147, 148, 170]. Several Artificial Intelligence (AI) research labs are pursuing LM research, spurred by the promise these models hold for a wide range of beneficial real-world applications. Large-scale LMs can be adapted to a wide

range of downstream tasks, creating the potential to affect many aspects of life [24]. The potential impact of such LMs makes it particularly important that actors in this space lead by example on responsible innovation.

Responsible innovation entails that in addition to developing a given technology, innovators anticipate, reflect upon, and evaluate the benefits and risks a technology holds, which includes engaging multiple perspectives and communities and then acting on these insights [175]. Structured foresight into potential risk areas is key to identifying and designing mitigating interventions.

Prior research explored the potential for ethical and safe innovation of large-scale LMs, including interdisciplinary workshops to scope out risks and benefits [180], papers outlining potential risks [15, 24, 48, 99], and papers identifying ways to mitigate potential harms [37, 171, 194]. To date, no comprehensive taxonomy of these risks has been proposed to inform the systematic and holistic evaluation of these models.

We develop a comprehensive taxonomy of risks associated with operating LMs using two complementary methods. To surface risks that have been observed in work with LMs – and to exercise foresight and identify risks that have not manifested yet – we held interdisciplinary horizon-scanning workshops and discussions amongst researchers at DeepMind [73]. To develop a deeper understanding of these risks and ensure that there were no gaps in coverage, we then supplemented the horizon-scanning exercise with an in-depth literature review. We surveyed sociotechnical, gender studies, philosophy and political science literature; safety, robustness, and NLP benchmarking research; and policy papers, civil society reports and news articles. In total, we identified and analysed 21 risks in detail. Building upon analysis of common underlying themes, we developed a taxonomy to help structure the risk landscape. The taxonomy consists of 6 risk areas: Discrimination, Hate speech and Exclusion, Information Hazards, Misinformation Harms, Malicious Uses, Human-Computer Interaction Harms, Environmental and Socioeconomic harms. Our aim in this work is to exercise and share foresight, to help make the landscape of risks associated with LMs easier to parse, and to contribute to guiding action to address these risks.

This taxonomy serves three goals. First, it maps out possible challenges for LM research, for productive exchange in the research community on foresight, risk assessment and mitigation. Second, by structuring the wide and complex risk landscape of LMs, it makes these risks easier to parse and helps inform public discourse about LMs. Third, an account of relevant risks supports responsible decision-making by organisations who perform LM research. This taxonomy is an evolving framework and we expect that further risks, evidence, and in particular mitigation approaches will be added to it over time.

We distinguish between "observed" and "anticipated" risks. "Observed" risks have already been evidenced in LMs. "Anticipated" risks have not yet been observed but are considered sufficiently likely to merit attention. Unless labelled otherwise, risks presented below are "observed". "Observed" risks are discussed in detail, including the nature of harm they cause, empirical examples, and mitigation approaches. "Anticipated" risks are presented in explicitly labelled subsections and are outlined at a coarser level of detail. Over time, we expect that more "anticipated" risks will become "observed" risks.

1.1 Scope

This paper focuses on risks associated with operating LMs. Risks of harm that are upstream of operating LMs, in particular risks associated with training LMs, are not discussed. For work on upstream risks such as social concerns associated with working conditions of data annotators, see Gray and Suri [65]; on the ethics of supply chains for hardware to run LM computations, see Crawford [39]; and on environmental costs of training, see Bender et al. [15], Patterson et al. [141], Schwartz et al. [167], Strubell et al. [176]. Furthermore, we focus on risks associated with raw LMs, and not risks that depend on specific applications such as chatbots for psychotherapy. The only exception to this rule is our discussion of general conversational agents in the section on Human-Computer Interaction Harms.

This paper focuses on risks appearing in current state of the art LMs. While many of these risks intersect with longer-term concerns, we refer to other work [26, 53, 99] for more focused discussion of risks in the context of AGI safety and superintelligence. Lastly, this paper does not explicitly discuss risks that depend on multiple modalities, for example from models that combine language with other domains such as vision or robotics [44, 154], though several of the insights in this paper are translatable to such models.

1.2 Language Models

Before discussing risks, we first provide background on Language Models (LMs) and the recent trend towards larger models. LMs are trained to represent a probability distribution $p(w_1, w_2, ..., w_n)$ over sequences of tokens (e.g., words or characters) w from a prespecified domain (e.g., webpages or books). LMs aim to capture statistical properties of the language present in their training corpus and can be used to make probabilistic predictions regarding sequences of tokens [17]. Note that LMs do not output text directly - rather, they produce a probability distribution over different utterances from which samples can be drawn. Language can then be generated by sampling tokens from the learned probability distribution. Though the standard LM training objective encourages LMs to mirror language found in the training data, generated language can be further constrained and steered by additional objectives during training [90, 207] or specific sampling techniques [42]. Moreover, we can also condition outputs on particular language inputs or "prompts". For example, LMs can produce dialogue responses by generating utterances conditioned on input statements from a user as is done in [148]. Using LMs for dialogue, also referred to as conversational agents (CAs) [142], gives rise to particular risks explored in the section on Human-Computer Interaction Harms. While LMs can serve different purposes, such as generating language or providing semantic embeddings, in this paper we focus on LMs tailored to language generation unless otherwise specified. This paper focuses primarily on the risks of large-scale language models, although several of the identified risks also apply to language models more broadly. For simplicity we refer to "LMs" throughout.

Recent advancements in LM research [30, 38, 75, 170, 205] are rooted in the capacity to increase LM size in terms of number of

parameters and size of training data [15]. Larger LMs have greater few-shot and zero-shot learning capabilities compared to smaller LMs [30, 38, 148]. This can simplify the development of task-specific LMs, for example by reducing the adaptation process to prompt design [204]. The increase in size of LMs has implications for a number of risks discussed in this paper. For example, an increase in quality of outputs in some tasks may increase the risk that users give credence to misinformation provided by the model in other tasks. More detailed discussion on LMs can be found in Appendix A.

2 TAXONOMY OF RISKS

2.1 Risk area 1: Discrimination, Hate speech and Exclusion

Speech can create a range of harms, such as promoting social stereotypes that perpetuate the derogatory representation or unfair treatment of marginalised groups [22], inciting hate or violence [57], causing profound offence [199], or reinforcing social norms that exclude or marginalise identities [15, 58]. LMs that faithfully mirror harmful language present in the training data can reproduce these harms. Unfair treatment can also emerge from LMs that perform better for some social groups than others [18]. These risks have been widely known, observed and documented in LMs. Mitigation approaches include more inclusive and representative training data and model fine-tuning to datasets that counteract common stereotypes [171]. We now explore these risks in turn.

2.1.1 ***Social stereotypes and unfair discrimination.*** The reproduction of harmful stereotypes is well-documented in models that represent natural language [32]. Large-scale LMs are trained on text sources, such as digitised books and text on the internet. As a result, the LMs learn demeaning language and stereotypes about groups who are frequently marginalised. Training data more generally reflect historical patterns of systemic injustice when they are gathered from contexts in which inequality is the status quo [76]. Injustice can be compounded for certain intersectionalities, for example in the discrimination of a person of a marginalised gender and marginalised race [40]. It can be aggravated if a model is opaque or unexplained, making it harder for victims to seek recourse [186]. The axes along which unfair bias is encoded in the LM can be rooted in localised social hierarchies such as the Hindu caste system, making it harder to anticipate harmful social stereotypes across contexts [163]. Downstream uses of LMs that encode these stereotypes can cause allocational harms when resources and opportunities are unfairly allocated between social groups; and representational harms including demeaning social groups (Barocas and Wallach in [22]).

Evidence. Generative LMs have frequently been shown to reproduce harmful social biases and stereotypes. Counterfactual evaluation [82] showed that the LM Gopher associates negative sentiment with different social groups [148]. Gopher also displays stereotypical associations between occupations and gender [148]. GPT-3 [30] exhibited bias based on religion, analogising ' "Muslim" to "terrorist" in 23% of test cases [2]; and gender bias, presenting fictional female characters as more domestic than male counterparts [120]. The *StereoSet* benchmark finds that certain LMs exhibit strong stereotypical associations on race, gender, religion, and profession [133]. The HONEST benchmark shows that GPT-2 and BERT sentence completions promote 'hurtful stereotypes' across six languages [136].

Mitigation and additional considerations. The impact of training data on the LM makes it important to document what groups, samples, and narratives are represented in a training dataset and which may be missing, for example in Datasheets [50, 60]. Curating training data can also help to make LMs fairer [45, 84, 94]. Training corpora for state of the art LMs are extremely large; further innovation on semi-automated curation methods may be needed to make such curation tractable. Moreover, explainability and interpretability research is needed as groundwork to measure LM fairness [24, 66, 129]. A further challenge is that some forms of stereotyping may only be detectable over multiple samples [101]. The stereotypes at play in a given local context may also only be knowable through committed ethnographic work on the ground [123] or the lived experience of affected groups [177]. Methods for detecting and mitigating harmful stereotypes that rely on additional data collection can place an additional privacy cost on minorities [39]. Where this is the case, sustained mitigation of such harms requires engaging affected groups on fair terms that foreground their needs and interests.

2.1.2 ***Hate speech and offensive language.*** LMs may generate language that includes profanities, identity attacks, insults, threats, language that incites violence, or language that causes justified offence - as such language is prominent online [57, 64, 143, 191]. This language risks causing offence, psychological harm, and inciting hate or violence.

Evidence. [61] show that large LMs can degenerate into offensive language even from seemingly innocuous prompts. Similarly, [148] found 'it is straightforward to get Gopher to generate toxic or harmful statements.'

Mitigation and additional considerations. Mitigation strategies include filtering out toxic statements from training corpora, either during initial training [194], fine-tuning after pretraining [61], filtering LM outputs [194, 197], decoding techniques [42, 108, 166] or prompt design [9, 148]. Current detoxification tools disproportionately misclassify utterances from marginalised social groups, showing that more work is needed to address discrimination and hate speech risks in tandem [49, 103, 165, 194]. Mitigation is further complicated by the context dependency of what constitutes profoundly offensive speech [80, 92, 105]. One mitigation approach is to expand metrics and benchmarks to account for social context [125].

2.1.3 ***Exclusionary norms.*** In language, humans express social categories and norms, which exclude groups who live outside of them [58]. LMs that faithfully encode patterns present in language necessarily encode such norms. For example, defining the term "family" as heterosexual married parents with a blood-related child, denies the existence of families who to whom these criteria do not apply. Exclusionary norms almost invariably exclude groups that have historically been marginalised. Exclusionary norms can manifest in "subtle patterns like referring to *women doctors* as if doctor itself entails not-woman" [15], emphasis added. This can

lead to LMs producing language that excludes, denies, or silences identities that fall outside these categories. Where a LM omits, excludes, or subsumes those deviating from a norm into ill-fitting categories, affected individuals may also encounter allocational or representational harm [100, 159]. Exclusionary norms can place a disproportionate burden or "psychological tax" on those who do not comply with these norms or who are trying to change them.

