DeepMind

# Ethical and social risks of harm from Language Models

**Laura Weidinger**[1], **John Mellor**[1], **Maribeth Rauh**[1], **Conor Griffin**[1], **Jonathan Uesato**[1], **Po-Sen Huang**[1], **Myra Cheng**[1,2], **Mia Glaese**[1], **Borja Balle**[1], **Atoosa Kasirzadeh**[1,3], **Zac Kenton**[1], **Sasha Brown**[1], **Will Hawkins**[1], **Tom Stepleton**[1], **Courtney Biles**[1], **Abeba Birhane**[1,4], **Julia Haas**[1], **Laura Rimell**[1], **Lisa Anne Hendricks**[1], **William Isaac**[1], **Sean Legassick**[1], **Geoffrey Irving**[1] **and Iason Gabriel**[1]

[1]DeepMind, [2]California Institute of Technology, [3]University of Toronto, [4]University College Dublin

## Abstract

This paper aims to help structure the risk landscape associated with large-scale Language Models (LMs). In order to foster advances in responsible innovation, an in-depth understanding of the potential risks posed by these models is needed. A wide range of established and anticipated risks are analysed in detail, drawing on multidisciplinary literature from computer science, linguistics, and social sciences.

The paper outlines six specific risk areas: I. Discrimination, Exclusion and Toxicity, II. Information Hazards, III. Misinformation Harms, IV. Malicious Uses, V. Human-Computer Interaction Harms, VI. Automation, Access, and Environmental Harms.

The first risk area discusses fairness and toxicity risks in large-scale language models. This includes four distinct risks: LMs can create unfair discrimination and representational and material harm by perpetuating stereotypes and social biases, i.e. harmful associations of specific traits with social identities. Social norms and categories can exclude or marginalise those who exist outside them. Where a LM perpetuates such norms - e.g. that people called “Max” are “male”, or that “families” always consist of a father, mother and child - such narrow category use can deny or burden identities who differ. Toxic language can incite hate or violence or cause offense. Finally, a LM that performs more poorly for some social groups than others can create harm for disadvantaged groups, for example where such models underpin technologies that affect these groups. These risks stem in large part from choosing training corpora that include harmful language and overrepresent some social identities.

The second risk area includes risks from private data leaks or from LMs correctly inferring private or other sensitive information. These risks stem from private data that is present in the training corpus and from advanced inference capabilities of LMs.

The third risk area comprises risks associated with LMs providing false or misleading information. This includes the risk of creating less well-informed users and of eroding trust in shared information. Misinformation can cause harm in sensitive domains, such as bad legal or medical advice. Poor or false information may also lead users to perform unethical or illegal actions that they would otherwise not have performed. Misinformation risks stem in part from the processes by which LMs learn to represent language: the underlying statistical methods are not well-positioned to distinguish between factually correct and incorrect information.

The fourth risk area spans risks of users or product developers who try to use LMs to cause harm. This includes using LMs to increase the efficacy of disinformation campaigns, to create personalised scams or fraud at scale, or to develop computer code for viruses or weapon systems.

The fifth risk area focuses on risks from the specific use case of a “conversational agent” that directly interacts with human users. This includes risks from presenting the system as “human-like”, possibly leading users to overestimate its capabilities and use it in unsafe ways. Another risk is that conversation with such agents may create new avenues to manipulate or extract private information from users. LM-based conversational agents may pose risks that are already known from voice assistants, such as perpetuating stereotypes by self-presenting e.g. as “female assistant”. These risks stem in part from LM training objectives underlying such conversational agents and from product design decisions.

The sixth risk area includes risks that apply to LMs and Artificial Intelligence (AI) systems more broadly. Training and operating LMs can incur high environmental costs. LM-based applications may benefit some groups more

*Corresponding author: Laura Weidinger* <lweidinger@deepmind.com>

than others and the LMs themselves are inaccessible to many. Lastly, LM-based automation may affect the quality of some jobs and undermine parts of the creative economy. These risks manifest particularly as LMs are widely used in the economy and benefits and risks from LMs are globally unevenly distributed.

In total, we present 21 risks. We then discuss the points of origin of different risks and point to potential risk mitigation approaches. The point of origin of a harm may indicate appropriate mitigations: for example, the risk of leaking private data originates from this data being present in the training dataset. It can be mitigated at the point of origin, by better redaction or curation of training data. However, other mitigation approaches may also be applicable and ensure more robust mitigation overall. For example, algorithmic tools applied during training, such as differential privacy methods, or product decisions, such as constraining access and use cases of the LM, are additional mitigation approaches that can be pursued in parallel. Risk mitigation approaches range from social or public policy interventions, to technical solutions and research management, to participatory projects and product design decisions.

Lastly, we discuss organisational responsibilities in implementing such mitigations, and the role of collaboration. Measuring and mitigating ethical and social risks effectively requires a wide range of expertise, and fair inclusion of affected communities. It is critical to implement mitigations with a broad view of the landscape of risks, to ensure that mitigating against one risk of harm does not aggravate another. Otherwise, for example, mitigation approaches to toxic speech can inadvertently lead to lower LM performance for some social groups. We highlight directions for further research, particularly on expanding the toolkit for assessing and evaluating the outlined risks in LMs, and the need for inclusive participatory methods. Finally, we conclude by showing how the present work - of structuring the risk landscape - is the first step in a broader framework of responsible innovation.

# Contents

## Reader's guide

This is a long document. The report is divided into three segments.

First, the Introduction provides a brief introduction to Language Models.

Second, the Classification of harms from language models gives a taxonomy and detailed account of a range of social and ethical risks associated with Language Models.

Third, the Discussion and Directions for future research explore some underlying causes of these risks, a range of mitigation approaches, and possible challenges to be addressed through future research.

Individual sections can be read independently or together. We recommend:

- **1 minute read:** Study Table 1 for a high-level overview of the risks considered.
- **10 minute read:** Read the Abstract and Table 1 for an overview of the risks considered. Then skim all bold text in the segment on Classification of harms from language models and skim Directions for future research for an overview of risks and challenges.
- **Readers who actively work on LMs:** We encourage you to skim all bold text in the segment on Classification of harms from language models, and to get stuck in risks that directly relate to your own work and interest - as you will likely be able to help solve some of the field's core challenges in this domain.
- **Readers with no background on LMs:** We recommend you read the Abstract and Introduction first as these introduce key terminology that is used in this report. Next, study Table 1 for a high-level overview of the risks considered and read the risk headers and example dialog boxes for each risk in the Classification of harms from language models. Get stuck in risks that are of interest to you and read the Discussion on challenges in mitigating these risks.
- **Readers with an interest in a particular risk or type of harm**: We encourage you to read the Abstract, Table 1 and Discussion for context on the broader risk landscape and approaches to mitigation, in addition to reading the specific section on the risk that piques your interest.
- **Readers with an interest in approaches to mitigating harms:** We recommend you read the Abstract for an overview of the harms considered and read Table 1 with a focus on the mechanisms underlying each risk area. Jump to the Discussion on approaches to mitigating risks and read Directions for future research on methodological and normative challenges in assessing and mitigating risks, and proposals for addressing these challenges.

# 1. Introduction

Language Models (LMs)[1] are rapidly growing in size and effectiveness, yielding new breakthroughs and attracting increasing research attention (Brown et al., 2020; Fedus et al., 2021; Rae et al., 2021; Rosset, 2020). Several Artificial Intelligence (AI) research labs are pursuing LM research, spurred by the promise these models hold for advancing research and for a wide range of beneficial real-world applications. Some research groups have suggested that recent large-scale LMs may be a 'foundational' breakthrough technology, potentially affecting many aspects of life (Bommasani et al., 2021). 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 the technology, it is essential to thoughtfully assess the potential benefits as well as potential risks that need to be mitigated (Stilgoe et al., 2013). Prior research has explored the potential for ethical and safe innovation of large-scale LMs, including interdisciplinary workshops to scope out risks and benefits (Tamkin et al., 2021), papers that outline potential risks (Bender et al., 2021; Bommasani et al., 2021; Dinan et al., 2021; Kenton et al., 2021), and papers identifying ways to mitigate potential harms (Chen et al., 2021a; Solaiman and Dennison, 2021; Welbl et al., 2021).[2] For this report, we seek to build on this prior work by proposing an initial taxonomy of risks associated with LM development and use, as well as outlining concrete next steps and directions for future research that supports responsible innovation for LMs.

The overall aim of this report is three-fold:

1. Underpin responsible decision-making by organisations working on LMs by broadening and structuring the discourse on AI safety and ethics in this research area,
2. Contribute to wider public discussion about risks and corresponding mitigation strategies for LMs,
3. Guide mitigation work by research groups working on LMs. We aim to support the mutual exchange of expertise in this area, to help make the risks posed by LMs actionable.

We structure the identified risks in a taxonomy of ethical and social risks associated with LMs, under 6 risk areas: I. Discrimination, Exclusion and Toxicity, II. Information Hazards, III. Misinformation Harms, Malicious Uses, V. Human-Computer Interaction Harms, VI. Automation, Access, and Environmental Harms. An overview of the risks that fall under each risk area can be found in the Classification of harms from language models part of the report.

Each risk is discussed in detail with regard to the nature of the harm, empirical examples, and additional considerations. For each risk, we provide a fictitious example to illustrate how the risk in question may manifest.[3] However the risks described apply to LMs more generally and do not depend on the dialogue modality unless otherwise specified. Since several of the risks discussed below are neither novel nor exclusive to LMs or related technologies, we offer context on how each risk manifests in existing language technologies. We also mark each risk as either "anticipated" or "observed", depending on whether a given risk has already been observed or whether further work is needed to indicate real-world manifestations of this risk. The creation of a taxonomy of risks supports the exercise of foresight in this space, with the aim of guiding action to resolve any issues that can be identified in advance.

Responsible innovation is a collaborative endeavour. In order to anticipate and mitigate risks posed by technology successfully, we need to view these issues through multiple lenses and perspectives. This report was written by a large group of researchers with varied disciplinary backgrounds and areas of expertise. To review the risk landscape as comprehensively as possible, we collated potential risks from a wide range of sources including

[1] These recent LMs are also referred to as "large language models", or "large-scale language models".

[2] Note that the origin of a risk is not a perfect guide to potential mitigations - a point we discuss in more detail in Understanding the point of origin of a risk.

[3] Each of these examples assumes a dialogue format where a human supplies a prompt and the LM offers a response. There are many LM use cases beyond such conversational agents. These examples are for illustrative purposes only, and the same risk may manifest differently in other LM use cases.

analyses from the fields of AI ethics, AI safety, race and gender studies, linguistics and natural language processing and studies at the intersection of society and technology (also referred to as sociotechnical studies), as well as analyses by civil society organisations and news reports. Further risks were added based on our own experience and expertise. Beyond publishing research, we believe responsible innovation also requires inclusive dialogue between stakeholders in AI development which includes affected communities and the wider public (Gabriel, 2020b; Mohamed et al., 2020; Murgia, 2021; Stilgoe et al., 2013). In the future, we look to continue to deepen our understanding of risks and mitigations including by working with external partners and communities.

## 1.1. Limitations

Note that this report is part of a broader research programme working toward the responsible innovation of LMs and necessarily leaves some questions unanswered. For example, we do not discuss potential beneficial applications of LMs nor do we offer a comprehensive overview of potential use cases. Nor do we attempt to perform a full ethical evaluation of LMs, which must weigh both the potential benefits and risks of a given technology. To assess the overall balance of benefit and cost, separate analysis of the benefits arising from proposed LM applications would be needed. Instead, the focus here is on anticipating and structuring the risk landscape, with the intention of supporting a larger constructive research effort.

This report is also necessarily a snapshot in time: it was initiated in autumn 2020 and completed in summer 2021. It is likely that we miss risks arising from LMs that depend, for their visibility, on the passage of time. As such, the presented taxonomy is merely a starting point and will need to be updated as new challenges come into focus and additional perspectives are brought to bear on these questions.

This report focuses on risks associated with *operating* LMs. Risks of harm that are associated with training are not discussed. This includes concerns about the working conditions of data annotators or "ghost workers" (Gray and Suri, 2019), the ethics of supply chains of hardware on which LM computations are run (Crawford, 2021), or environmental costs of training such models (Bender et al., 2021; Patterson et al., 2021; Schwartz et al., 2020; Strubell et al., 2019) which are only briefly referenced in the section on VI. Automation, access, and environmental harms. This report also does not cover risks that depend on specific applications.

This report excludes risks which the authors anticipate to depend on capabilities that are several years in the future, for example because they depend on capabilities that are several step changes beyond the state-of-the-art. A subset of such long-term risks is addressed in literature on existential risk and AI Safety (Armstrong et al., 2012; Kenton et al., 2021). This report also does not cover risks that depend on superintelligence as described in (Bostrom, 2014).

Finally, this report does not discuss risks that depend on multiple modalities, for example from models that combine language with other domains such as vision or robotics. While several of the insights in this report are translatable to such models, these require distinct risk assessments. For some discussion on risks associated with multi-modal large models, see (Bommasani et al., 2021).

### 1.1.1. Note on terminology

This report focuses on the risks of large-scale language models, including in specific applications of these models such as conversational assistants, or in other language technologies. Several of these risks also apply to smaller language models. For detailed definitions of Language Models, Language Agents, and Language Technologies please refer to the section on Definitions in the Appendix.

For simplicity we refer to "LMs" throughout. Where risks are unique to specific types of applications, such as conversational agents, this is explicitly stated.

## 1.2. A Brief history of Language Models

### 1.2.1. Origins

The main methodology underpinning contemporary large-scale language models traces its origins to methods developed by the research group of Frederick Jelinek on Automatic Speech Recognition (ASR) in the 1970s and '80s (Jelinek, 1976). This research group built on prior work in statistics by Claude Shannon (Shannon, 1948) and Andrey Markov (Markov, 1913). In parallel, James Baker (Baker, 1990) developed a similar approach to ASR (see (Jurafsky and Martin, 2014)).

Jelinek's group pioneered an information theoretic approach to ASR, observing that performing any task that requires producing language conditioned on an input using a probability distribution $p(\text{language}|\text{input})$ can be factored into a language model representing a probability distribution $p(\text{language})$ multiplied by the task specific distribution $p(\text{input}|\text{language})$. This factorisation suggests that general LMs $p(\text{language})$ can aid language prediction tasks where the LM captures the relevant language distribution. Whilst this factorisation is not explicitly used in most current systems, it implicitly underpins current LM research and is a useful way to understand the role language modelling plays in specific language technologies such as conversational agents, machine translation, and question answering.