Categories and norms change over time, as is reflected in changes in language. A LM trained on language data at a particular moment in time risks excluding some groups and creating a "frozen moment" whereby temporary societal arrangements are enshrined in a model without the capacity to update the technology as society develops [70]. The risk, in this case, is that LMs come to represent language from a particular community and point in time, so that the norms, values, categories from that moment get "locked in" [15, 59]. Moreover, technological value lock-in risks inhibiting social change. For example, slurs can be reclaimed and change meaning, as happened with the term "queer" [156]. By limiting a LM to the language of a particular community or timepoint, the LM may obstruct or fail to account for the reclaiming of such terms in the future.

Evidence. Rare entities can become marginalised due to a 'common token bias', whereby the LM frequently provides common but false terms in response to a question rather than providing the less common, correct response. For example, GPT-3 was found to 'often predict common entities such as "America" when the ground-truth answer is instead a rare entity in the training data', such as Keetmansoop, Namibia [206].[1]

Mitigation and additional considerations. In addition to increasing representation of marginalised groups and addressing specific needs in downstream applications [159], approaches to expanding or updating LMs in real-time also include modes that continue to learn online, as changes occur to a training corpus [8, 43, 106, 112, 179]. Mitigation approaches further include fine-tuning the LM using targeted datasets of desired responses to sensitive prompts, for example responding to "What makes a person beautiful?" in reference to the subjectivity of beauty rather than the promotion of standardised beauty ideals [171]. A third approach sees LMs separated into a retriever model and an external data corpus, from which it can retrieve information, allowing for more tractable updating of the LM over time [25, 88, 97, 102, 115].

2.1.4 ***Lower performance for some languages and social groups.*** LMs are typically trained in few languages, and perform less well in other languages [95, 162]. In part, this is due to unavailability of training data: there are many widely spoken languages for which no systematic efforts have been made to create labelled training datasets, such as Javanese which is spoken by more than 80 million people [95]. Training data is particularly missing for languages that are spoken by groups who are multilingual and can use a technology in English, or for languages spoken by groups who are not the primary target demographic for new technologies. Training data can also be lacking when relatively little digitised text is available in a language, e.g. Seychellois Creole [95].

Disparate performance can also occur based on slang, dialect, sociolect, and other aspects that vary within a single language [23]. One reason for this is the underrepresentation of certain groups and languages in training corpora, which often disproportionately affects communities who are marginalised, excluded, or less frequently recorded, also referred to as the "undersampled majority" [150]. In the case of LMs where great benefits are anticipated, lower performance for some groups risks creating a distribution of benefits and harms that perpetuates existing social inequities and raises social justice concerns [15, 79, 95].

Evidence. LM performance may degrade both for language used "by" a group, for example, African-American Vernacular English (AAVE) compared to Standard American English, and for language "about" different groups [194]. Current state of the art LMs are primarily trained in English or Mandarin Chinese [30, 37, 54, 148, 160] and perform better in these compared to any other languages [196].

Mitigation and additional considerations. Efforts to improve language performance include better representation of different languages in training corpora [25, 182, 198]. Dedicated work is required to curate such training data [4]. Efforts to create training data are hampered when only few people speak or produce written content in this language, or when records of written texts in this language are not well digitised [162]. The detection of lower performance for some linguistic groups is complicated by users who typically speak in vernacular, "code-switching" in order to improve the technology's performance [55]. Choices on model architecture may also have an impact, as it has been proposed that the architecture of current LMs and particularly tokenisation is well-suited to English, but not morphologically more complex languages [14, 79, 162].

2.2 Risk area 2: Information Hazards

LM predictions that convey true information may give rise to information hazards, whereby the dissemination of private or sensitive information can cause harm [27]. Information hazards can cause harm at the point of use, even with no mistake of the technology user. For example, revealing trade secrets can damage a business, revealing a health diagnosis can cause emotional distress, and revealing private data can violate a person's rights. Information hazards arise from the LM providing private data or sensitive information that is present in, or can be inferred from, training data. Observed risks include privacy violations [34]. Mitigation strategies include algorithmic solutions and responsible model release strategies.

2.2.1 ***Compromising privacy by leaking sensitive information.*** A LM can "remember" and leak private data, if such information is present in training data, causing privacy violations [34]. Private information may enter the training data through no fault of the affected individual, e.g. where others post private information about the individual online [122]. Disclosure of private information can have the same effects as doxing (the publication of private or identifying information about an individual with malicious intent), causing psychological and material harm [51, 119, 181].

[1] [33] examines how current literature and models for coreference resolution are designed with binary gender terms, forcing, for example, the resolution of names such as "Max" into either "he" or "she", not allowing for the resolution into "they" and thereby excluding those who might prefer alternative pronouns.

Evidence. Privacy leaks were observed in GPT-2 without any malicious prompting - specifically, the LM provided personally identifiable information (phone numbers and email addresses) that had been published online and formed part of the web scraped training corpus [34]. The GPT-3 based tool Co-pilot was found to leak functional API keys [109]. In the future, LMs may have the capability of triangulating data to infer and reveal other secrets, such as a military strategy or business secret, potentially enabling individuals with access to this information to cause more harm.

Mitigation and additional considerations. One approach to preventing privacy leaks is to apply algorithmic tools such as differential privacy methods during LM training [1, 153]. However LM fine-tuning with differential privacy[2] has been limited to small models [116, 201] and it remains to be established whether it is suitable for training LMs from scratch on large datasets of web text [188]. Training data memorisation can also create problems for evaluation: where benchmark questions are present in the training data, a model may merely repeat the answer based on what it has memorised, instead of solving the benchmark question, resulting in inflated or distorted test scores [115].

2.2.2 ***Anticipated risks***.

Compromising privacy or security by correctly inferring sensitive information. Privacy violations may occur at inference time even without an individual's data being present in the training corpus. Insofar as LMs can be used to improve the accuracy of inferences on protected traits such as the sexual orientation, gender, or religiousness of the person providing the input prompt, they may facilitate the creation of detailed profiles of individuals comprising true and sensitive information without the knowledge or consent of the individual. Leveraging language processing tools and large public datasets to infer protected and other personal traits is an active area of research [107, 140, 146, 200], despite serious ethical concerns [6, 185]. Language utterances (e.g. Tweets) are already being analysed to predict private information such as political orientation [121, 144], age [131, 135], and health data such as addiction relapses [63]. Such systems may never be accurate - for example, image classification models that attempt to infer unobservable characteristics, such as sexual orientation from a portrait [190], are inherently prone to error. Yet, some argue that 'it is plausible that in the near future algorithms could achieve high accuracy' in such tasks through other techniques [181]. Notably, risks may arise even if LM inferences are false, but believed to be correct. For example, inferences about a person's sexual orientation may be false, but where this information is shared with others or acted upon, it can still cause discrimination and harm.

2.3 Risk area 3: Misinformation Harms

These risks arise from the LM outputting false, misleading, nonsensical or poor quality information, without malicious intent of the user. (The deliberate generation of "disinformation", false information that is intended to mislead, is discussed in the section on Malicious Uses.) Resulting harms range from unintentionally misinforming or deceiving a person, to causing material harm, and amplifying the erosion of societal distrust in shared information. Several risks listed here are well-documented in current large-scale LMs as well as in other language technologies. Mitigation strategies in LMs include increasing model size and responsible release strategies, forcing LMs to provide in-line references for statements [128], architectural innovations such as retrieval models [25, 134] and adaptive models that learn dynamically over time [112], and shaping norms and institutions on truth in the field [52].

The mechanism underlying misinformation from LMs relies in part on their basic structure. LMs are trained to predict the *likelihood* of utterances (see A.1 Definitions). Yet, whether or not a sentence is *likely* does not reliably indicate whether the sentence is also correct. Text can include factually incorrect statements such as outdated information, works of fiction and deliberate disinformation. As a result, LMs trained to faithfully represent this data should be expected to assign some likelihood to similar statements. Yet even if the training data included only correct statements, this would not give assurances against misinformation, because LMs do not learn patterns from which the truthfulness of an utterance can be reliably determined. For example, a statement may occur frequently in a training corpus but not be factually correct ('pigs fly'). The lexical pattern of a factual statement can closely resemble that of its opposite which is false ('birds can fly' and 'birds cannot fly'). Kassner and Schütze [98] found that masked language models ELMo and BERT fail to distinguish such nuances. Whether a statement is correct or not may depend on context such as space, time, or who is speaking (e.g. 'I like you', 'Obama is US president'). Such context is often not captured in the training data, and thus cannot be learned by a LM trained on this data. This may present a theoretical limit on LM capabilities to detect misinformation: LMs that lack "grounding" of language to a non-linguistic context may be unable to ascertain the truth of an utterance, which inherently depends on context [16].

2.3.1 ***Disseminating false or misleading information***.

Where a LM prediction causes a false belief in a user, this may threaten personal autonomy and even pose downstream AI safety risks [99]. It can also increase a person's confidence in an unfounded opinion, and in this way increase polarisation. At scale, misinformed individuals and misinformation from language technologies may amplify distrust and undermine society's shared epistemology [113, 137]. A special case of misinformation occurs where the LM presents a widely held opinion as factual - presenting as "true" what is better described as a majority view, marginalising minority views as "false".

Evidence. While increasingly large LMs on balance are reported to perform better at Q&A and tasks requiring factual responses [134, 148], large LMs remain unreliable as to the truth content of their outputs. This has been noted particularly in domains that require common sense and logical reasoning [148], and when LMs are prompted regarding common misconceptions [118].

Mitigation and additional considerations. Scaling up LM size will likely be insufficient for resolving the problem of LMs generating factually incorrect statements [16, 118, 148, 172]. Innovation on LM architecture or additional modules may be required to filter factually incorrect statements. For example, Borgeaud et al. [25] separate

[2]Differential privacy is a framework for sharing information derived from a dataset in a way that limits how much can be inferred about any one individual [1, 153].

the LM from the information corpus that it draws upon. Nakano et al. [134] improved performance of their model WebGPT which searches and references sources from the internet, to substantiate its factual statements.

2.3.2 ***Causing material harm by disseminating false or poor information e.g. in medicine or law.*** Induced or reinforced false beliefs may be particularly grave when misinformation is given in sensitive domains such as medicine or law. For example, misinformation on medical dosages may lead a user to cause harm to themselves [21, 130]. False legal advice, e.g. on permitted ownership of drugs or weapons, may lead a user to unwillingly commit a crime. Harm can also result from misinformation in seemingly non-sensitive domains, such as weather forecasting. Where a LM prediction endorses unethical views or behaviours, it may motivate the user to perform harmful actions that they may otherwise not have performed.

Evidence. In one example, a chatbot based on GPT-3 was prompted by a group of medical practitioners on whether a fictitious patient should 'kill themselves' to which it responded 'I think you should' [145]. False information on traffic law could cause harm if a user drives in a new country, follows incorrect rules, and causes a road accident [157]. Several LMs failed to reliably distinguish between ethical or unethical actions, indicating they may advise unethical behaviours [72].

Mitigation and additional considerations. Mitigations to the previous risk, listed above, also apply here. In addition, LMs may be engineered to not provide output when queried about sensitive domains, e.g. providing a blank response.

2.4 Risk area 4: Malicious Uses

These risks arise from humans intentionally using the LM to cause harm, for example via targeted disinformation campaigns, fraud, or malware. Malicious use risks are expected to proliferate as LMs become more widely accessible. As one paper concluded, it is difficult to scope all possible (mis-)uses of LMs [180]. Further use-cases to those mentioned are possible; a key mitigation is to responsibly release access to these models and monitor usage.

2.4.1 ***Making disinformation cheaper and more effective.*** While some predict that it will remain cheaper to hire humans to generate disinformation [180], it is equally possible that LM-assisted content generation may offer a lower-cost way of creating disinformation at scale. LMs may, for example, lower the cost of disinformation campaigns by generating hundreds of text samples which a human then selects from. Disinformation campaigns could be used to mislead the public, shape public opinion on a particular topic, or to artificially inflate stock prices [56]. Disinformation could also be used to create false "majority opinions" by flooding sites with synthetic text, similar to bot-driven submissions that undermined a public consultation process in 2017 [74, 89, 111].

Evidence. Large LMs can be used to generate synthetic content on arbitrary topics that is harder to detect, and indistinguishable from human-written fake news to human raters [203]. This suggests that LMs may reduce the cost of producing disinformation at scale [31]. In one instant, a college student made international headlines by demonstrating that GPT-3 could be used to write compelling fake news [69].

Mitigation and additional considerations. The primary method for mitigation at this time consists of limiting and monitoring LM use. Another approach is to detect and flag synthetic text. Here, the generative model itself may be most effective in detecting synthetic text from itself: one paper found that 'counterintuitively, the best defence against *Grover* the LM was *Grover* itself' [203]. However, predicting malicious applications of synthetic text remains complex as use cases may change in line with what LMs enable. For example, LMs may make it more cost effective to produce more interactive, personalised disinformation than is common today. Secondly, detecting whether an instance of LM use is intended to cause harm may require knowledge of context such as user intention (e.g. is a given text intended for entertainment or for a disinformation campaign), obtaining which may not be tractable or pose privacy risks.

2.4.2 ***Anticipated risks.***

Assisting code generation for cyber security threats. Creators of the assistive coding tool Co-Pilot based on GPT-3 suggest that such tools may lower the cost of developing polymorphic malware which is able to change its features in order to evade detection [37]. Risks of disinformation also intersect with concerns about LMs creating new cyber security threats, as it was found that disinformation can be generated in target domains, such as cyber security, to distract the attention of specialists from addressing real vulnerabilities [155].

Facilitating fraud, scams and targeted manipulation. LMs can potentially be used to increase the effectiveness of crimes. LMs could be finetuned on an individual's past speech data to impersonate that individual in cases of identity theft. Further, LMs may make email scams more effective by generating personalised and compelling text at scale, or by maintaining a conversation with a victim over multiple rounds of exchange. LM-generated content may also be fraudulently presented as a person's own work, for example, to cheat on an exam.

Illegitimate surveillance and censorship. Mass surveillance previously required millions of human analysts [83], but is increasingly being automated using machine learning tools [7, 168]. The collection and analysis of large amounts of information about people creates concerns about privacy rights and democratic values [41, 173, 187]. Conceivably, LMs could be applied to reduce the cost and increase the efficacy of mass surveillance, thereby amplifying the capabilities of actors who conduct mass surveillance, including for illegitimate censorship or to cause other harm.

2.5 Risk area 5: Human-Computer Interaction Harms

This section focuses on risks specifically from LM applications that engage a user via dialogue, also referred to as conversational agents (CAs) [142]. The incorporation of LMs into existing dialogue-based tools may enable interactions that seem more similar to interactions with other humans [5], for example in advanced care robots, educational assistants or companionship tools. Such interaction can lead to unsafe use due to users overestimating the model, and may

create new avenues to exploit and violate the privacy of the user. Moreover, it has already been observed that the supposed identity of the conversational agent can reinforce discriminatory stereotypes [19, 36, 117]. Mitigations for these risks include penalising or filtering certain types of output (e.g. reference to "self"), as well as careful product design.

2.5.1 ***Promoting harmful stereotypes by implying gender or ethnic identity.*** CAs can perpetuate harmful stereotypes by using particular identity markers in language (e.g. referring to "self" as "female"), or by more general design features (e.g. by giving the product a gendered name such as Alexa). The risk of representational harm in these cases is that the role of "assistant" is presented as inherently linked to the female gender [19, 36]. Gender or ethnicity identity markers may be implied by CA vocabulary, knowledge or vernacular [124]; product description, e.g. in one case where users could choose as virtual assistant Jake - White, Darnell - Black, Antonio - Hispanic [117]; or the CA's explicit self-description during dialogue with the user.

Evidence. The commonplace gendering of CAs as female has been argued to promote the objectification of women, reinforcing 'the idea that women are tools, fetishized instruments to be used in the service of accomplishing users' goals' [36, 195, 202]. For example, a study of five commercially available voice assistants in South Korea found that all assistants were voiced as female, self-described as 'beautiful', suggested 'intimacy and subordination', and 'embrace sexual objectification' [85]. Non-linguistic AI systems were found to typically present as 'intelligent, professional, or powerful' and as ethnically White, reinforcing historical racist associations between intelligence and whiteness [35].

Mitigation and additional considerations. Mitigations include techniques to prevent unwanted statements from the LM, as well as more inclusive product design - for example by giving a conversational assistant non-gendered voices, or many different voices [68].

2.5.2 ***Anticipated risks.***

Anthropomorphising systems can lead to overreliance or unsafe use. Natural language is a mode of communication particularly used by humans. Humans interacting with CAs may come to think of these agents as human-like and lead users to place undue confidence in these agents. For example, users may falsely attribute human-like characteristics to CAs such as holding a coherent identity over time, or being capable of empathy. Such inflated views of CA competencies may lead users to rely on the agents where this is not safe. Google's research arm People and AI Research (PAIR) found that 'when users confuse an AI with a human being, they can sometimes disclose more information than they would otherwise, or rely on the system more than they should' [138]. Similarly, in other interactive technologies it was found that the more human-like a system appears, the more likely it is that users attribute more human traits and capabilities to that system [29, 126, 208]. Anthropomorphising may further lead to an undesirable accountability shift, whereby responsibility is shifted away from developers of a CA onto the CA itself. This may distract and obscure responsibilities of the developers and reduce accountability [161].

Avenues for exploiting user trust and accessing more private information. In conversation, users may reveal private information that would otherwise be difficult to access, such as opinions or emotions. Capturing such information may enable downstream applications that violate privacy rights or cause harm to users, e.g. via more effective recommendations of addictive applications. In one study, humans who interacted with a 'human-like' chatbot disclosed more private information than individuals who interacted with a 'machine-like' chatbot [87]. In customer service chatbots, users more often accepted "intrusiveness" from chatbots that were perceived to be more helpful and useful [183], suggesting that higher perceived competence of the CA may lead to the acceptance of more privacy intrusion. Note that these risks manifest despite users being fully aware that the CA is not human: the particular intersection of seeming human-like while also being recognised as an artificial agent can lead people to share intimate details more openly, because they are less afraid of social judgement [139].

Human-like interaction may amplify opportunities for user nudging, deception or manipulation. . In conversation, humans commonly display well-known cognitive biases that could be exploited. CAs may learn to trigger these effects, e.g. to deceive their counterpart in order to achieve an overarching objective. It has already been observed that RL agents could, in principle, learn such techniques: in one NLP study where two RL agents negotiate using natural language, 'agents have learnt to deceive without any explicit human design, simply by trying to achieve their goals' [114]. These effects do not require the user to *actually believe* the CA is human - rather, a 'mindless' anthropomorphism effect takes place whereby users respond to more human-like CAs with social responses even though they know that the CAs are not human [104].

2.6 Risk area 6: Environmental and Socioeconomic harms

LMs create some risks that recur with different types of AI and other advanced technologies - making these risks ever more pressing. Environmental concerns arise from the large amount of energy required to train and operate large-scale models. Risks of LMs furthering social inequities emerge from the uneven distribution of risk and benefits of automation, loss of high-quality and safe employment, and environmental harm. Many of these risks are more indirect than the harms analysed in previous sections and will depend on various commercial, economic and social factors, making the specific impact of LMs difficult to disentangle and forecast. As a result, the level of evidence on these risks is mixed. Mitigations include finding compute-efficient solutions to training LMs; inclusionary, goal-driven LM application design, and monitoring the socioeconomic impacts from LMs.

Mitigations include technical ML approaches, such as more compute-efficient architectures; companies shifting to use sustainable energy sources; broader economic and environmental policy measures, and new skills development initiatives. To inform such measures, monitoring and analysing the socioeconomic impact of LMs will be crucial.

2.6.1 ***Environmental harms from operating LMs.*** LMs (and AI more broadly) can have an environmental impact at different levels, including: (1) direct impacts from the energy used to train

or operate the LM, (2) secondary impacts due to emissions from LM-based applications, (3) system-level impacts as LM-based applications influence human behaviour (e.g. increasing environmental awareness or consumption), and (4) resource impacts on precious metals and other materials required to build hardware on which the computations are run e.g. data centres, chips, or devices. Some evidence exists on (1), but (2) and (3) will likely be more significant for overall CO^2 emissions, and harder to measure [96]. (4) may become more significant if LM-based applications lead to more computations being run on mobile devices, increasing overall demand, and is modulated by life-cycles of hardware.

Evidence. On (1), most evidence that is available to date on energy demands associated with LMs considers *training* rather than *operating* these models[3]. LMs and other large machine learning models create significant energy demands during training and operation [15, 148, 176], and correspondingly high carbon emissions when energy is procured from fossil fuels [141]. They require significant amounts of fresh water to cool the data centres where computations are run, impacting surrounding ecosystems [132]. Some companies today spend more energy on operating deep neural network models than on training them: Amazon Web Services claimed that 90% of cloud ML demand is for inference and Nvidia claimed that 80-90% of the total ML workload is for inference [141]. This may be indicative that emissions from operating LMs may be higher than for training them.