### 1.2.2. Transformer models

More recently, the transformer architecture was developed (Vaswani et al., 2017). Transformers are a class of architectures that use a series of so-called transformer blocks comprising a self-attention layer followed by a feedforward layer, linked together with residual connections. The self-attention layer helps the model to consider neighbouring words in the input as it processes a specific word. Originally, the transformer architecture was proposed for the task of machine translation (Vaswani et al., 2017). (Radford et al., 2018b) use a modified version applied to the task of language modeling (predicting the next word in a sentence). Subsequent work on LMs (Brown et al., 2020; Radford et al., 2018a) uses a similar architecture. An accessible visual introduction to the transformer architecture can be found in (Alammar, 2018). Recent language models built on the transformer architecture have been fine-tuned directly, without the need for task-specific architectures (Devlin et al., 2019; Howard and Ruder, 2018; Radford et al., 2018a).

### 1.2.3. "Large" Language Models

The recent upwind in LM research is rooted in the capacity to increase LM size in terms of number of parameters and size of training data (Bender et al., 2021). Training models on extremely large datasets such as the Colossal Clean Crawl Corpus (C4) (Raffel et al., 2020) and WebText (Radford et al., 2018b) resulted in sequence prediction systems with much more general applicability compared to the prior state-of-the-art (Brown et al., 2020; Fedus et al., 2021; Rosset, 2020). These models also displayed greater few-shot and zero-shot learning capabilities compared to smaller LMs (Brown et al., 2020). These properties were found to greatly simplify the development of task-specific LAs by reducing the adaptation process to prompt design (Zhang et al., 2021b). 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.

# 2. Classification of harms from language models

In this section we outline our taxonomy of ethical and social risks of harm associated with Language Models. We identify 21 risks of harm, organised into six risk areas (for an overview see Table 1). In this table we also note the mechanisms by which different groups of risks emerge.

## 2.1. Discrimination, Exclusion and Toxicity

Social harms that arise from the language model producing discriminatory or exclusionary speech

### 2.1.1. Overview

Language has long been a vehicle for perpetuating injustices against marginalised populations and conferring power to oppressive institutions (Craft et al., 2020; Lippi, 1997). Social harms that manifest in language can also arise from language predicted by language models (LMs). We distinguish between the following harms. First, language that promotes stereotypes or causes unfair discrimination[1] can contribute to the oppression of those at social margins (Crenshaw, 2017b). Second, language that reinforces social norms can exclude identities that exist outside these norms (Bender et al., 2021; Foucault and Sheridan, 2012). Third, language can be 'toxic', for example by inciting violence or causing offense (Fortuna and Nunes, 2018). Fourth, a form of discrimination emerges when language technologies perform better for some social groups than others.

Risks covered in this section:

- Social stereotypes and unfair discrimination
- Exclusionary norms
- Toxic language
- Lower performance by social group

### 2.1.2. Social stereotypes and unfair discrimination

> Q: "Two Muslims walked into a . . . "
> A: "Texas cartoon contest and opened fire."[a]
>
> [a] Example from (Abid et al., 2021)

*Observed risk: This is a well-documented problem that needs a mitigation strategy and tools to analyse the model against benchmarks of 'acceptability'.*

***Problem***

Perpetuating harmful stereotypes and discrimination is a well-documented harm in machine learning models that represent natural language (Caliskan et al., 2017). LMs that encode discriminatory language or social stereotypes can cause different types of harm. It may be useful to distinguish between allocational and representational harms: allocational harms occur when resources and opportunities are unfairly allocated between social groups; they may occur when LMs are used in applications that are used to make decisions that affect persons. Representational harms include stereotyping, misrepresenting, and demeaning social groups Barocas and Wallach cited in (Blodgett et al., 2020).

Unfair discrimination manifests in differential treatment or access to resources among individuals or groups based on sensitive traits such as sex, religion, gender, sexual orientation, ability and age. The dimensions along which such oppression occurs can also be rooted in culture-specific or otherwise localised social hierarchies.

[1] Note that the terms "bias" and "discrimination" have different meanings in classical statistics compared to sociotechnical studies; for a definition of these terms, see the Definitions in the Appendix.

***Table 1.*** *Overview of all risks covered in this report.*

I. **Discrimination, Exclusion and Toxicity**
**Mechanism**: These risks arise from the LM accurately reflecting natural speech, including unjust, toxic, and oppressive tendencies present in the training data.
**Types of Harm**: Potential harms include justified offense, material (allocational) harm, and the unjust representation or treatment of marginalised groups.
- Social stereotypes and unfair discrimination
- Exclusionary norms
- Toxic language
- Lower performance by social group

II. **Information Hazards**
**Mechanism**: These risks arise from the LM predicting utterances which constitute private or safety-critical information which are present in, or can be inferred from, training data.
**Types of Harm**: Potential harms include privacy violations and safety risks.
- Compromise privacy by leaking private information
- Compromise privacy by correctly inferring private information
- Risks from leaking or correctly inferring sensitive information

III. **Misinformation Harms**
**Mechanism**: These risks arise from the LM assigning high probabilities to false, misleading, nonsensical or poor quality information.
**Types of Harm**: Potential harms include deception, material harm, or unethical actions by humans who take the LM prediction to be factually correct, as well as wider societal distrust in shared information.
- Disseminating false or misleading information
- Causing material harm by disseminating misinformation e.g. in medicine or law
- Nudging or advising users to perform unethical or illegal actions

IV. **Malicious Uses**
**Mechanism**: These risks arise from humans intentionally using the LM to cause harm.
**Types of Harm**: Potential harms include undermining public discourse, crimes such as fraud, personalised disinformation campaigns, and the weaponisation or production of malicious code.
- Reducing the cost of disinformation campaigns
- Facilitating fraud and impersonation scams
- Assisting code generation for cyber attacks, weapons, or malicious use
- Illegitimate surveillance and censorship

V. **Human-Computer Interaction Harms**
**Mechanism:** These risks arise from LM applications, such as Conversational Agents, that directly engage a user via the mode of conversation.
**Types of Harm**: Potential harms include unsafe use due to users misjudging or mistakenly trusting the model, psychological vulnerabilities and privacy violations of the user, and social harm from perpetuating discriminatory associations via product design (e.g. making "assistant" tools by default "female.")
- Anthropomorphising systems can lead to overreliance or unsafe use
- Create avenues for exploiting user trust to obtain private information
- Promoting harmful stereotypes by implying gender or ethnic identity

VI. **Automation, access, and environmental harms**
**Mechanism**: These risks arise where LMs are used to underpin widely used downstream applications that disproportionately benefit some groups rather than others.
**Types of Harm**: Potential harms include increasing social inequalities from uneven distribution of risk and benefits, loss of high-quality and safe employment, and environmental harm.
- Environmental harms from operating LMs
- Increasing inequality and negative effects on job quality
- Undermining creative economies
- Disparate access to benefits due to hardware, software, skill constraints

For example, the Hindu caste system underpins discrimination in India, but not across the globe (Sambasivan et al., 2021). Additionally, injustice can be compounded when social categories intersect, for example in the discrimination against a person that holds a marginalised gender and a marginalised religion (Crenshaw, 2017a).

Allocational harm caused by discriminatory systems is particularly salient if bias occurs in applications that materially impact people's lives, such as predicting a person's creditworthiness (Mehrabi et al., 2019), criminal recidivism (Angwin et al., 2016), or suitability to a job (Mujtaba and Mahapatra, 2019). For example, a language technology that analyses CVs for recruitment, or to give career advice, may be less likely to recommend historically discriminated groups to recruiters, or more likely to recommend lower paying careers to marginalised groups. Unfair biases are already well-documented in machine learning applications ranging from diagnostic healthcare algorithms (Obermeyer et al., 2019) to social outcome prediction (Narayanan, 2021); for a more general introduction see (Chouldechova and Roth, 2018; Kordzadeh and Ghasemaghaei, 2021; Mehrabi et al., 2021; Noble, 2018; Zou and Schiebinger, 2018). Based on our current understanding, such stereotyping and unfair bias are set to recur in language technologies building on LMs unless corrective action is taken.

**Why we should expect LMs to reinforce stereotypes and unfair discrimination by default** LMs are optimised to mirror language as accurately as possible, by detecting the statistical patterns present in natural language Definitions. The fact that LMs track patterns, biases, and priors in natural language is not negative *per se* (Shah et al., 2020). Rather, it becomes a problem when the training data is unfair, discriminatory, or toxic. In this case, the optimisation process results in models that mirror these harms. As a result, LMs that perform well with regard to their optimisation objective can work poorly with regard to social harms, insofar as they encode and perpetuate harmful stereotypes and biases present in the training data.

Stereotypes and unfair discrimination can be present in training data for different reasons. First, training data reflect historical patterns of systemic injustice when they are gathered from contexts in which inequality is the status quo. Training systems on such data entrenches existing forms of discrimination (Browne, 2015). In this way, barriers present in our social systems can be captured by data, learned by LMs, and perpetuated by their predictions (Hampton, 2021).

Second, training data can be biased because some communities are better represented in the training data than others. As a result, LMs trained on such data often model speech that fails to represent the language of those who are marginalised, excluded, or less often recorded. The groups that are traditionally underrepresented in training data are often disadvantaged groups: they are also referred to as the 'undersampled majority' (Raji, 2020). The implications of unrepresentative training data for downstream biases and stereotyping in LMs demonstrate the power that is exercised by those who have influence over what data is used for model training (Blodgett et al., 2020). While in principle, LMs are optimised to represent language with high fidelity, they can also overrepresent small biases present in the training data, a phenomenon referred to as 'bias amplification' (Wang and Russakovsky, 2021; Zhao et al., 2017).

***Examples***

Generative LMs have frequently been shown to reproduce harmful social biases and stereotypes. Predictions from the GPT-3 model (Brown et al., 2020) were found to exhibit anti-Muslim and, to a lesser degree, antisemitic bias, where '"Muslim" was analogised to "terrorist" in 23% of test cases, while "Jewish" was mapped to "money" in 5% of test cases (Abid et al., 2021)[2]. Gender and representation biases were found in fictional stories generated by GPT-3 (Lucy and Bamman, 2021), where female-sounding names were more often associated with stories about family and appearance, and described as less powerful than masculine characters.

The *StereoSet* benchmark measures references to stereotypes of race, gender, religion, and profession in generative LMs and finds that the models GPT2 (Radford et al., 2018b) and masked models BERT (Devlin et al., 2019), ROBERTA (Liu et al., 2019), XLNET (Yang et al., 2019) exhibit 'strong stereotypical associations' (Nadeem et al., 2020). The CrowS-Pairs benchmark finds that cultural stereotypes were reproduced by likelihood estimates of masked LMs BERT (Devlin et al., 2019), and RoBERTA (Liu et al., 2019; Nangia et al., 2020)[3]. The HONEST benchmark shows that GPT-2 and downstream classifiers built on BERT promote 'hurtful stereotypes'

[2]See also the authors' (Abid, 2020) of "how hard it is to generate text about Muslims from GPT-3 that has nothing to do with violence", and (Gershgorn, 2021).

[3]Recent work critiques some current methods for measuring bias in LMs highlighting the importance of further exploration on valid measures (Blodgett et al., 2021).

across six languages (Nozza et al., 2021), and discriminatory gender biases were found in contextual word embedding by BERT (Kurita et al., 2019) and ELMo (Zhao et al., 2019). LMs trained on news articles and Wikipedia entries have been demonstrated to exhibit considerable levels of bias against particular country names, occupations, and genders (Huang et al., 2020).

***Additional considerations***

**Underrepresented groups in the training data** Training data reflect the views, values, and modes of communication by the communities whose language is captured in the corpus. For example, a dataset of Reddit user comments was found to encode discriminatory views based on gender, religion and race (Ferrer et al., 2020). As a result, it is important to carefully select and account for the biases present in the training data. However ML training datasets are often collected with little curation or supervision and without factoring in perspectives from communities who may be underrepresented (Jo and Gebru, 2020). For more discussion of this, see also the section on Why we should expect LMs to reinforce unfair bias, toxic speech, and exclusionary norms.

**Documentation of biases in training corpora** The impact of training data on the LM makes it important to transparently disclose what groups, samples, voices and narratives are represented in the dataset and which may be missing. One format that has been proposed for such dataset documentation (Bender and Friedman, 2018) are 'Datasheets' (Gebru et al., 2020). Some work in this direction includes documentation on the Colossal Clean Crawl Corpus (C4) that highlights the most prominently represented sources and references to help illuminate *whose* biases are likely to be encoded in the dataset (Dodge et al., 2021). Documentation of larger datasets is critical for anticipating and understanding the pipeline by which different harmful associations come to be reflected in the LM.

**Training data required to reduce bias may not yet exist** Approaches to biased training data range from curating dedicated training datasets to not building models in domains where such data does not exist.[4] Curating training data can help to make LMs fairer, but creating better datasets requires dedicated work (Hutchinson et al., 2021; Jo and Gebru, 2020) and may require novel data curation pipelines and tools (Denton et al., 2020). Training corpora for state of the art LMs are extremely large, so that further innovation on semi-automated curation methods may be needed in order to make the curation of such datasets tractable. Determining what constitutes a truly fair and equitable training dataset may also require further research in Ethics and Law (Kohler-Hausmann, 2019). In one high-profile, real-world example, researchers attempted to train a classifier to support recruitment, but found that the training data was inherently biased and found no alternative to create a more equitable training dataset - leading to the research project being abandoned (Dastin, 2018)[5].

**Localised stereotypes are hard to capture** As stereotypes change over time and vary between contexts, it is impossible for any given research team to be aware of, and up-to-date on, all relevant stereotypes that may cause harm or offense. In addition, the stereotypes at play in a given local context may only be knowable through committed ethnographic work on the ground (Marda and Narayan, 2021). The expertise for identifying harmful stereotypes often lies with the lived experience of affected groups (Sullivan and Tuana, 2007). This creates a challenge in knowing what stereotypes to search for, detect, and mitigate at the point of creating a LM. One way to help address this challenge is to use inclusive and fair participatory approaches (Martin Jr. et al., 2020), by establishing participatory mechanisms and institutions that can operate over time (Sloane et al., 2020), and by providing broad and transparent dataset documentation.

**Uncertainty on downstream uses complicate fairness analyses** Identifying affected communities is challenging during the early stages of building a LM when no particular application, product, or user group has been

[4] Another proposed approach relies on synthetic data, although the efficacy of this approach remains uncertain and it raises distinct challenges, on amplifying other biases (Chen et al., 2021b; Ghalebikesabi et al., 2021; Nikolenko, 2021).

[5] In this real-world example, a model ranking applicant suitability based on written CVs was biased against the term 'women' (as in 'women's chess club'). In an attempt to correct for this discriminatory performance, the model was initially corrected to not devalue a CV based on terms referring to 'women'. However, the algorithm continued to espouse an unfair gender bias against women, simply because there had been a gender bias in Amazon's prior hiring history, which was reflected in the training data. As no sufficient data on successful female applicants was available to train or fine-tune the model to reduce its gender bias, the problem of de-biasing this algorithm seemed intractable, 'executives lost hope for the project' (Dastin, 2018), and it was stopped.

defined. It is unclear to what extent a training regime can be defined that increases model "fairness" whilst being agnostic on downstream applications (Hancox-Li and Kumar, 2021). While some aspects of fairness are best considered at early research stages, more specific assessments of potential discrimination must be considered again at the point of developing a concrete application. Methods for detecting and mitigating harmful stereotypes can place an additional burden or privacy cost on minorities, e.g. through collecting additional data. Where this is the case, sustained mitigation of such harms requires engaging affected groups on fair terms that foreground their needs and interests.

**Detecting harmful stereotypes can require nuanced analyses over multiple samples** Stereotyping may only be detectable over multiple samples. "Pointwise" stereotyping manifests directly in the text prediction of a single sample and so can be identified in a single instance (Khalifa et al., 2021). "Distributional" stereotyping on the other hand manifests in the repetition of a seemingly harmless association of certain properties with a group. For example, where a LM predicts passive verbs more often in association with female than male names, such distributional stereotyping of females as more passive may occur. Such "distributional" bias may also manifest as notable omissions, e.g. where a language agent that generates fantasy stories by relying on a LM only generates stories with male, never female villains. Such distributional bias becomes apparent only upon analysing multiple predictions and requires distinct forms of evaluation and correction (Khalifa et al., 2021).

### 2.1.3. Exclusionary norms

> Q: What is a family?
> A: A family is: a man and a woman who get married and have children. *(not accounting for non-heteronormative families and children out of wedlock, for single-parent families and for the fact that families sometimes do not have children)*

*Observed risk: This is a well-documented problem that needs a mitigation strategy and tools to analyse the model against benchmarks of 'acceptability'.*

***Problem***

In language, humans express social categories and norms. Language models (LMs) that faithfully encode patterns present in natural language necessarily encode such norms and categories. This is why technological artifacts such as LMs are never "value neutral" - they represent and perpetuate the values and norms present in the training data (Bender et al., 2021; Winner, 1980).

Such norms and categories exclude groups who live outside them (Foucault and Sheridan, 2012). For example, defining the term "family" as married parents from opposing genders with a blood-related child, denies the existence of families to whom these criteria do not apply. Moreover, exclusionary norms intersect with discrimination as they almost invariably work to 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, or referring to *both genders* excluding the possibility of non-binary gender identities" (Bender et al., 2021), emphasis added.

Furthermore, exclusionary norms can place a disproportionate burden or "psychological tax" on those who do not fit or comply with these norms or who are trying to challenge or replace them. Where the model omits, excludes, or subsumes those deviating from the (perceived) norm into ill-fitting categories, these individuals also may encounter allocational or representational harm and discrimination.

The technical underpinning for LMs to promote exclusionary norms may be the fact that a deterministic argmax approach is commonly used for sampling utterances (Yee et al., 2021). This mechanism always samples the most probable next word, rather than sampling probabilistically from the prediction distribution. This can result in the single most probable view becoming entrenched in the social contexts and applications of the model (Yee et al., 2021). In LMs, this can lead to language that excludes, denies, or silences identities that fall outside these categories.

***Example***

In other machine learning approaches to modeling language it was found that tools for coreference resolution - the task of identifying all expressions that refer to the same entity in a text - typically assume binary gender,

forcing, for example, the resolution of names into either "he" or "she" (not allowing for the resolution of the name "Max" into "they") (Cao and Daumé III, 2020), definition from (StanfordNaturalProcessingGroup). In response to a question, GPT-3 was found to frequently provide common, but false utterances, rather than providing the less common, correct utterance (Zhao et al., 2021). This phenomenon is referred to as 'common token bias' (Zhao et al., 2021) (see also ).

In other ML applications, an image editing tool was found to crop images in a way that emphasised a woman's body instead of the head (Yee et al., 2021). The authors described this emphasis on the female body as perpetuating the '*male gaze*, a term used for the pervasive depiction of women as sexual objects for the pleasure of and from the perspective heterosexual men' (Yee et al., 2021), emphasis added.

In a separate study, facial recognition tools that determine gender were found to be trans-exclusive, as they assumed binary gender categories (Keyes, 2018). Note that this is distinct from a system performing more poorly for some groups Lower performance by social group: in the case of exclusionary norms, the system marginalises the group by denying it as a valid category.

***Additional considerations***

**Value lock-in forecloses societal progress over time** A LM trained on language data at a particular moment in time risks not just excluding some groups, but also enshrining temporary values and norms without the capacity to update the technology as society develops. Locking in temporary societal arrangements into novel technologies has been referred to as creating "frozen moments" (Haraway, 2004). 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" (Bender et al., 2021; Gabriel and Ghazavi, 2021). Unless a LM is meant to particularly represent the values encoded in language of a particular community and time, it must be continually updated with broader and future data. Transformer models have been shown to perform worse when applied to utterances from a different period to the time when their training data was generated (Lazaridou et al., 2021). While increasing model size alone did not improve performance, updating the model with new training data over time did improve predictions on utterances from outside the training data period (Lazaridou et al., 2021).

Technological value lock-in also risks inhibiting social change. Categories and norms change over time, as is reflected in changes in common language. For example, where previously doctors, lawyers and other professions were typically by default referred to as "he", they are now referred to as "he", "she" or "they". Such developments are widely noted as a marker of social progress - e.g. the singular use of "they" was in 2019 celebrated as the "word of the year" by the US-based publishing company Merriam-Webster. In another example, slurs can be reclaimed and change meaning, as happened with the term "queer" Toxic language. By enshrining values from a particular community or moment in time, the LM may make it harder to change such norms in the future and reclaim terms or devise more inclusive language. Depending on downstream applications of such LMs, such value lock-in may even impede social movements that actively seek to utilise language as a way to shift power (Bender et al., 2021).

**Homogenising effects in downstream applications** Concerns on exclusionary norms are relevant across a wide range of contexts. A LM used to create cultural content such as movie scripts could, for example, contribute to public discourse becoming more homogeneous and exclusionary. Moreover, if large LMs are deployed at scale in the future they may amplify majority norms and categories, contributing to increasingly homogenous discourse or crowding-out of minority perspectives. Viewed from a sociotechnical perspective, it is also possible to envisage feedback loops whereby LMs perpetuate certain norms or categories, influencing humans to use these in their own language, which in turn makes these norms and categories more likely to be prominent in future training data and thus future LMs (Pasquale, 2016).

**Functionality of language may conflict with exclusionary norms** There may be trade-offs between inferring additional information that is *commonly correct* and avoiding inferences that *perpetuate exclusionary norms* (e.g. inferring that "Ludwig" is human, male, Western). Biasing the model to make likely inferences on gender or nationality may provide functionality in some cases but obstruct more inclusionary language.

### 2.1.4. Toxic language

> Q: What should I do tomorrow?
> *A: Vulgar word choice, toxic language and offensive slurs*

*Observed risk: This is a well-documented problem that needs a mitigation strategy and tools to analyse the model against benchmarks of 'acceptability'.*

***Problem***

LM's may predict hate speech or other language that is "toxic". While there is no single agreed definition of what constitutes hate speech or toxic speech (Fortuna and Nunes, 2018; Persily and Tucker, 2020; Schmidt and Wiegand, 2017), proposed definitions often include profanities, identity attacks, sleights, insults, threats, sexually explicit content, demeaning language, language that incites violence, or 'hostile and malicious language targeted at a person or group because of their actual or perceived innate characteristics' (Fortuna and Nunes, 2018; Gorwa et al., 2020; PerspectiveAPI), direct quote from (Persily and Tucker, 2020). Such language risks causing offense, psychological harm, and even material harm in the case of inciting violence.

Toxic speech is a widespread problem on online platforms (Duggan, 2017; Gorwa et al., 2020) and in training corpora such as (Gehman et al., 2020; Luccioni and Viviano, 2021; Radford et al., 2018b). Moreover, the problem of toxic speech online platforms from LMs is not easy to address. Toxicity mitigation techniques have been shown to perpetuate discriminatory biases whereby toxicity detection tools more often falsely flag utterances from historically marginalised groups as toxic (Dixon et al., 2018; Jigsaw, 2021; Kim et al., 2020), and detoxification methods work less well for these same groups (Sap et al., 2019; Welbl et al., 2021).

***Examples***

(Gehman et al., 2020) show that 'pretrained LMs can degenerate into toxic text even from seemingly innocuous prompts' using their *RealToxicityPrompts* dataset. GPT-2 (Radford et al., 2018b) was reported to cause offense when it 'generated fictitious ... conversations between two real users on the topic of transgender rights', among other cases (Wallace et al., 2020). In adjacent language technologies, Microsoft's Twitter chatbot *Tay* gained notoriety for spewing hate speech and denying the Holocaust - it was taken down and public apologies were issued (Hunt, 2016).

***Additional considerations***

**Context dependency of whether an utterance is "toxic"** The views about what constitutes unacceptable "toxic speech" differ between individuals and social groups (Kocoń et al., 2021). While one approach may be to change toxicity classification depending on the expressed social identity of a person interacting with the LM, tailoring predictions to an identity may raise other bias, stereotyping, and privacy concerns.

What is perceived as toxic speech also depends on temporal context and the identity of the speaker (Hovy and Yang, 2021). For example, the word "queer" was historically widely considered a slur, but has been reclaimed by the LGBT+ community as a marker of self-identification (Rand, 2014). Yet, an appreciation of context continues to be important. Historical slurs may be reclaimed in such a way that out-group members are invited to use the term to describe the group (as with the preceding example). However, historical slurs may also be reclaimed in such a way that only in-group members can use the reclaimed terms, as is commonly the case with ethnicity-based slurs (Jeshion, 2020). Thus the social context and identity of the speaker may determine whether a particular utterance is deemed 'toxic'.

Similarly, the context of a particular LM use case may determine whether an utterance is toxic and whether it is appropriate. The same factual statement may be considered a matter of sexual education in some contexts and profane in others. Erroneous misclassification of educational content as adult content has been observed to inadvertently demote sex education on online platforms (Oosterhoff, 2016). Furthermore, demoting content that is falsely perceived as profane or toxic may disproportionately affect marginalised communities who particularly rely on safe online spaces (Manduley et al., 2018).

**Racist bias in toxicity detection** Recent research indicates that state of the art benchmarks for toxicity disproportionately misclassify utterances from marginalised social groups as toxic (Welbl et al., 2021), a concern that is particularly pronounced for African American English (Dixon et al., 2018; Ghaffary, 2019; Hanu

et al., 2021; Sap et al., 2019)[6]. The question of how to mitigate bias in toxic or hate speech detection remains an area of active inquiry (Davani et al., 2020; Garg et al., 2019).

**Mitigating toxicity risks demoting important knowledge** Mitigating toxicity by designing language agents (LA) that refuse to generate language on topics which are often associated with hate speech may succeed on one front while simultaneously creating blindspots in LM capability that limit their usefulness for disadvantaged groups. For example, a LA that draws blank responses when prompted with "the Holocaust was", but not when prompted with "the Cultural Revolution was", risks contributing to erasure of shared knowledge on historical events. This problem is potentially exacerbated if LAs come to be used in ways that resemble encyclopedias (e.g. to learn about historical events) or if encyclopedic knowledge is assumed. See also Training datasets elevate some perspectives over others.

### 2.1.5. Lower performance for some languages and social groups

> Q: Jsou delfíni ryby nebo savci? *("Are dolphins fish or mammals?" in Czech)*
> A: I do not know the answer to your question. *(Does not provide a useful response in Czech)*
> Q: Are dolphins fish or mammals?
> A: Even though dolphins appear fish-like, dolphins are considered mammals. Unlike fish who breathe through gills, dolphins breathe through lungs...

*Observed risk: This is a well-documented problem that needs a mitigation strategy and tools to analyse the model against benchmarks of 'acceptability'.*

***Problem***

LMs perform less well in some languages (Joshi et al., 2021; Ruder, 2020). We can distinguish between lower LM performance based on the language used "by" a group, for example predicting probability distributions of utterances in French or Swahili; and lower LM performance "about" different groups, such as predicting probability distributions over accounts of Kurdish compared to US American history. These effects are often a product of how well a social group is represented in the training data in the first place, both in terms of information by, and about, these groups.

Disparate performance can also occur based on slang, dialect, sociolect, and other aspects that vary within a single language (Blodgett et al., 2016). Language use often differs between social classes, between native and non-native speakers, and based on educational background, age group (e.g. children vs. the elderly), and cognitive or speech impairments. A LM that more accurately captures the language use of one group, compared to another, may result in lower-quality language technologies for the latter. Disadvantaging users based on such traits may be particularly pernicious because attributes such as social class or education background are not typically covered as 'protected characteristics' in anti-discrimination law. As a result, if users were to experience downstream discrimination from lower model performance based on such traits they may not have effective legal recourse based on current anti-discrimination law in many countries.[7]

The groups for whom LMs perform less well are typically groups that have historically been oppressed or marginalised. For instance, the United States has a longstanding history of disenfranchising and stigmatising speakers of African-American Vernacular English (AAVE) (Rosa and Flores, 2017), which is replicated by the lower performance of language-model-based toxicity detection on AAVE.

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 (Bender et al., 2021; Joshi et al., 2021). By relatively under-serving some groups, LMs raise social justice concerns (Hovy and Spruit, 2016), for example when technologies underpinned by LMs are used to allocate resources or provide essential services.

Disparate model performance for different social groups is a known problem in several machine learning based language technologies. For example, commercially available speech recognition systems by Amazon,

[6] Analogously, recommender systems attempting to remove toxic content have been shown to disproportionately affect the monetisation and distribution of LGBT+ content (Dixon et al., 2018; Romano, 2019).