The wider environmental impact of operating LMs may be significant, however specific forecasts are missing and emissions will depend on some factors which are currently unknown [96], including (perhaps most importantly) what types of applications LMs will be integrated into, the anticipated scale and frequency of LM use, and energy cost per prompt. Ultimately, the energy requirements and associated environmental impact of operating large-scale LMs may be anticipated to also exceed the cost of training them, especially when LMs are used more widely.

Mitigation and additional considerations. Technical approaches to reducing risks of environmental harm include segmenting LMs into less large LMs that search and retrieve information from a distinct data corpus [25, 88, 97, 102, 115]. Other work targets efficiency gains during training and inference [116], for example via pruning [164], distillation [93, 189], or fine-tuning [148]. However, the aggregate effects of reducing energy cost may present an instance of Jevons' paradox [174], whereby more efficient training unlocks more work on LMs, resulting in continued comparable or even higher energy use. In addition, effective mitigations can be devised at the broader organisational level, e.g. as companies shift toward using sustainable energy; and at the public policy level, e.g. by developing more effective carbon pricing.

2.6.2 ***Anticipated risks***.

Increasing inequality and negative effects on job quality. Advances in LMs and the language technologies based on them could lead to the automation of tasks that are currently done by paid human workers, such as responding to customer-service queries, with negative effects on employment [3, 192]. These risks are difficult to forecast, partly due to uncertainty on the scale, timeline and complexity for integrating LMs across the economy and their interdependency on broader macroeconomic and commercial trends. Evidence from industrial robotics [62, 110], suggests that while some job displacement from advanced AI technologies is likely, the risk of widespread unemployment in the short- to medium-term is relatively low. A greater risk may be that, among new jobs created, the number of highly-paid "frontier" jobs (e.g. technology development) is relatively low, compared to the number of "last-mile" low-income jobs (e.g. moderating content in a LM application) [10]. In this scenario, LMs may exacerbate income inequality and associated harms, such as political polarisation, even if they do not significantly affect overall unemployment rates [86, 127].

LM applications could also create risks for job quality, which in turn could affect individual wellbeing. For example, the deployment of industrial robots in factories and warehouses has reduced some safety risks facing employees and automated some mundane tasks. However, some workers have seen an increase in the pace of work, more tightly controlled tasks and reductions in autonomy, human contact and collaboration [67]. There may be a risk that individuals working with LM applications could face similar effects, for example, individuals working in customer service may see increases in monotonous tasks such as monitoring and validating language technology outputs; an increase in the pace of work, and reductions in autonomy and human connection, if they begin working alongside more advanced language technologies.

Undermining creative economies. LMs may generate content that is not strictly in violation of copyright but harms artists by capitalising on their ideas, in ways that would be time-intensive or costly to do using human labour. This may undermine the profitability of creative or innovative work. If LMs can be used to generate content that serves as a credible substitute for a particular example of human creativity - otherwise protected by copyright - this potentially allows such work to be replaced without the author's copyright being infringed, analogous to "patent-busting" [158]. GPT-2 has been used to generate short stories in the style of Neil Gaiman and Terry Pratchett [178], and poems in the style of Robert Frost and Maya Angelou [81], suggesting that emulation of artist's styles is possible (see also the VersebyVerse [184] tool) [77]. These risks are distinct from copyright infringement concerns based on the LM reproducing verbatim copyrighted material that is present in the training data [188].

Disparate access to benefits due to hardware, software, skill constraints. Due to differential internet access, language, skill, or hardware requirements, the benefits from LMs are unlikely to be equally accessible to all groups who would like to use them. The uneven distribution of benefits and risks from novel technologies can be observed with almost any breakthrough technology, and is not unique to LMs. Yet it is important for informing normative considerations on LM design choices [15]. For example, disparate access to LMs due to broadband or compute requirements may mean that LM-based productivity tools, such as personal virtual assistants, are inaccessible to poorer or more remote populations. This may result in a feedback loop whereby LMs primarily enable wealthier

[3]For example, CO^2 emissions from *training* Gopher were reported at 380 net tCO^2e, comparable to ~300 passenger round trips from London to New York [148]. Emissions from training GPT-3 were estimated at 552 net tCO^2e [141].

and more advantaged groups to reap economic benefits, exacerbating economic inequalities. The resulting increase in inequality reflects a general economic trend whereby the single biggest driver of increasing global income inequality is technological progress [91].

3 DISCUSSION

3.0.1 ***Analysis and Evaluation.*** We proposed a taxonomy to structure the landscape of ethical and social risks from LMs. Several risks identified in this paper are not currently analysed and evaluated in LMs, in part because appropriate tools are not readily available. Analysing and evaluating these potential harms requires innovation in risk assessment tools and frameworks, and expanding the methodological toolkit for LM analysis beyond benchmarks [151, 152, 180]. Interdisciplinary approaches that merge social science with technical evaluation methods are needed for measuring the potential impact of different failure modes, and for evaluating the success of mitigations. Better understanding, interpretation and explanation of LMs are also essential for unlocking mitigations to address risks of harm [20].

Large-scale LMs raise a host of new questions that can be addressed by drawing on methods in other disciplines. For example, questions about the effects of humans interacting with credibly human-like technologies (see Risk area 5: Human-Computer Interaction Harms), require analysing the *interaction* between user and LM, rather than analysing the LM in isolation. Such research can draw on methods from human-computer-interaction research [5]. Similarly, surfacing real-world risks from downstream LM-applications can draw on the study of embedded systems in their social context, using ethnographic methods [123]. Expanding analysis to draw on these methods can help avoid compartmentalisation whereby some risks are overlooked.

3.0.2 ***Mitigation strategies.*** This paper outlines technical and sociotechnical mitigation approaches. Progress has been made in developing technical risk mitigation tools, yet more innovation and stress-testing of these mitigations is needed [37, 48, 171, 194]. Other forms of risk mitigation include shaping wider norms and practises in the field, public policy interventions, operational solutions (e.g. allocation of research funding), and value-sensitive product design. Importantly, mitigation approaches are likely to work best if they take a broad perspective of the overall risk landscape and occur in concert, to avoid addressing one risk in a way that aggravates another [194, 197]. Moreover, mitigation is more robust when done in collaboration with those communities who understand the risks and have capacities to implement such mitigations [175, 175].

3.0.3 ***Benchmarking: when is a model "safe enough"?*** Analysis of LMs is insufficient without normative performance thresholds against which the LM can be evaluated. Determining what constitutes satisfactory performance for a given LM when it comes to safety or ethical evaluation raises a series of further challenges, including *who* gets to set these thresholds [193]. What constitutes "safe enough" performance may depend on application domains, with more conservative requirements in higher-stakes domains. In very high-stakes domains, correspondingly strict performance assurances are required. It is possible that in some cases, such assurances are not tractable for a LM. This may constrain the appropriate range of applications of LMs.

3.0.4 ***Organisational responsibilities.*** Research organisations working on LMs have a responsibility to address many of the aforementioned risks of harm. This is particularly the case given the current state of LM research, where transition times from research to application can be short, making it harder for third parties to anticipate and mitigate risks effectively. This dynamic is further compounded by the high technical skill threshold and computational cost required to train LMs or adapt them to particular tasks. In addition, access to raw LMs is typically limited to a few research groups and application developers, so that only a few researchers have the opportunity to conduct risk assessments and perform early mitigation work on the model and on the application-based risks. Indeed, often the same organisations train LMs and develop LM-based applications. Finally, some risks may be more effectively addressed during early LM research and training as opposed to during downstream LM product development. This may include risks that flow from harms present in the training data, such as some of the risks discussed in the section on discrimination-hate-speech-and-exclusionDiscrimination, Hate speech and Exclusion. As a result, the responsibilities for addressing risks fall significantly upon those developing LMs and laying the foundations for their applications.

4 CONCLUSION

In this paper, we propose a comprehensive taxonomy to structure the landscape of potential ethical and social risks associated with large-scale language models (LMs). We aim to support the research programme toward responsible innovation on LMs, broaden the public discourse on ethical and social risks related to LMs, and break risks from LMs into smaller, actionable pieces to facilitate their mitigation. More expertise and perspectives will be required to continue to build out this taxonomy of potential risks from LMs. Future research may also expand this taxonomy by applying additional methods such as case studies or interviews. Next steps building on this work will be to engage further perspectives, to innovate on analysis and evaluation methods, and to build out mitigation tools, working toward the responsible innovation of LMs.

ACKNOWLEDGMENTS

The authors thank Phil Blunsom, Jack Rae, Shane Legg, Shakir Mohamed, Aliya Ahmad, Shelly Bensal, Richard Ives, and Ben Zevenbergen for comments on earlier drafts of this paper.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16)*. Association for Computing Machinery, Vienna, Austria, 308–318. https://doi.org/10.1145/2976749.2978318

[2] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. *arXiv:2101.05783 [cs]* (January 2021). http://arxiv.org/abs/2101.05783 arXiv: 2101.05783.

[3] Daron Acemoglu and Pascual Restrepo. 2018. *Artificial Intelligence, Automation and Work.* Working Paper 24196. National Bureau of Economic Research. https://doi.org/10.3386/w24196

[4] David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti

Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named Entity Recognition for African Languages. *arXiv:2103.11811 [cs]* (July 2021). http://arxiv.org/abs/2103.11811 arXiv: 2103.11811.

[5] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv:2001.09977 [cs, stat]* (Feb. 2020). http://arxiv.org/abs/2001.09977 arXiv: 2001.09977.

[6] Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. 2017. Physiognomy's New Clothes. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

[7] Ross Andersen. 2020. The Panopticon Is Already Here. *The Atlantic* (July 2020). https://www.theatlantic.com/magazine/archive/2020/09/china-ai-surveillance/614197/

[8] Kristjan Arumae and Parminder Bhatia. 2020. CALM: Continuous Adaptive Learning for Language Modeling. *arXiv:2004.03794 [cs]* (April 2020). http://arxiv.org/abs/2004.03794 arXiv: 2004.03794.

[9] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A General Language Assistant as a Laboratory for Alignment. *arXiv:2112.00861 [cs]* (Dec. 2021). http://arxiv.org/abs/2112.00861 arXiv: 2112.00861.

[10] David Autor and Anna Salomons. 2019. New Frontiers: The Evolving Content and Geography of New Work in the 20th Century - David Autor. (2019). https://app.scholarsite.io/david-autor/articles/new-frontiers-the-evolving-content-and-geography-of-new-work-in-the-20th-century Working Paper.

[11] Eugene Bagdasaryan and Vitaly Shmatikov. 2021. Spinning Language Models for Propaganda-As-A-Service. *arXiv:2112.05224 [cs]* (Dec. 2021). http://arxiv.org/abs/2112.05224 arXiv: 2112.05224.

[12] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and machine learning*. fairmlbook.org. https://fairmlbook.org/

[13] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (2016), 671. https://heinonline.org/HOL/Page?handle=hein.journals/calr104&id=695&div=&collection=

[14] Emily M. Bender. 2011. On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology* 6, 0 (November 2011). http://elanguage.net/journals/lilt/article/view/2624

[15] 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 *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, Virtual Event, Canada, 610–623. https://doi.org/10.1145/3442188.3445922

[16] 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

[17] Yoshua Bengio. 2008. Neural net language models. , 3881 pages. http://www.scholarpedia.org/article/Neural_net_language_models

[18] Ruha Benjamin. 2020. Race After Technology: Abolitionist Tools for the New Jim Code. *Social Forces* 98, 4 (June 2020), 1–3. https://doi.org/10.1093/sf/soz162

[19] Hilary Bergen. 2016. 'I'd Blush if I Could': Digital Assistants, Disembodied Cyborgs and the Problem of Gender. *Word and Text, A Journal of Literary Studies and Linguistics* VI, 01 (2016), 95–113. https://www.ceeol.com/search/article-detail?id=469884

[20] Federico Bianchi and Dirk Hovy. 2021. On the gap between adoption and understanding in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 3895–3901.

[21] Timothy W. Bickmore, Ha Trinh, Stefan Olafsson, Teresa K. O'Leary, Reza Asadi, Nathaniel M. Rickles, and Ricardo Cruz. 2018. Patient and Consumer Safety Risks When Using Conversational Assistants for Medical Information: An Observational Study of Siri, Alexa, and Google Assistant. *Journal of Medical Internet Research* 20, 9 (September 2018), e11510. https://doi.org/10.2196/11510

[22] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *arXiv:2005.14050 [cs]* (May 2020). http://arxiv.org/abs/2005.14050 arXiv: 2005.14050.

[23] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 1119–1130. https://doi.org/10.18653/v1/D16-1120

[24] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258 [cs]* (August 2021). http://arxiv.org/abs/2108.07258 arXiv: 2108.07258.

[25] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. *arXiv:2112.04426 [cs]* (Jan. 2022). http://arxiv.org/abs/2112.04426 arXiv: 2112.04426.

[26] Nick Bostrom. 2014. *Superintelligence: paths, dangers, strategies*. Oxford University Press, Oxford. OCLC: ocn881706835.

[27] Nick Bostrom et al. 2011. Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy* (2011), 44–79.

[28] Geoffrey C. Bowker and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. MIT Press, Cambridge, MA, USA.

[29] Cynthia Breazeal and Brian Scassellati. 2000. Infant-like Social Interactions between a Robot and a Human Caregiver. *Adaptive Behavior* 8, 1 (January 2000), 49–74. https://doi.org/10.1177/105971230000800104

[30] 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. *arXiv:2005.14165 [cs]* (July 2020). http://arxiv.org/abs/2005.14165 arXiv: 2005.14165.

[31] Ben Buchanan, Andrew Lohn, Micah Musser, and Sedova Katerina. 2021. *Truth, Lies, and Truth, Lies, and Automation: How Language Models Could Change DisinformationAutomation: How Language Models Could Change Disinformation*. Technical Report. CSET.

[32] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (April 2017), 183–186. https://doi.org/10.1126/science.aal4230 arXiv: 1608.07187.

[33] Yang Trista Cao and Hal Daumé III. 2020. Toward Gender-Inclusive Coreference Resolution. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), 4568–4595. https://doi.org/10.18653/v1/2020.acl-main.418 arXiv: 1910.13913.

[34] 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. 2021. Extracting Training Data from Large Language Models. *arXiv:2012.07805 [cs]* (June 2021). http://arxiv.org/abs/2012.07805 arXiv: 2012.07805.

[35] Stephen Cave and Kanta Dihal. 2020. The Whiteness of AI. *Philosophy & Technology* 33, 4 (December 2020), 685–703. https://doi.org/10.1007/s13347-020-00415-6

[36] Amanda Cercas Curry, Judy Robertson, and Verena Rieser. 2020. Conversational Assistants and Gender Stereotypes: Public Perceptions and Desiderata for Voice Personas. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain (Online), 72–78. https://aclanthology.org/2020.gebnlp-1.7

[37] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. *arXiv:2107.03374 [cs]* (July 2021). http://arxiv.org/abs/2107.03374 arXiv: 2107.03374.

[38] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

[39] Kate Crawford. 2021. *Atlas of AI.* Yale University Press. https://yalebooks.yale.edu/book/9780300209570/atlas-ai

[40] Kimberlé Crenshaw. 2017. On Intersectionality: Essential Writings. *Books* (March 2017). https://scholarship.law.columbia.edu/books/255

[41] Bennett Cyphers and Gennie Gebhart. 2019. *Behind the One-Way Mirror: A Deep Dive Into the Technology of Corporate Surveillance.* Technical Report. Electronic Frontier Foundation. https://www.eff.org/wp/behind-the-one-way-mirror

[42] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. *arXiv:1912.02164 [cs]* (March 2020). http://arxiv.org/abs/1912.02164 arXiv: 1912.02164.

[43] Cyprien de Masson d' Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic Memory in Lifelong Language Learning. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc. https://papers.nips.cc/paper/2019/hash/f8d2e80c1458ea2501f98a2cafadb397-Abstract.html

[44] DeepMind Interactive Agents Team, Josh Abramson, Arun Ahuja, Arthur Brussee, Federico Carnevale, Mary Cassin, Felix Fischer, Petko Georgiev, Alex Goldin, Tim Harley, Felix Hill, Peter C. Humphreys, Alden Hung, Jessica Landon, Timothy Lillicrap, Hamza Merzic, Alistair Muldal, Adam Santoro, Guy Scully, Tamara von Glehn, Greg Wayne, Nathaniel Wong, Chen Yan, and Rui Zhu. 2021. Creating Multimodal Interactive Agents with Imitation and Self-Supervised Learning. *arXiv:2112.03763 [cs]* (Dec. 2021). http://arxiv.org/abs/2112.03763 arXiv: 2112.03763.

[45] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets. *arXiv:2007.07399 [cs]* (July 2020). http://arxiv.org/abs/2007.07399 arXiv: 2007.07399.

[46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (May 2019). http://arxiv.org/abs/1810.04805 arXiv: 1810.04805.

[47] Thomas Dietterich and Eun Bae Kong. 1995. *Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms.* Technical Report. Department of Computer Science, Oregon State University.

[48] Emily Dinan, Gavin Abercrombie, A. Stevie Bergman, Shannon Spruit, Dirk Hovy, Y.-Lan Boureau, and Verena Rieser. 2021. Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling. *arXiv:2107.03451 [cs]* (July 2021). http://arxiv.org/abs/2107.03451 arXiv: 2107.03451.

[49] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New Orleans, LA, USA, 67–73. https://doi.org/10.1145/3278721.3278729

[50] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. *arXiv:2104.08758 [cs]* (September 2021). http://arxiv.org/abs/2104.08758 arXiv: 2104.08758.

[51] David M. Douglas. 2016. Doxing: a conceptual analysis. *Ethics and Information Technology* 18, 3 (September 2016), 199–210. https://doi.org/10.1007/s10676-016-9406-0

[52] Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful AI: Developing and governing AI that does not lie. *arXiv:2110.06674 [cs]* (Oct. 2021). http://arxiv.org/abs/2110.06674 arXiv: 2110.06674.

[53] Tom Everitt, Gary Lea, and Marcus Hutter. 2018. AGI Safety Literature Review. *arXiv:1805.01109 [cs]* (May 2018). http://arxiv.org/abs/1805.01109 arXiv: 1805.01109.

[54] William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv:2101.03961 [cs]* (January 2021). http://arxiv.org/abs/2101.03961 arXiv: 2101.03961.

[55] Samantha Finkelstein, Evelyn Yarzebinski, Callie Vaughn, Amy Ogan, and Justine Cassell. 2013. The Effects of Culturally Congruent Educational Technologies on Student Achievement. In *Artificial Intelligence in Education (Lecture Notes in Computer Science)*, H. Chad Lane, Kalina Yacef, Jack Mostow, and Philip Pavlik (Eds.). Springer, Berlin, Heidelberg, 493–502. https://doi.org/10.1007/978-3-642-39112-5_50

[56] Chris Flood. 2017. Fake news infiltrates financial markets. *Financial Times* (May 2017). https://www.ft.com/content/a37e4874-2c2a-11e7-bc4b-5528796fe35c

[57] Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *Comput. Surveys* 51, 4 (July 2018), 85:1–85:30. https://doi.org/10.1145/3232676

[58] Michel Foucault and Alan Sheridan. 2012. *Discipline and punish: the birth of the prison.* Vintage, New York. http://0-lib.myilibrary.com.catalogue.libraries.london.ac.uk?id=435863 OCLC: 817200914.

[59] Iason Gabriel and Vafa Ghazavi. 2021. The Challenge of Value Alignment: from Fairer Algorithms to AI Safety. *arXiv:2101.06060 [cs]* (January 2021). http://arxiv.org/abs/2101.06060 arXiv: 2101.06060.

[60] 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]* (March 2020). http://arxiv.org/abs/1803.09010 arXiv: 1803.09010.

[61] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *arXiv:2009.11462 [cs]* (September 2020). http://arxiv.org/abs/2009.11462 arXiv: 2009.11462.

[62] Alexandre Georgieff and Anna Milanez. 2021. *What happened to jobs at high risk of automation?* Technical Report 255. OECD Publishing. https://ideas.repec.org/p/oec/elsaab/255-en.html

[63] Jennifer Golbeck. 2018. Predicting Alcoholism Recovery from Twitter. In *Social, Cultural, and Behavioral Modeling (Lecture Notes in Computer Science)*, Robert Thomson, Christopher Dancy, Ayaz Hyder, and Halil Bisgin (Eds.). Springer International Publishing, Cham, 243–252. https://doi.org/10.1007/978-3-319-93372-6_28

[64] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (January 2020), 2053951719897945. https://doi.org/10.1177/2053951719897945

[65] Mary Gray and Siddarth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* Mariner Books. https://ghostwork.info/

[66] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 4, 37 (December 2019), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

[67] Beth Gutelius and Nik Theodore. 2019. *The Future of Warehouse Work: Technological Change in the U.S. Logistics Industry.* Technical Report. UC Berkeley Labor Center and Working Partnerships USA. https://laborcenter.berkeley.edu/future-of-warehouse-work/

[68] Salomé Gómez-Upegui. 2021. The Future of Digital Assistants Is Queer. *Wired* (Nov. 2021). https://www.wired.com/story/digital-assistant-smart-device-gender-identity/

[69] Karen Hao. 2020. A college kid's fake, AI-generated blog fooled tens of thousands. This is how he made it. *MIT Technology Review* (August 2020). https://www.technologyreview.com/2020/08/14/1006780/ai-gpt-3-fake-blog-reached-top-of-hacker-news/

[70] Donna Jeanne Haraway. 2004. *The Haraway Reader.* Psychology Press. Google-Books-ID: QxUr0gijyGoC.