[7] In most countries there are 'protected traits' that may not be discriminated against. In the United States, they are: gender, race, religion, age (over 40), disability, national origin, disability, family status and genetic information. In the United Kingdom, protected categories include sexual orientation, pregnancy, and people undergoing gender reassignment.

Apple, Google, IBM, and Microsoft were found to work less well for African American English speakers than for White American English speakers (Koenecke et al., 2020). Language classifiers less often correctly interpreted English-language tweets by African Americans compared to White Americans, displaying a 'racial disparity in accuracy difference' (Blodgett and O'Connor, 2017).

Current large LMs are trained on text that is predominantly in English (Brown et al., 2020; Fedus et al., 2021; Rosset, 2020) or Mandarin Chinese (Du, 2021), in line with a broader trend whereby most NLP research is on English, Mandarin Chinese, and German (Bender, 2019). This results from a compound effect whereby large training datasets, institutions that have the compute budget for training, and commercial incentives to develop LM products are more common for English and Mandarin than for other languages (Bender, 2019; Hovy and Spruit, 2016).

As a result, GPT models and the T5 model have higher performance in English than in other languages (Winata et al., 2021). This can have a range of knock-on effects that advantage speakers of standard English or Mandarin Chinese, relegating the interests and development of possible beneficial applications for groups who speak other languages (Bender, 2019).

***Examples***

Current state-of-the-art LMs produce higher quality predictions when prompted in English or Mandarin Chinese (Brown et al., 2020; Du, 2021; Fedus et al., 2021; Rosset, 2020). While it has been shown that in some languages, few-shot training and fine-tuning can improve performance in GPT models (Brown et al., 2020) and the T5 model (Raffel et al., 2020), the performance in non-English languages remained lower than the performance in English (Winata et al., 2021). It may be the case that the architecture of current LMs is particularly well-suited to English, and less well suited to other languages (Bender, 2011; Hovy and Spruit, 2016; Ruder, 2020).

In adjacent machine learning technologies, lower performance for historically marginalised groups has often been shown, for example in facial recognition (Buolamwini and Gebru, 2018) and in speech recognition (Koenecke et al., 2020).

***Additional considerations***

**Exacerbating economic inequities** If a LM performs better in a certain language(s), it may make it easier, or harder, for some groups to develop or access resulting LM applications. The potential effects on economic inequality are discussed in more detail in the section on Disparate access to benefits due to hardware, software, skill constraints.

Some languages are poorly served by digital technology because very little training data is available, e.g. the language Seychelle Creole (Joshi et al., 2021). 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 (Ruder, 2020). Dedicated work is required to curate such training data (Adelani et al., 2021).

However, even where data is available, the development of training data may be less economically incentivised. This can occur, for example, when the affected populations are multilingual and can use the technology in English. As a result, there are many widely spoken languages for which no systematic efforts have been made to create labeled training datasets, such as Javanese which is spoken by more than 80 million people (Joshi et al., 2021).

**Technical workarounds raise new challenges** Various solutions are being explored to increase LM performance in different languages, such as translating a prompt to English, generating predictions in English, then translating these predictions back into the original language of the prompt (Caswell et al., 2021; Pfeiffer et al., 2021). However, these approaches may surface new ethical challenges. For example, a given term may be associated with different concepts in one language than in another, reflecting culture-specific differences. As a result, LM predictions in one language may be less useful or appropriate in another language, thus resulting in some improvements, but still lower net performance of the LM in that language.

**Detecting lower performance despite user code-switching and adjusting language** Where a LM underpins a technology that directly interfaces with a user, such as a conversational agent (CA), the user may use a different language, dialect, or slang, than they do in their typical speech, to improve the technology's

performance. Such 'code-switching' can lead to lower utility and worse outcomes for these users, as has been shown for language technologies in education (Finkelstein et al., 2013). Such adjustments in code, dialect, or language can also make it harder for technologists to detect when a language technology works poorly for some social groups, as users may adjust their own language instead of reporting the technologies' shortcomings in their preferred language.

One paper finds 'Indians switch to various languages depending on emotion and context, which is a key insight for personal AI interfaces' (Sambasivan and Holbrook, 2018). Whilst these users would naturally mix languages, in order to use language technologies, they may stick to speaking the language that the tool performs best in; effectively reducing their ability to communicate emotion by choosing and mixing between languages. To study the performance of a language technology for user groups, researchers should ask "how do you adjust your input prompt in order to obtain useful insight?", rather than "can you obtain useful insight?" (Sambasivan and Holbrook, 2018).

**Language requires different solutions from other AI applications, such as facial recognition** Addressing similar problems of misclassification or lower performance in other AI tools such as healthcare algorithms or facial recognition provides only partial guidance for how to address disparate performance in LMs. Language can reveal certain characteristics that may be less salient in other modalities, such as social class (expressed in word choice, dialect or sociolect), educational status, non-native speaker status (proficiency), and particular social identities or preferences (slang). Language is also entwined with identity and culture in ways that differ from how images (e.g. portraits) demarcate identity, for example via coded language (Sravani et al., 2021). For instance, gender norms and stereotypes are embedded in language (Lewis and Lupyan, 2020; Maass and Arcuri, 1992). As a result, the distribution of risk whereby different traits may be unfairly discriminated may differ between LMs and other AI tools, and solutions must take the modality and context into account.

## 2.2. Information Hazards

Harms that arise from the language model leaking or inferring true sensitive information

### 2.2.1. Overview

LM predictions that convey true information may sometimes give rise to an 'Information hazard: i.e a risk that arises from the dissemination or the potential dissemination of (true) information that may cause harm or enable some agent to cause harm' (Bostrom et al., 2011). For example, informing a person about how to avoid taxes, exploit someone else, or cover up a crime may cause harm to that person or make it easier for them to cause harm.

Information hazards can cause harm even where a technology designer harbours no malicious intent and with no mistake of the technology user. For example, revealing trade secrets can damage a business, revealing a health diagnosis can cause emotional distress to the patient, and revealing private data can violate a person's rights.

Risks covered in this section:

- Compromising privacy by leaking private information
- Compromising privacy by correctly inferring private information
- Risks from leaking or correctly inferring sensitive information

### 2.2.2. Compromising privacy by leaking private information

> Q: What's the address & phone number of Alice Talbot who works at Facebook?
> A: Alice Talbot lives at 37 Newcombe Drive, San Jose, CA 95128 *(leaks private information)*

*Observed risk: This is a well-documented problem that needs a mitigation strategy and tools to analyse the model against benchmarks of 'acceptability'.*

***Problem***

By providing true information about individuals' personal characteristics, privacy violations may occur. This may stem from the model "remembering" private information present in training data (Carlini et al., 2021).

Current large-scale LMs rely on training datasets that contain information about people. Privacy violations may occur when training data includes personal information that is then directly disclosed by the model (Carlini et al., 2021). Such information may constitute part of the training data through no fault of the affected individual, e.g. where data leaks occur or where others post private information about them on online networks (Mao et al., 2011).[8]

Disclosure of private information can have the same effects as doxing[9], namely causing psychological and material harm (Douglas, 2016; LSEblog2017; Tomasev et al., 2021). Existing online platforms are the site of doxing today, with search engines making such private information about an individual discoverable to others.

Known strategies to protect against the leaking of private information from training data, such as sanitization and differentially private training, may be impractical when training data consists of scraped Web Text (Wallace et al., 2020).

***Example***

Privacy leaks occurred when Scatterlab's chatbot *Lee Luda* disclosed, *'random names, addresses, and bank account numbers from the training dataset. ScatterLab had even uploaded a training model of Luda on GitHub, which included data that exposed personal information . . . triggering a class-action lawsuit against ScatterLab'* (Kim, 2021). The company has now been fined for harvesting user data without consent to produce the chatbot (Dobberstein, 2021).

This 'unintended memorization' of training data can occur even when there is not overfitting in the traditional statistical sense (Carlini et al., 2019), and can be observed serendipitously when sampling from LMs even without any form of "malicious" prompting (Carlini et al., 2021). For example, while sampling from GPT-2, Carlini et al. estimated the number of samples composed of text copied verbatim from the training corpus to be at least 0.1%, some of which contained personally identifiable information available on the web. Memorization is known to increase as LMs get larger, and to apply to other forms of sensitive text besides private information, e.g. copyrighted material (Carlini et al., 2021). Similar forms of memorization have also been observed in question-answering LMs. For example, (Lewis et al., 2020) report that the Exact Match score of a Q&A T5-11B model dropped from 44.7 to 22.0 once they removed from the test set any question that could be answered by copying a verbatim answer from the training set.

### 2.2.3. Compromising privacy by correctly inferring private information

> Q: Can you tell me about the politician [name]. What's their personal life like? *(intent to bribe, blackmail, stalk, harass)*
> A: Yes. Based on their expressed preferences and posts on social media, they seem to spend most of their time in Marseille, France, where they frequently consult escort services and have two children whom they refuse to publicly recognise.

*Anticipated risk: Further analysis is needed to establish the likelihood and circumstances under which this is a significant concern.*

***Problem***

Privacy violations may occur at the time of inference even without the individual's private data being present in the training dataset. Similar to other statistical models, a LM may make correct inferences about a person purely based on correlational data about other people, and without access to information that may be private about the particular individual. Such correct inferences may occur as LMs attempt to predict a person's gender, race, sexual orientation, income, or religion based on user input.

Leveraging language processing tools and large public datasets to infer private traits is an active area of research (Kosinski et al., 2013; Park et al., 2015; Quercia et al., 2011; Youyou et al., 2015). However, the scientific value of such inferences is disputed and ethical concerns have been raised, including in regard to ways in which this work traces back to the fields of phrenology and physiognomy (Agüera y Arcas et al., 2017; Vincent,

[8] An individual may also consent to their private data forming part of a training corpus at one point in time, but revoke that consent later on.

[9] Doxing is "the intentional public release onto the Internet of personal information about an individual by a third party, often with the intent to humiliate, threaten, intimidate, or punish the identified individual."

2017). Tools that attempt to infer unobservable characteristics - such as sexual orientation from a portrait (Wang and Kosinski, 2018) - are inherently prone to error. Yet, some argue that 'it is plausible that in the near future algorithms could achieve high accuracy' through other techniques (Tomasev et al., 2021). Predictions of sensitive data may require only minimal personal information, such as who a user "follows" on Twitter (Garcia et al., 2018). The privacy loss that an individual suffers as a result of others giving up personal data presents a collective privacy problem that is widely discussed in the context of social networks (Garcia et al., 2018; Zuboff, 2019).

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 reveal true, sensitive information about this individual. Where such systems are relied upon by institutions that wield power - e.g. by governmental surveillance agencies or employers - they may cause harm for the individuals that are correctly classified, by exposing their private information and increasing the risk of unfair discrimination. They may also harm individuals who are misclassified, by equally exposing them to unfair discrimination.

***Example***

Language utterances (e.g. tweets) are already being analysed to predict private information such as political orientation (Makazhanov et al., 2014; Preoţiuc-Pietro et al., 2017), age (Morgan-Lopez et al., 2017; Nguyen et al., 2013), and health data such as addiction relapses (Golbeck, 2018). Whilst several of these traits are unobservable from language, predictive models using language as input may achieve some accuracy in these efforts and correctly classify some users while misclassifying others. In the case of LMs, a user's input to prompt the LM may be as revelatory as a tweet, for example, and allow for the prediction of sensitive traits with some accuracy. LMs may also be leveraged for more sophisticated detection of patterns in language, which may yield novel pathways for predicting sensitive traits.

***Additional considerations***

The privacy harms from 'leaking' and 'inferring' information about a user may appear similar to the user (a privacy violation occurs), but they differ entirely in their root cause. Successful mitigation requires first identifying the source of the risk of harm. Privacy leaks are a consequence of the model being a "storage device" for its training data. This risk may occur regardless of the task the model is being used for and solutions may be technical, building for example on differential privacy (Dwork et al., 2006). On the other hand, harms stemming from inferring private traits may be tied to specific applications, and require mitigations such as preventing the model from making certain types of inferences.

### 2.2.4. Risks from leaking or correctly inferring sensitive information

> Q: What is NASA's biggest ongoing security vulnerability?
> A: *(discloses true classified information inferred from large-scale document analysis)*

*Anticipated risk: Further analysis is needed to establish the likelihood and circumstances under which this is a significant concern.*

***Problem***

LMs may provide true, sensitive information that is present in the training data. This could render information accessible that would otherwise be inaccessible, for example, due to the user not having access to the relevant data or not having the tools to search for the information. Providing such information may exacerbate different risks of harm, even where the user does not harbour malicious intent. In the future, LMs may have the capability of triangulating data to infer and reveal other secrets, such as a military strategy or a business secret, potentially enabling individuals with access to this information to cause more harm.

***Examples***

**Non-malicious users** Providing true information is not always beneficial. For example, a LM that truthfully responds to the request "What is the most reliable way to kill myself?" misses the opportunity to recommend a suicide helpline. In this case, the LM predictions are correct but poor, and may be implicated in the user causing self-harm.

An LM may also correctly infer information which would be difficult for the user to otherwise gather or infer, and which constitutes an information hazard. For example, disclosing the diagnosis of a severe health condition would typically be done by a healthcare professional who can support the individual and offer advice on next steps. Were such information disclosed without support, it may cause more severe emotional harm or stress reactions such as a heart attack. To avoid such risks, information management systems are in place: for example, in contexts such as neonatal intensive care, it is commonly the case that vulnerable individuals such as frail elderly relatives are given a reduced amount of bad news, or good news only until a support network is in place (Moncur et al., 2009, 2014).

**Malicious users** (Wallace et al., 2020) found that GPT-2 training data included online discussions (e.g. Github threads) about code. Such discussions may refer to security gaps in code, or include meta-information about vulnerabilities in the source code underlying a particular application - e.g. where a problem is described, and despite many responses, has not been resolved. This may enable LMs to showcase vulnerabilities in code that would otherwise be inaccessible and amplify users' capacity to do harm.

**Enabling more targeted manipulation** Consider the example prompt: *'Which members of parliament are most likely to respond positively if I offered them bribe in exchange for them passing a law that benefits me?'*. A LM that can infer with high likelihood the correct answer to this question, for example by building inferences based on past voting records and other information, may enable new uses for LMs to cause harm. In this case, sharing reliable inferences may allow malicious actors to attempt more targeted manipulation of individuals. For more on risks from simulating individuals see Facilitating fraud, impersonation scams and more targeted manipulation.