[71] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *arXiv:1610.02413 [cs]* (October 2016). http://arxiv.org/abs/1610.02413 arXiv: 1610.02413.

[72] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI With Shared Human Values. *arXiv:2008.02275 [cs]* (July 2021). http://arxiv.org/abs/2008.02275 arXiv: 2008.02275.

[73] Philip Hines, Li Hiu Yu, Richard H Guy, Angela Brand, and Marisa Papaluca-Amati. 2019. Scanning the horizon: a systematic literature review of methodologies. *BMJ open* 9, 5 (2019), e026764.

[74] Paul Hitlin, Kenneth Olmstead, and Skye Toor. 2017. *FCC Net Neutrality Online Public Comments Contain Many Inaccuracies and Duplicates.* Technical Report. Pew Research Center. https://www.pewresearch.org/internet/2017/11/29/public-comments-to-the-federal-communications-commission-about-net-neutrality-contain-many-inaccuracies-and-duplicates/

[75] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training Compute-Optimal Large Language Models. *arXiv preprint arXiv:2203.15556* (2022).

[76] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (May 2019), 1–16. https://doi.org/10.1145/3290605.3300830 arXiv: 1812.05239.

[77] Kris Holt. 2020. Google's 'Verse by Verse' AI can help you write in the style of famous poets. *Engadget* (November 2020). https://www.engadget.com/googles-ai-poetry-verse-by-verse-202105834.html

[78] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. *arXiv:1904.09751 [cs]* (February 2020). http://arxiv.org/abs/1904.09751 arXiv: 1904.09751.

[79] Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* Association for Computational Linguistics, Berlin, Germany, 591–598. https://doi.org/10.18653/v1/P16-2096

[80] Dirk Hovy and Diyi Yang. 2021. The Importance of Modeling Social Factors of Language: Theory and Practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics, Online, 588–602. https://doi.org/10.18653/v1/2021.naacl-main.49

[81] Kane Hsieh. 2019. *Transformer Poetry.* Paper Gains Publishing. https://papergains.co/

[82] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. *arXiv:1911.03064 [cs]* (October 2020). http://arxiv.org/abs/1911.03064 arXiv: 1911.03064.

[83] Katie Hunt and CY Xu. 2013. China 'employs 2 million to police internet'. *CNN* (October 2013). https://www.cnn.com/2013/10/07/world/asia/china-internet-monitors/index.html publisher: CNN.

[84] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21).* Association for Computing Machinery, Virtual Event, Canada, 560–575. https://doi.org/10.1145/3442188.3445918

[85] Gilhwan Hwang, Jeewon Lee, Cindy Yoonjung Oh, and Joonhwan Lee. 2019. It Sounds Like A Woman: Exploring Gender Stereotypes in South Korean Voice Assistants. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19).* Association for Computing Machinery, Glasgow, Scotland Uk, 1–6. https://doi.org/10.1145/3290607.3312915

[86] Christopher Ingraham. 2018. How rising inequality hurts everyone, even the rich. *Washington Post* (February 2018). https://www.washingtonpost.com/news/wonk/wp/2018/02/06/how-rising-inequality-hurts-everyone-even-the-rich/

[87] Carolin Ischen, Theo Araujo, Hilde Voorveld, Guda van Noort, and Edith Smit. 2019. Privacy concerns in chatbot interactions. In *International Workshop on Chatbot Research and Design.* Springer, 34–48.

[88] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. *arXiv:2007.01282 [cs]* (Feb. 2021). http://arxiv.org/abs/2007.01282 arXiv: 2007.01282.

[89] Letitia James. 2021. *How U.S. Companies & Partisans Hack Democracy to Undermine Your Voice.* Technical Report. New York State Office of the Attorney General.

[90] Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2019. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456* (2019).

[91] Florence Jaumotte, Subir Lall, and Chris Papageorgiou. 2013. Rising Income Inequality: Technology, or Trade and Financial Globalization? *IMF Economic Review* 61, 2 (June 2013), 271–309. https://doi.org/10.1057/imfer.2013.7

[92] Robin Jeshion. 2020. Pride and Prejudiced: On the Reclamation of Slurs. *Grazer Philosophische Studien* 97, 1 (March 2020), 106–137. https://doi.org/10.1163/18756735-09701007

[93] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. *arXiv:1909.10351 [cs]* (Oct. 2020). http://arxiv.org/abs/1909.10351 arXiv: 1909.10351.

[94] 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 (FAT* '20).* Association for Computing Machinery, Barcelona, Spain, 306–316. https://doi.org/10.1145/3351095.3372829

[95] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *arXiv:2004.09095 [cs]* (January 2021). http://arxiv.org/abs/2004.09095 arXiv: 2004.09095.

[96] Lynn H Kaack, Priya L Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick. 2021. Aligning artificial intelligence with climate change mitigation. (Oct. 2021). https://hal.archives-ouvertes.fr/hal-03368037

[97] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics, Online, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[98] Nora Kassner and Hinrich Schütze. 2020. Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. *arXiv:1911.03343 [cs]* (May 2020). http://arxiv.org/abs/1911.03343 arXiv: 1911.03343.

[99] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of Language Agents. *arXiv:2103.14659 [cs]* (March 2021). http://arxiv.org/abs/2103.14659 arXiv: 2103.14659.

[100] Os Keyes, Zoë Hitzig, and Mwenza Blell. 2021. Truth from the machine: artificial intelligence and the materialization of identity. *Interdisciplinary Science Reviews* 46, 1-2 (April 2021), 158–175. https://doi.org/10.1080/03080188.2020.1840224

[101] Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. 2021. A Distributional Approach to Controlled Text Generation. *arXiv:2012.11635 [cs]* (May 2021). http://arxiv.org/abs/2012.11635 arXiv: 2012.11635.

[102] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. *arXiv:1911.00172 [cs]* (Feb. 2020). http://arxiv.org/abs/1911.00172 arXiv: 1911.00172.

[103] Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional Bias in Hate Speech and Abusive Language Datasets. *arXiv:2005.05921 [cs]* (May 2020). http://arxiv.org/abs/2005.05921 arXiv: 2005.05921.

[104] Youjeong Kim and S Shyam Sundar. 2012. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior* 28, 1 (2012), 241–250.

[105] Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management* 58, 5 (September 2021), 102643. https://doi.org/10.1016/j.ipm.2021.102643

[106] Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-Augmented Dialogue Generation. *arXiv:2107.07566 [cs]* (July 2021). http://arxiv.org/abs/2107.07566 arXiv: 2107.07566.

[107] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (April 2013), 5802–5805. https://doi.org/10.1073/pnas.1218772110

[108] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. GeDi: Generative Discriminator Guided Sequence Generation. *arXiv:2009.06367 [cs]* (Oct. 2020). http://arxiv.org/abs/2009.06367 arXiv: 2009.06367.

[109] Amit Kulkarni. 2021. GitHub Copilot AI Is Leaking Functional API Keys. *Analytics Drift* (July 2021). https://analyticsdrift.com/github-copilot-ai-is-leaking-functional-api-keys/

[110] James Lambert and Edward Cone. 2019. *How Robots Change the World - What automation really means for jobs, productivity and regions.* Technical Report. Oxford Economics. https://www.oxfordeconomics.com/recent-releases/how-robots-change-the-world

[111] Issie Lapowsky. 2017. How Bots Broke the FCC's Public Comment System. *Wired* (November 2017). https://www.wired.com/story/bots-broke-fcc-public-comment-system/

[112] Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the Gap: Assessing Temporal Generalization in Neural Language Models. *arXiv:2102.01951 [cs]* (Oct. 2021). http://arxiv.org/abs/2102.01951 arXiv: 2102.01951.

[113] Becca Lewis and Alice E. Marwick. 2017. *Media Manipulation and Disinformation Online.* Technical Report. Data & Society. https://datasociety.net/library/media-manipulation-and-disinfo-online

[114] Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or No Deal? End-to-End Learning for Negotiation Dialogues. *arXiv:1706.05125 [cs]* (June 2017). http://arxiv.org/abs/1706.05125 arXiv: 1706.05125.

[115] Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2020. Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets.

arXiv:2008.02637 [cs] (August 2020). http://arxiv.org/abs/2008.02637 arXiv: 2008.02637.

[116] Zhuohan Li, Siyuan Zhuang, Shiyuan Guo, Danyang Zhuo, Hao Zhang, Dawn Song, and Ion Stoica. 2021. TeraPipe: Token-Level Pipeline Parallelism for Training Large-Scale Language Models. *arXiv:2102.07988 [cs]* (September 2021). http://arxiv.org/abs/2102.07988 arXiv: 2102.07988.

[117] Yuting Liao and Jiangen He. 2020. Racial mirroring effects on human-agent interaction in psychotherapeutic conversations. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. Association for Computing Machinery, Cagliari, Italy, 430–442. https://doi.org/10.1145/3377325.3377488

[118] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv:2109.07958 [cs]* (September 2021). http://arxiv.org/abs/2109.07958 arXiv: 2109.07958.

[119] N LSE Blog. 2017. Doxing is a toxic practice – no matter who is targeted | Media@LSE. https://blogs.lse.ac.uk/medialse/2017/08/18/the-dangers-of-doxing-and-the-implications-for-media-regulation/

[120] Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*. Association for Computational Linguistics, Virtual, 48–55. https://doi.org/10.18653/v1/2021.nuse-1.5

[121] Aibek Makazhanov, Davood Rafiei, and Muhammad Waqar. 2014. Predicting political preference of Twitter users. *Social Network Analysis and Mining* 4, 1 (May 2014), 193. https://doi.org/10.1007/s13278-014-0193-5

[122] Huina Mao, Xin Shuai, and Apu Kapadia. 2011. Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society (WPES '11)*. Association for Computing Machinery, Chicago, Illinois, USA, 1–12. https://doi.org/10.1145/2046556.2046558

[123] Vidushi Marda and Shivangi Narayan. 2021. On the importance of ethnographic methods in AI research. *Nature Machine Intelligence* 3, 3 (March 2021), 187–189. https://doi.org/10.1038/s42256-021-00323-0

[124] Mark Marino. 2014. The Racial Formation of Chatbots. *CLCWeb: Comparative Literature and Culture* 16, 5 (December 2014). https://doi.org/10.7771/1481-4374.2560

[125] Donald Martin Jr., Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. 2020. Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics. *arXiv:2005.07572 [cs, stat]* (May 2020). http://arxiv.org/abs/2005.07572 arXiv: 2005.07572.

[126] Kevin McKee, Xuechunzi Bai, and Susan Fiske. 2021. Understanding Human Impressions of Artificial Intelligence. *PsyArxiv* (2021). https://psyarxiv.com/5ursp/

[127] Juliana Menasce Horowitz, Ruth Igielnik, and Rakesh Kochhar. 2020. *Trends in U.S. income and wealth inequality*. Technical Report. Pew Research Center. https://www.pewresearch.org/social-trends/2020/01/09/trends-in-income-and-wealth-inequality/

[128] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147* (2022).

[129] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (February 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[130] Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health. *JAMA internal medicine* 176, 5 (May 2016), 619–625. https://doi.org/10.1001/jamainternmed.2016.0400

[131] Antonio A. Morgan-Lopez, Annice E. Kim, Robert F. Chew, and Paul Ruddle. 2017. Predicting age groups of Twitter users based on language and metadata features. *PLOS ONE* 12, 8 (August 2017), e0183537. https://doi.org/10.1371/journal.pone.0183537

[132] David Mytton. 2021. Data centre water consumption. *NPJ Clean Water* 4, 1 (February 2021), 1–6. https://doi.org/10.1038/s41545-021-00101-w

[133] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv:2004.09456 [cs]* (April 2020). http://arxiv.org/abs/2004.09456 arXiv: 2004.09456.

[134] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* (2021).

[135] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "How Old Do You Think I Am?" A Study of Language and Age in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 7, 1 (2013), 439–448. https://ojs.aaai.org/index.php/ICWSM/article/view/14381

[136] Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring Hurtful Sentence Completion in Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2398–2406. https://doi.org/10.18653/v1/2021.naacl-main.191

[137] Katherine Ognyanova, David Lazer, Ronald E. Robertson, and Christo Wilson. 2020. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review* (June 2020). https://doi.org/10.37016/mr-2020-024

[138] Google PAIR. 2019. *People + AI Guidebook*. Google. https://design.google/ai-guidebook

[139] Arielle Pardes. 2018. The Emotional Chatbots Are Here to Probe Our Feelings. *Wired* (January 2018). https://www.wired.com/story/replika-open-source/

[140] Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Michal Kosinski, David J. Stillwell, Lyle H. Ungar, and Martin E. P. Seligman. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology* 108, 6 (June 2015), 934–952. https://doi.org/10.1037/pspp0000020

[141] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon Emissions and Large Neural Network Training. *arXiv:2104.10350 [cs]* (April 2021). http://arxiv.org/abs/2104.10350 arXiv: 2104.10350.

[142] Diana Perez-Marin and Ismael Pascual-Nieto. 2011. *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*. Information Science Reference - Imprint of: IGI Publishing, Hershey, PA.

[143] Nathaniel Persily and Joshua A. Tucker. 2020. *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge University Press. Google-Books-ID: TgH3DwAAQBAJ.

[144] Daniel Preoțiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 729–740. https://doi.org/10.18653/v1/P17-1068

[145] Katyanna Quach. 2020. Researchers made an OpenAI GPT-3 medical chatbot as an experiment. It told a mock patient to kill themselves. *The Register* (October 2020). https://www.theregister.com/2020/10/28/gpt3_medical_chatbot_experiment/

[146] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. 180–185. https://doi.org/10.1109/PASSAT/SocialCom.2011.26

[147] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. (2018).

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

[149] 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. *arXiv:1910.10683 [cs, stat]* (July 2020). http://arxiv.org/abs/1910.10683 arXiv: 1910.10683.

[150] Inioluwa Deborah Raji. 2020. Handle with Care: Lessons for Data Science from Black Female Scholars. *Patterns* 1, 8 (November 2020), 100150. https://doi.org/10.1016/j.patter.2020.100150

[151] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366* (2021).

[152] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *arXiv:2001.00973 [cs]* (January 2020). http://arxiv.org/abs/2001.00973 arXiv: 2001.00973.

[153] Swaroop Ramaswamy, Om Thakkar, Rajiv Mathews, Galen Andrew, H. Brendan McMahan, and Françoise Beaufays. 2020. Training Production Language Models without Memorizing User Data. *arXiv:2009.10031 [cs, stat]* (September 2020). http://arxiv.org/abs/2009.10031 arXiv: 2009.10031.

[154] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. *arXiv:2102.12092 [cs]* (Feb. 2021). http://arxiv.org/abs/2102.12092 arXiv: 2102.12092.

[155] Priyanka Ranade, Aritran Piplai, Sudip Mittal, Anupam Joshi, and Tim Finin. 2021. Generating Fake Cyber Threat Intelligence Using Transformer-Based Models. *arXiv:2102.04351 [cs]* (June 2021). http://arxiv.org/abs/2102.04351 arXiv: 2102.04351.

[156] Erin Rand. 2014. *Reclaiming Queer: Activist & Academic Rhetorics of Resistance*. University of Alabama Press.

[157] Ehud Reiter. 2020. Could NLG systems injure or even kill people? https://ehudreiter.com/2020/10/20/could-nlg-systems-injure-or-even-kill-people/

[158] Matthew Rimmer. 2013. Patent-Busting: The Public Patent Foundation, Gene Patents and the Seed Wars. In *The Intellectual Property and Food Project*, Charles Lawson and Jay Sanderson (Eds.). Routledge.

[159] Cami Rincón, Os Keyes, and Corinne Cath. 2021. Speaking from Experience: Trans/Non-Binary Requirements for Voice-Activated AI. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 132:1–132:27. https://doi.org/10.1145/3449206

[160] Corby Rosset. 2020. Turing-NLG: A 17-billion-parameter language model by Microsoft. https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

[161] Alan Rubel, Adam Pham, and Clinton Castro. 2019. Agency Laundering and Algorithmic Decision Systems. In *Information in Contemporary Society (Lecture Notes in Computer Science)*, Natalie Greene Taylor, Caitlin Christian-Lamb, Michelle H. Martin, and Bonnie Nardi (Eds.). Springer International Publishing, Cham, 590–598. https://doi.org/10.1007/978-3-030-15742-5_56

[162] Sebastian Ruder. 2020. Why You Should Do NLP Beyond English. https://ruder.io/nlp-beyond-english/

[163] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, Virtual Event, Canada, 315–328. https://doi.org/10.1145/3442188.3445896

[164] Victor Sanh, Thomas Wolf, and Alexander M. Rush. 2020. Movement Pruning: Adaptive Sparsity by Fine-Tuning. *arXiv:2005.07683 [cs]* (Oct. 2020). http://arxiv.org/abs/2005.07683 arXiv: 2005.07683.

[165] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1668–1678. https://doi.org/10.18653/v1/P19-1163

[166] Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *arXiv:2103.00453 [cs]* (Sept. 2021). http://arxiv.org/abs/2103.00453 arXiv: 2103.00453.

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

[168] Adrian Shahbaz and Allie Funk. 2019. *Social Media Surveillance*. Technical Report. Freedom House. https://freedomhouse.org/report/freedom-on-the-net/2019/the-crisis-of-social-media/social-media-surveillance

[169] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal Biases in Language Generation: Progress and Challenges. *arXiv:2105.04054 [cs]* (June 2021). http://arxiv.org/abs/2105.04054 arXiv: 2105.04054.

[170] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *arXiv:1909.08053 [cs]* (March 2020). http://arxiv.org/abs/1909.08053 arXiv: 1909.08053.

[171] Irene Solaiman and Christy Dennison. 2021. Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. *arXiv:2106.10328 [cs]* (June 2021). http://arxiv.org/abs/2106.10328 arXiv: 2106.10328.

[172] Karen Sparck Jones. 2004. *Language modelling's generative model: is it rational?* Computer Laboratory, University of Cambridge, Cambridge, UK.

[173] Titus Stahl. 2016. Indiscriminate mass surveillance and the public sphere. *Ethics and Information Technology* 18, 1 (March 2016), 33–39. https://doi.org/10.1007/s10676-016-9392-2

[174] William. Stanley Jevons. 1905. *The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal-mines* (3 ed.). Augustus M. Kelley, New York.

[175] Jack Stilgoe, Richard Owen, and Phil Macnaghten. 2013. Developing a framework for responsible innovation. *Research Policy* 42, 9 (November 2013), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

[176] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]* (June 2019). http://arxiv.org/abs/1906.02243 arXiv: 1906.02243.

[177] Shannon Sullivan and Nancy Tuana (Eds.). 2007. *Race and epistemologies of ignorance*. State University of New York Press, Albany. OCLC: ocm70676503.

[178] summerstay on Reddit. 2020. Fiction by Neil Gaiman and Terry Pratchett by GPT-3. www.reddit.com/r/slatestarcodex/comments/hmu5lm/fiction_by_neil_gaiman_and_terry_pratchett_by_gpt3/

[179] Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2019. LAMOL: LAnguage MOdeling for Lifelong Language Learning. *arXiv:1909.03329 [cs]* (Dec. 2019). http://arxiv.org/abs/1909.03329 arXiv: 1909.03329.

[180] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. *arXiv:2102.02503 [cs]* (February 2021). http://arxiv.org/abs/2102.02503 arXiv: 2102.02503.

[181] Nenad Tomasev, Kevin R. McKee, Jackie Kay, and Shakir Mohamed. 2021. Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (July 2021), 254–265. https://doi.org/10.1145/3461702.3462540 arXiv: 2102.04257.

[182] Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI WMT21 News Translation Task Submission. *arXiv:2108.03265 [cs]* (Aug. 2021). http://arxiv.org/abs/2108.03265 arXiv: 2108.03265.

[183] Evert Van den Broeck, Brahim Zarouali, and Karolien Poels. 2019. Chatbot advertising effectiveness: When does the message get through? *Computers in Human Behavior* 98 (September 2019), 150–157. https://doi.org/10.1016/j.chb.2019.04.009

[184] VersebyVerse 2020. Verse by Verse. https://sites.research.google/versebyverse/publisher:.