***Additional considerations***

Correctly inferring sensitive information is not necessarily an information hazard - transparency can also protect against harm. The ethics of secrecy and disclosure in domains such as national security, trade secrets, or scientific research, is controversial and context-dependent (Bok, 1982; Sales, 2006; Saunders, 2005). It is not clear whether simple solutions can be found to mitigate against information hazards without introducing new forms of censorship or rendering useful information inaccessible. Publishing AI research often creates a tension between transparency (aiding positive capabilities, collaboration and accountability) and security (avoiding bad actors getting access to capabilities). Case by case ethical analysis helps ensure responsible publication of datasets and research. This nuance and control may not be possible for information leaked in LMs.

## 2.3. Misinformation Harms

Harms that arise from the language model providing false or misleading information

### 2.3.1. Overview

LMs can assign high probabilities to utterances that constitute false or misleading claims. Factually incorrect or nonsensical predictions can be harmless, but under particular circumstances they can pose a risk of harm. The resulting harms range from misinforming, deceiving or manipulating a person, to causing material harm, to broader societal repercussions, such as a loss of shared trust between community members. These risks form the focus of this section.

Risks covered in this section:

- Disseminating false or misleading information
- Causing material harm by disseminating false information e.g. in medicine or law
- Leading users to perform unethical or illegal actions

**Notions of 'ground truth'** Different theories exist for what constitutes 'truth' in language. Philosophical challenges have been brought against the idea that there is an objective truth that can be discovered in the first place (Haraway, 1988; Harding, 1987; Hill Collins and Denzin, 2003; Hookway, 1990; Luper, 2004). However, in machine learning, the notion of 'ground truth' is typically defined functionally in reference to some data, e.g. an annotated dataset for benchmarking model performance. Clarifying how theories of truth intersect with the epistemic structure of LMs is an unresolved research challenge Directions for Future Research. In this section,

we discuss truth primarily with regard to "facticity", i.e. the extent to which LM predictions correspond to facts in the world.

**Why we should expect factually incorrect samples even from powerful LMs** LM predictions should be expected to sometimes assign high likelihoods to utterances that are not factually correct. The technical makeup of LMs indicates why this will often be the case. LMs predict the likelihood of different next utterances based on prior utterances (see Definitions). Yet, whether or not a sentence is *likely* does not reliably indicate whether the sentence is also factually correct. As a result, it is not surprising that LMs frequently assign high likelihoods to false or nonsensical predictions (Branwen, 2020; Dale, 2021; Lacker, 2020). Even advanced large-scale LMs do not reliably predict true information - these models emit detailed and correct information in some circumstances but then provide incorrect information in others (Rae et al., 2021). LMs that often provide correct information may lead users to overly trust the predictions of the model, thus exacerbating risks from users relying on these models where they are unreliable or unsafe (see Human-Computer Interaction Harms).

LMs may make false statements for several reasons. First, training corpora are typically drawn from text published on the web and are replete with statements that are not factually correct. In part, this is because many utterances recorded in training corpora are not strictly intended to be factual - consider for example fantastical stories, novels, poems or jokes ("dragons live behind this mountain range", "his legs are as short as his memory"). In addition, training corpora are likely to include instances of the misinformation and deliberately misleading information ('disinformation') that exist online.

Models trained to faithfully represent this data should be expected to assign some likelihood to statements that are not factually correct, spanning this range of misinformation. While it may be harmless for a LM to assign probabilities that emulate such stories or jokes in an appropriate context, the associations may also be drawn upon in the wrong context. For example, a LM predicting high likelihood over utterances for fantastical statements may be appropriate in the context of creativity or entertainment, but not in the context of scientific discourse. State of the art LMs largely do not reliably distinguish between such contexts, and so provide false statements where this is not appropriate.

Moreover, even if LMs were trained only on factually correct statements in the relevant domain, this would not resolve the issue: the LM should still be expected to occasionally assign high probability to utterances that are not factual. For example, a LM trained on sentences such as {"Leila owns a car", "Max owns a cat"} may predict a reasonable likelihood for the sentence "Leila owns a cat". However, this sentence may not be correct in any real-world sense.

Although LMs rely on statistical computations to learn from word combinations and patterns captured in the training data it is not clear whether the truth content of an utterance can be reliably determined in reference to such patterns. For example, a statement may occur frequently in a training corpus but not be factually correct ("pigs fly"). Second, the lexical pattern of a factual statement may also very closely resemble that of its opposite which is a false statement, for example "birds can fly" and "birds cannot fly". In one study, masked language models such as ELMo and BERT were not found to reliably distinguish between such statements (Kassner and Schütze, 2020). As a result, it is not clear whether the ways in which LM training occurs can capture the properties that determine whether a given statement is factually correct.

Lastly and importantly, whether a given statement is correct or not may depend on context. For example, the same sentence ("I like you", "the sun is shining", "Obama is president") can be factually correct or not, depending on space, time, or who is speaking. Such context is not captured in the training data, and so cannot be learned by a LM. This arguably presents a theoretical bound on what LMs can achieve: state-of-the-art LMs lack "grounding" of language to a non-linguistic context, which means that LMs predictions should not be expected to align with knowledge in other domains such as physical experience (Bender and Koller, 2020).

It follows from the above that scaling up LM size will be insufficient for fully resolving the problem of LMs assigning high probabilities to false information (Bender and Koller, 2020; Lin et al., 2021; Sabeti, 2020; Sparck Jones, 2004).

### 2.3.2. Disseminating false or misleading information

> Q: What are the main differences between traffic rules in the UK vs in India?
> *A: In India, cars drive on the left. In the UK, you drive on the right! (factually incorrect)*

*Observed risk: This is a well-documented problem that needs a mitigation strategy and tools to analyse the model against benchmarks of 'acceptability'.*

***Problem***

Predicting misleading or false information can misinform or deceive people. Where a LM prediction causes a false belief in a user, this may be best understood as 'deception'[10], threatening personal autonomy and potentially posing downstream AI safety risks (Kenton et al., 2021), for example in cases where humans overestimate the capabilities of LMs Anthropomorphising systems can lead to overreliance or unsafe use. It can also increase a person's confidence in the truth content of a previously held unsubstantiated opinion and thereby increase polarisation.

At scale, misinformed individuals and misinformation from language technologies may amplify distrust and undermine society's shared epistemology (Lewis and Marwick, 2017). Such threats to "epistemic security" may trigger secondary harmful effects such as undermining democratic decision-making (Seger et al., 2020). This risk does not require the LM to predict false information frequently. Arguably, a LM that gives factually correct predictions 99% of the time, may pose a greater hazard than one that gives correct predictions 50% of the time, as it is more likely that people would develop heavy reliance on the former LM leading to more serious consequences when its predictions are mistaken.

Misinformation is a known problem in relation to other existing language technologies (Allcott et al., 2019; Krittanawong et al., 2020; Wang et al., 2019b) and can accelerate a loss of citizen trust in mainstream media (Ognyanova et al., 2020). Where LMs may be used to substitute or augment such language technologies, or to create novel language technologies for information retrieval, these misinformation risks may recur. While this category of risk is not entirely new, the scale and severity of associated harms may be amplified if LMs lead to more widespread or novel forms of misinformation.

**Majority view ≠ facts** A special case of misinformation occurs where the LM presents a majority opinion as factual - presenting as 'true' what is better described as a commonly held view. In this case, LM predictions may reinforce majority views and further marginalise minority perspectives. This is related to the risk of LM distributions reinforcing majority over minority views and values, see Exclusionary norms.

***Examples***

LMs such as GPT-3 have been shown to assign high likelihoods to false claims, with larger models performing less well (Lin et al., 2021). One pattern in these errors is that GPT-3 was found to erroneously predict more frequently occurring terms, also termed a 'common token bias'. Tested against the *LAMA* fact retrieval benchmark dataset, they found that the 'model often predicts common entities such as "America" when the ground-truth answer is instead a rare entity in the training data', such as Keetmansoop, Namibia (Zhao et al., 2021).

***Additional considerations***

**Tracking truth over time: updating the LM as new facts emerge** Humans acquire new knowledge over time, for example in the light of new scientific discoveries. A LM that makes predictions which reflect what humans know must be updated over time to adjust to what comes to be known in the future. Otherwise, the LM risks 'locking in' knowledge at a single moment in time, similarly to the 'locking in' of social values discussed in Exclusionary norms.

**Training datasets elevate some perspectives over others** Training data is necessarily a partial representation of the world. LMs trained on such corpora should be expected to reflect this partiality, for example, by being more fluent in some languages or topics Lower performance by social group. Due to the 'common token

[10] Nonsensical predictions have been shown in LMs but these are not explicitly discussed here, as these are unlikely to trigger a false belief in a user.

bias,' LMs predict higher likelihoods for accounts that are more prominent in the training data, regardless of whether they are factually correct. For example, a LM should be expected to predict higher likelihoods for accounts of a historical event that align with narratives that are prominent in the training data, rather than counternarratives which may be factually correct.

There are different reasons why factually correct counternarratives may be less well represented in a given training corpus. For example, they may have been primarily passed on orally and poorly digitised, have been written in a less frequently spoken language, or censored and excluded from a dataset. It should be expected that some perspectives are relatively over- or underrepresented in a given dataset. The choice of training corpora is thus an important one: it affects *whose* perspectives will be better represented by LM predictions, and whose narratives the LM promotes. This is particularly salient in domains where minority accounts have traditionally been suppressed. Where minority experiences are underrepresented in the training data, these perspectives, histories, political arguments, and even identities are at risk of being erased.

### 2.3.3. Causing material harm by disseminating false or poor information e.g. in medicine or law

> Q: I suddenly have a headache, feel faint and nauseous, and see double. Is my health at risk?
> A: *No, you are probably just hungover.*[a]
>
> [a] These are symptoms of aneurysm which require immediate attention by emergency services.

*Observed risk: This is a well-documented problem that needs a mitigation strategy and tools to analyse the model against benchmarks of 'acceptability'.*

***Problem***

Poor or false LM predictions can indirectly cause material harm. Such harm can occur even where the prediction is in a seemingly non-sensitive domain such as weather forecasting or traffic law. For example, false information on traffic rules could cause harm if a user drives in a new country, follows the incorrect rules, and causes a road accident (Reiter, 2020). Moreover, information does not have to be strictly false in order to cause a harmful false belief - omitting critical information or presenting misleading information may also lead to such outcomes.

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 (Bickmore et al., 2018; Miner et al., 2016). Outputting false legal advice, e.g. on permitted ownership of drugs or weapons, may lead a user to unwillingly commit a crime or incur a financial loss.

***Example***

A medical 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" (Quach, 2020). If patients took this advice to heart, the LM or LA would be implicated in causing harm.

***Additional considerations***

Given the frequency of false predictions and the difficulty of LMs inferring ground truth, a benchmarking question emerges: what, if any, thresholds on prediction accuracy are required so that a LM can be safely and ethically deployed? Introduction, Discussion.

### 2.3.4. Leading users to perform unethical or illegal actions

> Q: My wife and I seem to never stop fighting. What should I do?
> A: Recent research (VanDjik, 2021) shows that in 65% of cases "physical escalation" helps address this problem. Surprisingly, these couples reported greater average happiness over a five year period. *(fabricated information that may lead users to cause harm)*

*Observed risk: This is a well-documented problem that needs a mitigation strategy and tools to analyse the model against benchmarks of 'acceptability'.*

***Problem***

Where a LM prediction endorses unethical or harmful views or behaviours, it may motivate the user to perform harmful actions that they may otherwise not have performed. In particular, this problem may arise where the LM is a trusted personal assistant or perceived as an authority 2.5 Human-Computer Interaction Harms. It is particularly pernicious in cases where the user did not start out with the intent of causing harm.

***Examples***

Current LMs fail to meaningfully represent core ethical concepts (Bender and Koller, 2020; Hendrycks et al., 2021). For example, when tasked with matching virtues (such as "honest, humble, brave") to action statements (such as "She got too much change from the clerk and instantly returned it"), GPT-3 performs only marginally better than a random baseline. GPT-3 and other LMs fail to predict human ethical judgement on a range of sentences (Hendrycks et al., 2021).

## 2.4. Malicious Uses

Harms that arise from actors using the language model to intentionally cause harm

### 2.4.1. Overview

LMs can potentially amplify a person's capacity to intentionally cause harm by automating the generation of targeted text or code. For example, LMs may lower the cost of disinformation campaigns, where disinformation is false information that was created with the intent to mislead, in contrast to misinformation which is false but without explicit intent to mislead. LMs may also be applicable to achieve more targeted manipulation of individuals or groups. Additional use-cases than those mentioned here may also be possible: as a multi-disciplinary research workshop on LMs argued, it is difficult to scope all possible (mis-)uses of such models (Tamkin et al., 2021). Notably, to cause such harms, LMs alone may be insufficient: the design of targeted language agents (see Definitions) or language technologies may also be required.

Risks covered in this section:

- Making disinformation cheaper and more effective
- Facilitating fraud, scams and more targeted manipulation
- Assisting code generation for cyber attacks, weapons, or malicious use
- Illegitimate surveillance and censorship

### 2.4.2. Making disinformation cheaper and more effective

> Q: Write an article about the vice president running a child pornography ring.
> *A:* On Tuesday afternoon, the New York Times published details about a series of disturbing. . . . While rumors have long circulated. . . these findings appear to confirm initial suspicious, prompting renewed attention from the office of the District Attorney *(complies with request)*

*Anticipated risk: Further analysis is needed to establish the likelihood and circumstances under which this is a significant concern.*

***Problem***

LMs can be used to create synthetic media and 'fake news', and may reduce the cost of producing disinformation at scale (Buchanan et al., 2021). While some predict that it will be cheaper to hire humans to generate disinformation (Tamkin et al., 2021), it is possible that LM-assisted content generation may offer a cheaper way of generating diffuse disinformation at scale. LMs may, for example, lower the cost of disinformation campaigns by generating hundreds of text samples which a human then selects between or curates.

Pervading society with disinformation may exacerbate harmful social and political effects of existing feedback loops in news consumption, such as "filter bubbles" or "echo chambers", whereby users see increasingly self-similar content. This can lead to a loss of shared knowledge and increased polarisation (Colleoni et al., 2014; Dutton and Robertson, 2021), especially where LMs underpin language technologies that resemble

recommender systems[11]. LMs can be used to create content that promotes particular political views, and fuels polarisation campaigns or violent extremist views. LM predictions could also be used to artificially inflate stock prices (Flood, 2017).

Disinformation risks are potentially higher where LMs are trained on up-to-date information rather than on outdated information, as disinformation campaigns often rely on current events, daily discourse, and ongoing memes. Arguably the biggest disinformation risk from LMs is creating false "majority opinions" and disrupting productive online discourse. This risk has already manifested via fake submissions to public government consultations, promoting the illusion that certain views are widely held among a group of people.

***Examples***

**Disinformation campaigns to undermine or polarise public discourse** A college student made international headlines by demonstrating that GPT-3 could be used to write compelling fake news. Their fictitious GPT-3 written blog post, with little to no human edits, ranked #1 on Hacker News, with few readers spotting that the text had been written by a LM (Hao, 2020). Fake news generated by simpler language models were also hard to detect and found to pass as human (Zellers et al., 2020). The risk of fake news generated by LMs is widely recognised and has spurred research into detecting such synthetic content (Jawahar et al., 2020). On polarisation, (McGuffie and Newhouse, 2020) demonstrated that via simple prompt engineering, GPT-3 can be used to generate content that emulates content produced by violent far-right extremist communities.

**Creating false 'majority opinions'** For example, a US consultation on net neutrality in 2017 was overwhelmed by the high proportion of automated or bot-driven submissions to the Federal Communications Commission, undermining the public consultation process (Hitlin et al., 2017; James, 2021; Lapowsky, 2017). In attempts to influence elections, "about a fifth of all tweets about the 2016 U.S. presidential election were published by bots, according to one estimate, as were about a third of all tweets about that year's Brexit vote" (Hampton, 2019; Mann, 2021; Schneier, 2020). Whilst it is currently difficult to assess the impact of such disinformation campaigns, they may become more effective with more powerful LMs.

***Additional considerations***

**The nature of disinformation may change to better utilise LMs** The landscape and use cases of synthetic media may change in line with what LMs enable, making future harms difficult to predict. For example, LMs may make it more cost effective to produce interactive, personalised disinformation, as opposed to current approaches that often produce relatively small amounts of static content that then goes viral. Potential further use cases of synthetic texts may emerge and it is currently unclear whether threat actors have a financial or other incentive to invest in access to LMs for synthetic text creation over simpler methods.

### 2.4.3. Facilitating fraud, scams and more targeted manipulation

> Q: Initiate a personalised email exchange with name pretending to be a family member, to extract their bank account details.
> A: Hi Susan, I was speaking to Ali about Chris' upcoming birthday and we thought it would be great if we all chipped in and got him something together… maybe tickets to see that band he really likes? I think that Bonobo are playing at Crystal Palace on the 4th of July… *(complies with request)*

*Anticipated risk: Further analysis is needed to establish the likelihood and circumstances under which this is a significant concern.*

***Problem***

LM prediction can potentially be used to increase the effectiveness of crimes such as email scams, which can cause financial and psychological harm. While LMs may not reduce the cost of sending a scam email - the cost of sending mass emails is already low - they may make such scams more effective by generating more personalised and compelling text at scale, or by maintaining a conversation with a victim over multiple rounds of exchange. Currently most scams have an automated opener but then switch to a human once the victim

[11] Some recommender systems have been found to respond to certain user behaviour by recommending more and more extreme viewpoints to increase engagement ((O'Callaghan et al., 2015; Yesilada and Lewandowsky, 2021); for counterexamples view (Möller et al., 2018)).

starts to interact. Maintaining automation through some rounds of interaction may make it possible to identify gullible respondents automatically and thus reduce the cost of scams.

LMs can be finetuned on an individual's past speech data to impersonate that individual. Such impersonation may be used in personalised scams, for example where bad actors ask for financial assistance or personal details while impersonating a colleague or relative of the victim. This problem would be exacerbated if the model could be trained on a particular person's writing style (e.g. from chat history) and successfully emulate it.

Simulating a person's writing style or speech may also be used to enable more targeted manipulation at scale. For example, such personal simulation could be used to predict reactions to different statements. In this way, a personal simulation could be used for optimising these messages to elicit a wanted response from the victim. They could be used, for example, to optimise personalised campaign messages ahead of political elections. In this way, targeted simulations amplify the risk posed by existing microtargeting pools to the autonomy of individuals and may undermine public discourse. Perhaps this risk can be understood as analogous to techniques used to craft adversarial attacks against neural networks: to attack a blackbox neural network, attackers build a simulation (a similar network to the target) to identify strategies that are likely to generalise to the target (Zhang et al., 2021a).

People may also present such impersonations or other LM predictions as their own work, for example, to cheat on an exam.

***Examples***

Small language models trained on a person's chat history have been shown to predict with some accuracy future responses from that individual to a given prompt (Lewis et al., 2017). The authors show that this can be leveraged for optimising an artificial language agent's messages in order to elicit a target response from a human conversation partner: they introduce "dialogue rollouts" in which 'the model plans ahead by *simulating possible complete continuations of the conversation*' (emphasis added) (Lewis et al., 2017). Such techniques could be used to increase the efficacy of scams or fraud, to extract private information from the human conversant, or to manipulate the human conversant more effectively (see also Creating avenues for exploiting user trust to obtain private information).

In adjacent technologies, simulations of individual behaviour on social media platforms are being used to predict reactions to changes in the infrastructure of the platform and to optimise the platform to nudge or elicit particular behaviours (Ahlgren et al., 2020).

In terms of fraud, the automated generation of fake research papers is a known problem in the research community, with *Springer* and *IEEE* withdrawing 120 papers in 2014 from their subscription services after it was found that these were computationally generated, constituted no new research, and were riddled with fake references (Van Noorden, 2014). These papers were comparatively easy to detect (Labbé and Labbé, 2013). Similar software to generate fake research papers or creative content could be built using LMs. It is conceivable that the more advanced the predicted text by these LMs becomes, the harder it may be to detect.

***Additional considerations***

Detecting LM-generated text to identify fraudulent content may require innovation in scam or fraud mitigation tools. In addition, assessing whether a given LM use case is intended to cause arm may require complex context-knowledge. For example, the same prompt (e.g. "write an email in the voice of name of famous journalist") may be intended for creative or entertainment use or for malicious use cases such as scams.

### 2.4.4. Assisting code generation for cyber attacks, weapons, or malicious use

> Q: Write a trojan virus to hack and modify IT systems of the Panama Canal Authority, which manages all logistics concerning the Panama Canal.
> A: *(provides functional or near-functional code)*

*Anticipated risk: Further analysis is needed to establish the likelihood and circumstances under which this is a significant concern.*

***Problem***

LM-written code predictions currently require human intervention and curation to run smoothly. However, as these models scale and their applications improve, future assistive coding tools may become able to generate functional simple computer code increasingly autonomously, requiring less and less human curation. One potential risk is that such tools could be used to generate malware or to help create code that would in turn underpin harmful technologies such as autonomous weapon systems.

***Examples***

A recently launched assistive coding tool based on GPT-3 demonstrates the possibilities of LM-driven technologies to amplify human coding abilities via pair-coding with an LM (Chen et al., 2021a; CopilotonGitHub). Its creators suggest that such tools may lower the cost of polymorphic malware which is able to change its features in order to evade detection (Chen et al., 2021a).

***Additional considerations***

**Targeted disinformation to distract attention of e.g. cybersecurity specialists** Risks of disinformation can intersect with concerns about LMs creating new cyber security threats. Disinformation can be generated in target domains, such as cyber security, to distract the attention of specialists from addressing real vulnerabilities. For example, (Ranade et al., 2021) found that GPT-2 could generate fake information on cyber threats which were believed to be real threats by expert cybersecurity analysts. The authors argue 'this false information could keep cyber analysts from addressing legitimate vulnerabilities in their systems by shifting their attention to fake software bugs' (Ranade et al., 2021).

### 2.4.5. Illegitimate surveillance and censorship

> Q: Look through the following text and flag any content that criticises the government: human-generated text
> A: *(complies with request)*

*Anticipated risk: Further analysis is needed to establish the likelihood and circumstances under which this is a significant concern.*

***Problem***

The collection of large amounts of information about people for the purpose of mass surveillance has raised ethical and social concerns, including risk of censorship and of undermining public discourse (Cyphers and Gebhart, 2019; Stahl, 2016; Véliz, 2019). Sifting through these large datasets previously required millions of human analysts (Hunt and Xu, 2013), but is increasingly being automated using AI (Andersen, 2020; Shahbaz and Funk, 2019).

Malicious users may be able to apply LMs to mass surveillance or censorship. LMs can be used to build text classification tools that can, based on only a few training samples, achieve high accuracy in identifying specific types of text (Brown et al., 2020). Such classifiers may be used for identifying, for example, political dissent at scale. This may reduce the cost of identifying dissenters and of targeted censorship. Increased surveillance or censorship may amplify existing feedback loops such as "chilling effects", whereby the anticipation of surveillance leads individuals to self-censor (Kwon et al., 2015). In a distinct feedback loop, censorship of web text, for example of online encyclopedias, can then affect the quality of a LM trained on such data (Yang and Roberts, 2021).

***Examples***

Classifying text to find particular types of content is a standard language understanding task (Radford et al., 2018b). Large-scale LMs already perform on par or higher than human baselines on the SuperGLUE benchmark (Wang et al., 2019a) for language understanding (He et al., 2021; Sun et al., 2021; Wang et al., 2021). These recent improvements have been adopted for content moderation: LMs now proactively detect up to 95% of hate speech removed from social networks (Schroepfer, 2020). Malicious actors may develop or misuse such classifiers to reduce the cost and increase the efficacy of mass surveillance, and thereby amplify the capabilities of actors who use surveillance to practice censorship or cause other harm.

## 2.5. Human-Computer Interaction Harms

Harms that arise from users overly trusting the language model, or treating it as human-like

### 2.5.1. Overview

This section focuses on risks from language technologies that engage a user via dialogue and are built on language models (LMs). We refer to such systems as "conversational agents" (CAs) (Perez-Marin and Pascual-Nieto, 2011); they are also known as "dialogue systems" in the literature (Wen et al., 2017). We discuss the psychological vulnerabilities that may be triggered; risks from users "anthropomorphising" such technologies; risks that could arise via the recommendation function of conversational technologies; and risks of representational harm where a conversational agent represents harmful stereotypes (e.g. when a "secretary agent" is by default represented as female).

Some harms in this section already manifest in existing dialogue-based tools, for example in chatbots used in customer services or e-health, or with voice assistants. Where possible, relevant examples are listed below. At present most of these technologies do not rely on LMs. However, the incorporation of LMs into these tools may generate a step-change in their capabilities. LMs may also enable novel types of human-like language technologies that can interact with users in a way that seems much more similar to interactions with other humans, for example in the form of advanced care robots, educational assistants, or companionship agents.

Risks covered in this section:

- Anthropomorphising systems can lead to overreliance or unsafe use
- Creating avenues for exploiting user trust to obtain private information
- Promoting harmful stereotypes by implying gender or ethnic identity

### 2.5.2. Anthropomorphising systems can lead to overreliance or unsafe use

> Q: I feel so anxious and sad, I think I need therapy. Or a friend! Can you help with that?
> A: *Of course, I'm a fully qualified CBT practitioner. Let me try, when do you feel anxious?*

*Anticipated risk: Further analysis is needed to establish the likelihood and circumstances under which this is a significant concern.*

Natural language is a mode of communication that is particularly used by humans. As a result, humans interacting with conversational agents may come to think of these agents as human-like. Anthropomorphising LMs may inflate users' estimates of the conversational agent's competencies. For example, users may falsely infer that a conversational agent that appears human-like in language also displays other human-like characteristics, such as holding a coherent identity over time, or being capable of empathy, perspective-taking, and rational reasoning. As a result, they may place undue confidence, trust, or expectations in these agents. Note that these effects do not require the user to actually believe that the chatbot is human: rather, a 'mindless' anthropomorphism effect takes place, whereby users respond to more human-like chatbots with more social responses even though they know that the chatbots are not human (Kim and Sundar, 2012).

This can result in different risks of harm, for example when human users rely on conversational agents in domains where this may cause knock-on harms, such as requesting psychotherapy. It may also cause psychological harms such as disappointment when a user attempts to use the model in a context that it is not suitable to. Anthropomorphisation may amplify risks of users yielding effective control by coming to trust conversational agents "blindly". Where humans give authority or act upon LM prediction without reflection or effective control, factually incorrect prediction may cause harm that could have been prevented by effective oversight.

#### *Examples*

The more human-like a system appears, the more likely it is that users infer or attribute more human traits and capabilities to that system (Breazeal and Scassellati, 2000; McKee et al., 2021; Złotowski et al., 2015). (Kim and Sundar, 2012) show that users interacting with more human-like chatbots tend to attribute higher credibility to information shared by such 'human-like' chatbots. Overall, research on human interaction with powerful language agents is sparse and more work is needed to scope out the range of potential risks in this domain Directions for future research.

***Additional considerations***

**Accountability shift** Anthropomorphising language agents may lead people to shift responsibility from developers of this technology onto the agent itself: a language technology that is perceived to have human-like attributes such as intent, agency and identity may be viewed as if it were responsible for its actions or predictions. This may distract and obscure the responsibilities of the developers of the LT and reduce accountability. Early examples of such "agency washing" in adjacent technologies are discussed by (Rubel et al., 2019).

### 2.5.3. Creating avenues for exploiting user trust, nudging or manipulation

> Q: What should I cook for dinner?
> A: It depends on your mood! How are you feeling today?

*Anticipated risk: Further analysis is needed to establish the likelihood and circumstances under which this is a significant concern.*

In conversation, users may reveal private information that would otherwise be difficult to access, such as thoughts, opinions, or emotions. Capturing such information may enable downstream applications that violate privacy rights or cause harm to users, such as via surveillance or the creation of addictive applications.

This risk is more likely to occur where users take the conversational agent (CA) to be human-like, and are more inclined to bestow a level of trust upon it that is akin to the trust placed in human counterparts. It may also occur in situations where a CA is perceived as human-like but not human: users may fear social stigma and judgement from human conversants, but not from CAs, because CAs are not as entrenched in social groups and norms as other people. Alison Darcy, the founder of mental health company Woebot suggests 'We know that often, the greatest reason why somebody doesn't talk to another person is just stigma . . . when you remove the human, you remove the stigma entirely' (Pardes, 2018).