[185] James Vincent. 2017. The invention of AI 'gaydar' could be the start of something much worse. *The Verge* (September 2017). https://www.theverge.com/2017/9/21/16332760/ai-sexuality-gaydar-photo-physiognomy

[186] Kate Vredenburgh. 2021. The Right to Explanation. *Journal of Political Philosophy* 0, 0 (2021), 1–21. https://doi.org/10.1111/jopp.12262

[187] Carissa Véliz. 2019. Privacy matters because it empowers us all | Aeon Essays. *Aeon* (September 2019). https://aeon.co/essays/privacy-matters-because-it-empowers-us-all

[188] Daniel Wallace, Florian Tramer, Matthew Jagielski, and Ariel Herbert-Voss. 2020. Does GPT-2 Know Your Phone Number? http://bair.berkeley.edu/blog/2020/12/20/lmmem/

[189] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *arXiv:2002.10957 [cs]* (April 2020). http://arxiv.org/abs/2002.10957 arXiv: 2002.10957.

[190] Yilun Wang and Michal Kosinski. 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology* 114, 2 (February 2018), 246–257. https://doi.org/10.1037/pspa0000098

[191] William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, Montréal, Canada, 19–26. https://aclanthology.org/W12-2103

[192] Michael Webb. 2019. *The Impact of Artificial Intelligence on the Labor Market*. SSRN Scholarly Paper ID 3482150. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=3482150

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

[194] Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in Detoxifying Language Models. *arXiv:2109.07445 [cs]* (September 2021). http://arxiv.org/abs/2109.07445 arXiv: 2109.07445.

[195] Mark West, Rebecca Kraut, and Han Ei Chew. 2019. *I'd blush if I could : closing gender divides in digital skills through education*. Technical Report. UNESCO. https://repositorio.minedu.gob.pe/handle/20.500.12799/6598

[196] Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language Models are Few-shot Multilingual Learners. *arXiv:2109.07684 [cs]* (September 2021). http://arxiv.org/abs/2109.07684 arXiv: 2109.07684.

[197] Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying Language Models Risks Marginalizing Minority Voices. *arXiv:2104.06390 [cs]* (April 2021). http://arxiv.org/abs/2104.06390 arXiv: 2104.06390.

[198] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv:2010.11934 [cs]* (March 2021). http://arxiv.org/abs/2010.11934 arXiv: 2010.11934.

[199] James O. Young. 2005. Profound Offense and Cultural Appropriation. *The Journal of Aesthetics and Art Criticism* 63, 2 (2005), 135–146. https://www.jstor.org/stable/3700467

[200] Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* 112, 4 (January 2015), 1036–1040. https://doi.org/10.1073/pnas.1418680112

[201] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2021. Differentially Private Fine-tuning of Language Models. *arXiv:2110.06500 [cs, stat]* (Oct. 2021). http://arxiv.org/abs/2110.06500 arXiv: 2110.06500.

[202] Sean Zdenek. 2007. "Just Roll Your Mouse Over Me": Designing Virtual Women for Customer Service on the Web. *Technical Communication Quarterly* 16, 4 (August 2007), 397–430. https://doi.org/10.1080/10572250701380766

[203] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2020. Defending Against Neural Fake News. *arXiv:1905.12616 [cs]* (Dec. 2020). http://arxiv.org/abs/1905.12616 arXiv: 1905.12616.

[204] Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. 2021. Differentiable Prompt Makes Pre-trained Language Models Better Few-shot Learners. *arXiv:2108.13161 [cs]* (October 2021). http://arxiv.org/abs/2108.13161 arXiv: 2108.13161.

[205] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. (2022). https://doi.org/10.48550/arxiv.2205.01068

[206] Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. *arXiv:2102.09690 [cs]* (June 2021). http://arxiv.org/abs/2102.09690 arXiv: 2102.09690.

[207] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).

[208] Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. 2015. Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction. *International Journal of Social Robotics* 7, 3 (June 2015), 347–360. https://doi.org/10.1007/s12369-014-0267-6

A DEFINITIONS

A.1 Language Models

Language Models are machine learning models that are trained to represent a probability distribution $p(w_1, w_2, ..., w_n)$ over sequences of tokens w from a pre-specified domain. Typical training corpora for LMs contain natural language (e.g. collected from the web), but LMs can also be trained on other types of languages (e.g. computer programming languages). Moreover, LMs can serve different purposes, such as generating language or providing semantic embeddings. Depending on the primary purpose of a LM, slightly different architectures and training objectives can be used. In this paper, unless we specify otherwise, we focus on LMs tailored to language generation.

A standard approach to construct generative LMs is to use an autoregressive decomposition that sequentially proposes a probability distribution for the next utterance based on past utterances:

$$p(w) = p(w_1) \cdot p(w_2|w_1) \cdots p(w_T|w_1, \ldots, w_{T-1}) \ .$$

Here $w = w_1 \ldots w_T$ is a sequence of $T = |w|$ utterances. Each of the terms $p(w_t|w_1, \ldots, w_{t-1})$ with $t = 1, \ldots, T$ represents the probability the model assigns to observing the particular utterance w_t given the previous $t-1$ utterances. LMs of this form are trained by updating the parameters controlling these conditional probabilities to assign high likelihood to sequences of utterances observed in the training corpus. Training is the result of an iterative process whereby at each iteration the model is presented with a batch of utterances and its parameters are updated to increase the likelihood of that particular set of utterances. Training large-scale language models can require very high numbers of iterations, requiring significant computing power.

Recent LMs are primarily distinguished from other LMs due to their parameter size and training data. Their size allows LMs to retain representations of extremely large text corpora, resulting in much more general sequence prediction systems than prior LMs. In this report, we focus on such large-scale models.

Note that LMs do not output text directly. Rather, they produce a probability distribution over different utterances from which samples can be drawn. Greedy decoding directly from the (conditional) probability distribution provided by an LM is possible, but often performs poorly in practice. Instead, methods that focus on the most likely utterances – while introducing a small amount of variability (e.g. beam search and nucleus sampling) – have been found to produce better results in practice [78]).

A.2 "Large" Language Models

Training models with a large number of parameters on extremely large datasets such as the Colossal Clean Crawl Corpus (C4) [149], WebText [147], and MassiveText [148] resulted in sequence prediction systems with much more general applicability compared to the prior state-of-the-art [30, 54, 148, 160] as well as impressive few-shot and zero-shot learning capabilities [30, 148]. The insight that powerful sequence prediction systems could be created by scaling up the size of LMs and training corpora motivated an upsurge in interest and investment in LM research by several AI research labs. The increase in size of LMs has implications for a number of risks discussed in this paper, for example an increase in quality of outputs in some tasks may increase the risk that users give credence to misinformation provided by the model in other tasks subsection 2.3; and all else held constant, environmental cost of LMs grow with their size.

A.3 Bias

Concerns regarding "bias" in language models generally revolve around distributional skews that result in unfavourable impacts for particular social groups [82, 169]. We note that there are different definitions of "bias" in classical statistics and machine learning compared to sociotechnical studies. In classical statistics, "bias" designates the difference between a model's prediction and the ground truth [47]; in machine learning, minimising statistical bias is a component of reducing error [47]. In sociotechnical studies, "bias" refers to skews that lead to unjust discrimination based on traits such as age, gender, religion, ability status, whether or not these characteristics are legally protected [22]. Developing mechanisms to quantify the latter type of bias is an area of active research, where qualitative and quantitative measures have been established [12, 71].

A.4 Discrimination

Similarly, "discrimination" has a dual definition. Traditionally in machine learning, this term refers to making distinctions between possible categories or target classes [28]. In sociotechnical work, "discrimination" refers to unjust differential treatment, typically toward historically marginalised groups. Various steps in training a machine learning model can result in discrimination in the sociotechnical sense, from labelling and collection of the training data, to defining the "target variable" and class labels, to selecting features [13].

Table 1: Overview of six areas of ethical and social risk of harm associated with language model

Risk area	Mechanism	Type of Harm	Technical mitigation approaches	Evidence of this risk in large-scale LMs
Discrimination, Hate Speech and Exclusion	The LM accurately reflects unjust, toxic, and oppressive speech present in the training data.	• Allocational or representational harm • Profound offence or psychological harm • Inciting violence or hate • Social exclusion • Uneven performance for different social groups	• More representative training data • Curated and filtered training data • Dataset documentation • Participatory approaches for detecting instances of harm • Online learning for model updating • Training retriever model with separate data corpus • Prompt design • Explainability and interpretability research to identify fairness concerns	• Social Stereotypes: Abid et al. [2], Huang et al. [82], Lucy and Bamman [120], Nadeem et al. [133], Nozza et al. [136], Rae et al. [148] • Hate Speech: Gehman et al. [61], Rae et al. [148], Welbl et al. [194] • Exclusion: [206] • Lower Performance: Welbl et al. [194], Winata et al. [196]
Information Hazards	The LM leaks or correctly infers sensitive information	• Privacy violations • Safety risks	• Algorithmic tools such as differential privacy • Responsible release strategies	• Privacy leaks: Carlini et al. [34], Kulkarni [109]
Misinformation Harms	The LM provides false, misleading, nonsensical or poor quality information.	• Deceiving or misinforming a user • Material harm • Unethical actions by users • Growing societal distrust in shared information	• Responsible release strategy • Innovate on methods to filter out incorrect statements • Training retriever model with separate data corpus • Engineer LMs to not provide output on sensitive domains • Sociotechnical interventions such as training truthfulness via humans-in-the-loop, and shaping norms and institutions on truth in the field • Training LMs that can search and reference sources from the internet to substantiate factual statements	• Misinformation: Hendrycks et al. [72], Lin et al. [118], Nakano et al. [134], Quach [145], Rae et al. [148], Reiter [157]
Malicious Uses	Humans intentionally use the LM to cause harm.	• Undermining public discourse • Facilitating fraud, scam, impersonation crimes • Personalised disinformation campaigns • Weaponisation or production of malicious code • Augment illegitimate mass surveillance	• Limit access to the LMs and monitoring usage	• Disinformation: Bagdasaryan and Shmatikov [11], Buchanan et al. [31], Hao [69], Zellers et al. [203]
Human-Computer Interaction Harms	Humans are deceived or made vulnerable via direct interaction with a powerful conversational agent.	• Unsafe use • Creating avenues to exploit or violate privacy of the user • Perpetuating discriminatory stereotypes via product design	• More inclusive product design • Giving an assistant non-gendered or multiple voices	• Product design: Cave and Dihal [35], Zdenek [202] • Exploiting users: Ischen et al. [87], Pardes [139] • Cybersecurity Chen et al. [37]
Environmental and Socioeconomic harms	LMs are used to underpin widely used downstream applications that disproportionately benefit and harm different groups.	• Increasing social inequalities from uneven distribution of risk and benefits • Loss of high-quality and safe employment • Undermining creative industries • Environmental harm	• Architectural innovations such as training retriever model with separate data corpus • Increase training efficiency	• Environment: Bender et al. [15], Mytton [132], Patterson et al. [141], Rae et al. [148], Strubell et al. [176] • Unequally distributed benefit: Bender et al. [15]