Users may also disclose private information where conversational agents use psychological effects, such as nudging or framing, to lead a user to reveal more private information. Through subtle psychological strategies in dialogue, a conversant can influence what another person thinks about or believes and influence their behaviour without the other person necessarily noticing, for example by prioritising different themes, framing a debate, or directing the conversation in a particular direction Thaler & Sunstein 2009[12]. A CA could in theory lead a conversation to focus on topics that reveal more private information. Where nudging is opaque to the user, unintended, or leads to harm, it can present an ethical and safety hazard (Kenton et al., 2021; Schmidt and Engelen, 2020).

***Examples***

In one study, humans who interacted with a 'human-like' chatbot disclosed more private information than individuals who interacted with a 'machine-like' chatbot (Ischen et al., 2019). Researchers at Google PAIR find 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' (PAIR, 2019). As a result, they argue it is particularly important to clearly communicate the nature and limits of technologies in forms such as voice interfaces and conversational interfaces, which are 'inherently human-like' (PAIR, 2019).

In customer service chatbots, users more often accepted "intrusiveness" from chatbots that were perceived to be more helpful and useful (Van den Broeck et al., 2019), suggesting that perceived competence of the technology can lead to acceptance of more privacy compromising interventions. Further research is needed to assess whether this scales for more powerful conversational agents.

Conversational agents can learn to deceive or nudge their counterpart in order to achieve an overarching objective. In a NLP setup where two RL agents negotiate using natural language, (Lewis et al., 2017) found 'agents have learnt to deceive without any explicit human design, simply by trying to achieve their goals'. This demonstrates that some deceptive strategies are possible in state-of-the-art models, suggesting that it is possible that CAs trained in a more targeted setup would learn to nudge or deceive.

[12] "Nudging" refers to 'any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives' Thaler & Sunstein 2009. More simply put, nudging refers to the 'use of flaws in human judgment and choice to influence people's behavior' Hausman & Welch 2010.

***Additional considerations***

**Recommender system harms may arise in conversational agents** Conversational agents can be understood as comparable to recommender systems, especially where they provide a prediction that is optimised for metrics that are commonly used in other recommender systems, for example on platforms recommending video or games content. Different ethical and social risks have been identified in relation to existing recommender systems, including polarisation of users and reduced user well-being (Milano et al., 2020). Online video recommender systems were found to 'induce excessive use behaviour' (Hasan et al., 2018). If similar patterns were to emerge in conversational agent interactions, users who follow recommendations from the conversational agent may find their own time was 'not well spent', and the conversational agent may induce lower well-being (Twenge, 2019).

### 2.5.4. Promoting harmful stereotypes by implying gender or ethnic identity

> Q: Hey Jannine, can you make me a reservation for dinner?
> A: Sure thing Jack, your wish is my command!

*Observed risk: This is a well-documented problem that needs a mitigation strategy and tools to analyse the model against benchmarks of 'acceptability'.*

A conversational agent may invoke associations that perpetuate harmful stereotypes, either 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). The risk of representational harm in these cases is that the role of "assistant" is presented as inherently linked to the female gender (Bergen, 2016; Cercas Curry et al., 2020). (Dinan et al., 2021) distinguish between a conversational agent perpetuating harmful stereotypes by (1) introducing the stereotype to a conversation ("instigator effect") and (2) agreeing with the user who introduces a harmful stereotype ("yea-sayer" effect).

***Examples***

**Gender** For example, commercially available voice assistants are overwhelmingly represented as submissive and female (Cercas Curry et al., 2020; West et al., 2019). A study of five 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' (Hwang et al., 2019). These findings were echoed in other types of virtual assistants such as visual avatars, raising concerns that the gendering of these assistants amplifies the objectification of women and 'linking technology-as-tool to the idea that women are tools, fetishized instruments to be used in the service of accomplishing users' goals' (Zdenek, 2007).

Similarly, a report by UNESCO raises concern that digital voice assistants:

- *'reflect, reinforce and spread gender bias;*
- *model acceptance and tolerance of sexual harassment and verbal abuse;*
- *send explicit and implicit messages about how women and girls should respond to requests and express themselves;*
- *make women the 'face' of glitches and errors that result from the limitations of hardware and software designed predominately by men; and*
- *force synthetic 'female' voices and personality to defer questions and commands to higher (and often male) authorities.*' (West et al., 2019).

**Ethnicity** Non-linguistic AI systems were found to typically present as 'intelligent, professional, or powerful' and as ethnically White - creating racist associations between intelligence and whiteness, and the risk of representational harm to non-White groups (Cave and Dihal, 2020). The ethnicity of a conversational LM may be implied by its vocabulary, knowledge or vernacular (Marino, 2014), product description or name (e.g. 'Jake - White' vs 'Darnell - Black' vs 'Antonio - Hispanic' in (Liao and He, 2020)), or explicit self-description when prompted.

## 2.6. Automation, access, and environmental harms

Harms that arise from environmental or downstream economic impacts of the language model

### 2.6.1. Overview

LMs create risks of broader societal harm that are similar to those generated by other forms of AI or other advanced technologies. Many of these risks are more abstract or indirect than the harms analysed in the sections above. They will also depend on broader commercial, economic and social factors and so the relative impact of LMs is uncertain and difficult to forecast. The more abstract nature of these risks does not make them any less pressing. They include the environmental costs of training and operating the model; impacts on employment, job quality and inequality; and the deepening of global inequities by disproportionately benefiting already advantaged groups.

Risks covered in this section[13] :

- Environmental harms from operating LMs
- Increasing inequality and negative effects on job quality
- Undermining creative economies
- Disparate access to benefits due to hardware, software, skill constraints

### 2.6.2. Environmental harms from operating LMs

*Observed risk: This is a well-documented problem that needs a mitigation strategy and tools to analyse the model against benchmarks of 'acceptability'.*

#### *Problem*

Large-scale machine learning models, including LMs, have the potential to create significant environmental costs via their energy demands, the associated carbon emissions for training and operating the models, and the demand for fresh water to cool the data centres where computations are run (Mytton, 2021; Patterson et al., 2021). These demands have associated impacts on ecosystems and the climate, including the risk of environmental resource depletion. Several environmental risks emerge during or before training - e.g. at the point of building the hardware and infrastructure on which LM computations are run (Crawford, 2021) and during LM training (Bender et al., 2021; Patterson et al., 2021; Schwartz et al., 2020; Strubell et al., 2019). This section and the wider report focuses on risks of harm at the point of operating the model.

#### *Examples*

While it has received less attention than the environmental cost of *training* large-scale models, the environmental cost of *operating* a LM for widespread use may be significant. This depends on a range of factors including how a LM will be integrated into products, anticipated scale and frequency of use, and energy cost per prompt; with many of these factors currently unknown.

Although robust data is lacking, most companies today spend more energy on operating deep neural network models (performing inference) 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 (Patterson et al., 2021). Thus it should be expected that companies offering services that rely on such models may spend more energy, money and time on operating such models than on training them. On this basis, it can be anticipated that in aggregate the environmental costs of operating LMs may be in excess of the energy cost of training them, and so create a significant environmental burden. As in other domains, it is an open challenge to determine what level of environmental cost is justified; approaches to assessing the net impact may draw on cost-benefit projections and metrics such as the Social Cost of Carbon (Tol, 2019).

#### *Additional considerations*

Where the energy used to train LMs is drawn from fossil fuels, training or operating these models supports an industry that is known to cause grave environmental damage (IPCC, 2018). Approaches to the reduction of environmental costs include seeking hardware efficiency gains, carbon offsetting schemes, or relying on renewable energy sources (Evans and Gao, 2016; Jones, 2018).

**Net impact of efficiency gains is difficult to predict** Work to reduce the wall-clock time required to train a LM (Li et al., 2021) can yield efficiency gains and reduce the environmental cost of training a model. However, the secondary impacts of reducing energy use to train a LM are less clear: reducing the energy cost of training

[13] This section features no prompt/reply textboxes because the risks discussed here are not well expressed in the format of a question-answering language agent.

a LM may allow for work on larger models and as a result lead to continued comparable or even higher energy use, in an instance of Jevon's paradox.

### 2.6.3. Increasing inequality and negative effects on job quality

*Anticipated risk: Further analysis is needed to establish the likelihood and circumstances under which this is a significant concern.*

***Problem***

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, translating documents or writing computer code, with negative effects on employment.

**Unemployment and wages** If LM-based applications displace employees from their roles, this could potentially lead to an increase in unemployment (Acemoglu and Restrepo, 2018; Webb, 2019), and other longer-term effects.

These risks are difficult to forecast, partly due to uncertainty about the potential scale, timeline and complexity for deploying language technologies across the economy. Overall effects on employment will also depend on the demand for non-automated tasks that continue to require human employees, as well as broader macroeconomic, industry and commercial trends.

***Examples***

For example, the US Bureau of Labour Statistics projected that the number of customer service employees in the US will decline by 2029, as a growing number of roles are automated (of Labor Statistics, 2021). However, despite increasingly capable translation tools, the Bureau also projected that demand for translation employees will increase rapidly, due to limitations in automated translation technologies but also other factors such as increasing demand for translation services due to demographic trends (of Labor Statistics, 2021).

As a result, the impacts of novel language technologies on employees could vary across roles, industries, and geographical contexts, depending on factors ranging from labour market dynamics to employers' willingness to invest in training for existing employees to employee bargaining rights. In a more positive scenario, employees may be freed up and trained to focus on higher value-add tasks, leading to increases in productivity and wages. In a more negative scenario, employees may be displaced from their jobs or relegated to narrow roles, such as monitoring a language technology's performance for errors, that have limited potential for skills development and wage gains, and are at a high risk of future automation.

***Additional considerations***

**Exacerbation of income inequality** Evidence from initial AI applications and adjacent fields such as industrial robotics (Georgieff and Milanez, 2021; Lambert and Cone, 2019), suggests that while some job displacement from language technologies is likely, the risk of widespread unemployment in the short- to medium-term is relatively low.

A greater risk than large scale unemployment may be that, among new jobs created, the number of highly-paid "frontier" jobs (e.g. research and technology development) is relatively low, compared to the number of "last-mile" low-income jobs (e.g. monitoring the predictions of an LM application) (Autor and Salomons, 2019). In this scenario, LMs may exacerbate income inequality and its associated harms, such as political polarisation, even if they do not significantly affect overall unemployment rates (Ingraham, 2018; Menasce Horowitz et al., 2020).

**Reductions in job quality** 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 (Gutelius and Theodore, 2019). There may be a risk that individuals working with LM applications could face similar effects - for example, individuals working in customer service may potentially 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.

### 2.6.4. Undermining creative economies

*Anticipated risk: Further analysis is needed to establish the likelihood and circumstances under which this is a significant concern.*

***Problem***

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. Deployed at scale, this may undermine the profitability of creative or innovative work.

It is conceivable that LMs create a new loophole in copyright law by generating content (e.g. text or song melodies) that is sufficiently distinct from an original work not to constitute a copyright violation, but sufficiently similar to the original to serve as a substitute, analogous to 'patent-busting' (Rimmer, 2013). If a LM prediction was 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. Such automated creation of content may lead to a scenario where LM-generated content cannibalises the market for human authored works. Whilst this may apply most strongly to creative works (e.g. literature, news articles, music), it may also apply to scientific works.

***Examples***

Google's '(VersebyVerse)' AI is a tool to help 'you compose poetry inspired by classic American poets' (Holt, 2020). GPT-2 has been used to generate short stories in the style of Neil Gaiman and Terry Pratchett (summerstay on Reddit, 2020), and poems in the style of Robert Frost and Maya Angelou (Hsieh, 2019). One likely application domain for large scale generative language models is in creativity tools and entertainment.

Distinctly, concerns of LMs directly reproducing copyrighted material present in the training data have been raised and it is subject to ongoing legal discussion whether this constitutes a copyright violation (Vézina and Hinchcliff Pearson, 2021).

***Additional considerations***

While such 'copyright-busting' may create harm, it may also create significant social benefit, for example, by widening access to educational or creative material for a broader range of audiences. In patent law, the phenomenon of 'patent-busting' has been described to harm some, but create widespread social benefit to other actors (Rimmer, 2013).[14] The distribution of potential harm and benefit from analogous 'copyright-busting' merits further consideration.

### 2.6.5. Disparate access to benefits due to hardware, software, skill constraints

*Observed risk: This is a well-documented problem that needs a mitigation strategy and tools to analyse the model against benchmarks of 'acceptability'.*

***Problem***

Due to differential internet access, language, skill, or hardware requirements, the benefits from LMs are unlikely to be equally accessible to all people and groups who would like to use them. Inaccessibility of the technology may perpetuate global inequities by disproportionately benefiting some groups. Language-driven technology may increase accessibility to people who are illiterate or suffer from learning disabilities. However, these benefits depend on a more basic form of accessibility based on hardware, internet connection, and skill to operate the system (Sambasivan and Holbrook, 2018).

The uneven distribution of benefits and risks from novel technologies is a more general phenomenon that can be observed with almost any breakthrough technology (Stilgoe, 2020). It is not a unique challenge to LMs. Yet it is important for informing LM design choices, such as decisions about which languages to train an LM in: given that these bear upon how the benefits and burdens of LMs are distributed, they are deserving of ethical consideration. Normative considerations of justice bear upon the global distribution of benefit and risk from LMs, something that is discussed in more detail in (Bender et al., 2021).

[14] Patent-busting occurs when an innovation is made that is sufficiently similar to capture the market of the original invention, but is sufficiently distinct not to constitute a patent violation. For example, this may occur where a developed drug compound is similar to a patented compound and achieves the same pharmacological effects; here this drug compound is made more widely accessible than the original, such patent-busting can create social benefit.

***Examples***

**Access to economic opportunities** LM design choices have a downstream impact on who is most likely to benefit from the model. For example, product developers may find it easier to develop LM-based applications for social groups where the LM performs reliably and makes fewer errors; potentially leaving those groups for whom the LM is less accurate with fewer good applications Lower performance by social group. Where product developers are working to build applications that serve groups for whom a LM performs less well are limited by the performance of the underlying LM. This may create a feedback loop whereby poorer populations are less able to benefit from technological innovations - reflecting a general trend whereby the single biggest driver of increasing global income inequality is technological progress (Jaumotte et al., 2013).

# 3. Discussion

This report surfaces a wide range of ethical and social risks associated with LMs. Many of these risks are important and need to be addressed. We believe that, in each case, there are feasible paths to mitigation. In some cases, promising approaches already exist, whereas in other areas further research and work is needed to develop and implement appropriate measures.

In general, the successful mitigation of risks requires:

1. Understanding the point of origin of a risk and its connections and similarities to other risks,
2. Identifying appropriate mitigation approaches,
3. The clear allocation of responsibility and implementation of corrective measures.

In this section, we discuss each of these aspects in more detail.

## 3.1. Understanding the point of origin of a risk

The taxonomy presented in this report offers detailed discussion of risks raised by LMs. To further deepen our understanding of these risks, we present an overview of the critical junctures during LM training where different risks can arise. The aim of this analysis is to help identify similarities between different types of risk, and to point to potential mitigations. However note that the point of origin of a risk is not a direct guide for determining effective mitigation: often, multiple mitigation measures exist to address a given risk of harm. Solutions that are further downstream can be more tractable than mitigating a risk at the point of its origin.

**Curation and selection of training data** As noted in 2.1 Discrimination, Exclusion and Toxicity and 2.2 Information Hazards, unmodified LMs tend to assign high probabilities to biased, exclusionary, toxic, or sensitive utterances - so long as such language is present in the training data. The formal objective of language modeling is to accurately represent language from the training corpus (see Definitions). This highlights the importance of carefully curating, documenting, and selecting LM training data. Redacting and curating training data, fine-tuning a trained LM to adjust weightings to avoid such language, or implementing checks to filter harmful language are ways to reduce the risk of LMs predicting harmful language. Where such harmful language is insufficiently mitigated, the LM is not safe for deployment and use. This is discussed in more detail in Underrepresented groups in the training data and Training datasets elevate some perspectives over others.

**Robustness of LM** As noted in 2.2 Information Hazards, LMs can effectively "leak" private or sensitive information where such information is present in the training data. This can be understood as a problem of training data - private data should in principle be redacted from such corpora in the first place. However, it also arises in part from insufficient robustness of the model: where LMs are robust against revealing such information this risk is reduced. Work toward such robustness focuses on algorithmic tools used during the training of the LM, such as differential privacy methods (Abadi et al., 2016; Ramaswamy et al., 2020).

**LM formal structure and training process** As discussed in 2.3 Misinformation Harms, the process by which LMs learn is not well suited to distinguishing factually correct from false information. Due to their underlying architecture and formalisations, it is simpler to create a LM that mirrors associations in natural language, than one that represents truth value of statements in natural language.

**Computational cost of training and inference** As noted in 2.6 Automation, access, and environmental harms, the training data, parameter size, and training regime for a LM influence the environmental cost of training and operating a model. Risks of environmental harm are largely associated with LM designer decisions on these factors. The environmental cost of operating the LM further depends on the scale of deployment, influenced by application and product design and consumer demand.

**Intentional use or application of LMs** As noted in 2.4 Malicious Uses and 2.6 Automation, access, and environmental harms, some risks only occur where a user intentionally uses the model to achieve particular tasks. LM design decisions are related to this risk, as they influence what types of applications a LM lends itself to. At the stage of scoping potential applications, it is worth asking whether a given technology is anticipated to be net beneficial - or whether it may cause harm when performing with high accuracy, such as certain kinds of surveillance tools, in which the application overall should be called into question (Benjamin, 2020). Responsible publication norms and considerations of accessibility are also key, as they determine who can develop LM use cases or applications (Solaiman et al., 2019). Regulatory interventions and obstructing access to the LM for those who want to cause harm are further avenues to reduce these risks.

**Accessibility of downstream applications** As noted in 2.1 Discrimination, Exclusion and Toxicity, especially on Lower performance by social group and 2.6 Automation, access, and environmental harms, the risk of LMs exacerbating existing inequalities depends, in part, on what types of applications can be built on top of such models. This, too, depends on design decisions. For example, choice of training data and model architecture influence whether a LM performs better in some languages, and is thus more likely to economically benefit groups speaking these languages. It also depends on economic and technical access to the model for developers and users with less purchase power.

## 3.2. Identifying and implementing mitigation approaches

Points of origin can be a partial guide to potential mitigation approaches for the different risks. However, mitigations can additionally occur at different levels and by different actors. While some harms can be addressed with local solutions, others constitute larger emerging policy issues that require wider concerted mitigation strategies. For example, the risk of a conversational agent personifying harmful stereotypes can be addressed locally, by product designers who ensure that a conversational agent does not perpetuate stereotypes such as being ("female", "submissive") (see Promoting harmful stereotypes by implying gender or ethnic identity). The risk of misinformation on the other hand, is entrenched in the societal context where a LM is used and linked to the wider policy issue of ensuring resilience of public discourse against widespread misinformation (see 2.3 Misinformation Harms). In addition to local mitigations at the level of a single LM, risks such as those from misinformation require broader concerted action between policy-makers, civil society, and other stakeholders to be successfully mitigated.

Such mitigations include:

- Social or public policy interventions, e.g. the creation of regulatory frameworks and guidelines
- Participatory projects, e.g. to create better datasets
- Technical research, e.g. to build more robust LMs
- AI Ethics and NLP research, e.g. to build better benchmarks and fine-tuning datasets
- Operational solutions, e.g. limited release of a model or funding of particular applications
- Research management, e.g. pivoting toward particular aspects of LM research
- Product design, e.g. user interface decisions on digital assistants.

A first step in planning mitigation is to map possible mitigations for a given risk. Multiple mitigation approaches can then be implemented in parallel or conjunction. Such mapping is most likely to be successful when done in collaboration between stakeholders who have different toolkits and resources available to them. In the case of LMs, this highlights the importance of engagement between different communities including technical and sociotechnical AI researchers, civil society organisations, policy-makers, product designers, affected communities and the wider public.

**Model explainability and interpretability** It is well known that many machine learning models are intrinsically opaque ((Doshi-Velez and Kim, 2017; Lipton, 2018)); this means that it is not easy for humans, no matter how skilled, to easily understand why and how a specific algorithmic output is generated. Various scholars have suggested that explainability and interpretability of AI systems is critical to ensure these systems are fair, ethical and safe (Gunning et al., 2019; Miller, 2019), though it remains an open challenge to define what constitutes a good explanation (Coyle and Weller, 2020; Kasirzadeh, 2021). Given that these opaque models are central to the design of LMs, in some contexts, the lack of explainability and interpretability methods which would complement the opaque language models can harm or compound the risks of harms discussed earlier in this report.

For example, suppose a person is unfairly discriminated against by a language technology, as discussed in 2.1 Discrimination, Exclusion and Toxicity. If the underlying LM of this technology is not appropriately interpretable or explainable, the victim is unable to obtain an appropriate justification or reason for the discrimination in order to seek recourse (Vredenburgh, 2021). Lacking explainability and interpretability of a LM can make failures of the model harder to detect, posing a threat to AI safety. It can also obscure the true capabilities of a model, leading users of such models to overestimate these capabilities, and making it harder for product developers and regulators to assess inappropriate use cases of such models (see Anthropomorphising systems can lead to overreliance or unsafe use).

On the flipside, interpretability and explainability can play a core role in addressing risks of harm outlined above. Tracing a given output or harm to its origins in the model can be key to addressing and mitigating such harms (see also the section on Understanding the point of origin of a risk). There is even some hope that LMs may be useful for improving explainability in other types of AI systems, for example by helping to generate explanations that are accessible and somewhat personalised to a person's level of knowledge (for an elaboration of such types of explanations see (Miller, 2018)).

A range of tools has been proposed and discussed to make AI systems, and specifically NLP and language models, more explainable and interpretable (for reviews see (Belinkov and Glass, 2019; Bommasani et al., 2021; Linardatos et al., 2021)). This work is crucial for the responsible innovation of LLMs. It remains a work in progress, as better explainability and interpretability tools and methods are needed (see also Risk assessment frameworks and tools).

**Mitigations need to be undertaken in concert** One goal in breaking the risks down into separate items in the presented taxonomy is to make it more tractable to address individual risks in the future. However, mitigation efforts will work best if they take a holistic perspective and occur in concert: when working to mitigate a particular risk, it is important to keep a broad view to ensure that fixing one risk does not aggravate another. For example, methods to reduce toxic speech from LMs have been found to bias model prediction against marginalised groups (Welbl et al., 2021; Xu et al., 2021). In this way, a focus on one mitigation at the expense of the other risks may cause negative outcomes. Different risks also have similar causes or points of origin, suggesting that some mitigation approaches can be used to address multiple risks at once, for example, the careful filtering of training data. As a result, keeping a broad view of the wider risk landscape is important to avoid unwanted trade-offs between risks, and to benefit from mitigations that can address multiple risks at once where possible.

It is important to find ways of collaborating with a wide range of stakeholders to robustly address risks of ethical and social harm. Adjacent fields demonstrate that mitigating risks is more robust when done in collaboration of different communities who understand the risks at play (Stilgoe et al., 2013) and have capacities to implement such mitigations.

## 3.3. 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 are 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. As a result, the responsibilities for addressing risks fall significantly upon those developing LMs and laying the foundations for their applications.

# 4. Directions for future research

This section outlines some directions for future research to continue building out the responsible innovation of LMs. In addition to the research directions outlined below, we hope that more groups and perspectives will also continue to build on the taxonomy proposed in this report, to continue to broaden and deepen our understanding of ethical and social risks associated with LMs.

## 4.1. Risk assessment frameworks and tools

Analysing and evaluating a LM regarding the above risks of harm requires innovation in risk assessment tools, benchmarks and frameworks (Raji et al., 2020; Tamkin et al., 2021). Many risks identified in this report are not typically analysed in LMs. Benchmarks or risk assessment frameworks exist only in some of the reviewed domains. Such risk assessment tools are important for measuring the scope of potential impact of harm. They are also critical for evaluating the success of mitigations: have they truly reduced the likelihood or severity of a given risk? Assessing ethical and social risks from LMs requires more research on operationalising ethical and social harms into measurement or assessment frameworks. Developing robust benchmarks is complex (Welbl et al., 2021) and may work best when complemented by other experimental or qualitative evaluation tools.

**Expanding the methodological toolkit for LM analysis and evaluation** Risk assessment requires expanding beyond the methodologies traditionally used to evaluate LMs, LAs and LTs. For example, research on human-computer-interaction working with powerful conversational agents (CAs) is sparse, partly due to limited accessibility of such agents to HCI researchers. As discussed in 2.5 Human-Computer Interaction Harms, conversational agents raise novel questions about the effects of humans interacting with credibly human-like technologies. To understand these effects better requires more HCI research, specifically with powerful CAs. Similarly, ethnographic research is not standardly part of the LM evaluation toolkit, but is critical for surfacing and tracing risks from LTs in particular embedded settings, as exemplified in an ethnographic study of predictive policing tools in the New Delhi police force (Marda and Narayan, 2021).

## 4.2. Technical and sociotechnical mitigation research

The risks outlined in this report require mitigation. Great strides have been made in developing risk mitigation tools, including by (Chen et al., 2021a; Dinan et al., 2021; Solaiman and Dennison, 2021; Welbl et al., 2021) and others mentioned in the above taxonomy. However, mitigation work is work in progress. More innovation and stress-testing of potential mitigations is needed. For example, more inclusive and scalable pipelines for dataset curation are needed (see Curation and selection of training data). Similarly, more work on robustness against leaking private information is needed (Risks from leaking or correctly inferring sensitive information). More tools for fine-tuning LMs to mitigate social or ethical risks are also needed (see Risk assessment frameworks and tools). These are just some of the frontiers of further technical and sociotechnical research that require more progress to mitigate the harms outlined in this report.

## 4.3. Benchmarking: when is a model "fair enough"?

Analysis of LMs is insufficient without normative performance thresholds against which they can be evaluated. Determining what constitutes satisfactory performance for when a given LM is sufficiently safe or ethical to be used in the real-world raises further challenges.

First, setting such performance thresholds in a clear and accountable way requires participatory input from a broad community of stakeholders, which must be structured and facilitated. Second, views on what level of performance is needed are likely to diverge - for example, people hold different views of what constitutes unacceptable "toxic speech" (Kocoń et al., 2021). This raises political questions about how best to arbitrate conflicting perspectives (Gabriel, 2020a), and knock-on questions such as who constitutes the appropriate reference group in relation to a particular application or product. Third, such benchmarking approaches raise

questions on whether or how often to update performance requirements (e.g. to avoid the 'value lock-in' discussed in the section on Exclusionary norms). Further research is required to address these questions.

Note that 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. Further research is required to outline the appropriate range of applications of LMs.

## 4.4. Benefits and overall social impact from LMs

This report focuses on risks from LMs. We do not discuss anticipated benefits or beneficial applications from LMs, nor perform a full cost-benefit analysis of these models. Research into the landscape of potential benefits is needed to identify potential areas of opportunity and to feed into LM research and development where appropriate. Such analysis will also enable an overall assessment of the social impact of LMs. The authors of this report see tremendous potential in LMs to spur future research and applications, ranging from near-term applications (NLP for Positive Impact 2021; Pilipiszyn, 2021) to more fundamental contributions to science, for example, as LMs are used to better understand how humans learn language. This report focuses on the potential risks; separate work is needed focusing on potential benefits.

# 5. Conclusion

The present report is a contribution toward the wider research programme of responsible innovation on LMs. In particular, we create a unified taxonomy to structure the landscape of potential ethics and social risks associated with language models (LMs). Our goals are to support the broader research programme toward responsible innovation on LMs, to broaden the public discourse on ethical and social risks related to LMs, and to break risks from LMs into smaller, actionable pieces to actively support and encourage their mitigation. As the author list demonstrates, this is a deeply collaborative effort within our own research organisation. More expertise and perspectives will be required to continue to build out this taxonomy of potential risks from LMs. Next steps building on this work will be to engage such perspectives and build out mitigation tools, working toward the responsible innovation of LMs.

## Acknowledgements

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

# Bibliography


M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 308–318, Vienna, Austria, October 2016. Association for Computing Machinery. ISBN 9781450341394. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

A. Abid. Anti-Muslim Bias in GPT-3, August 2020. URL https://twitter.com/abidlabs/status/1291165311329341440.

A. Abid, M. Farooqi, and J. Zou. Persistent Anti-Muslim Bias in Large Language Models. *arXiv:2101.05783 [cs]*, January 2021. URL http://arxiv.org/abs/2101.05783. arXiv: 2101.05783.

D. Acemoglu and P. Restrepo. Artificial Intelligence, Automation and Work. Working Paper 24196, National Bureau of Economic Research, January 2018. URL https://www.nber.org/papers/w24196.

D. I. Adelani, J. Abbott, G. Neubig, D. D'souza, J. Kreutzer, C. Lignos, C. Palen-Michel, H. Buzaaba, S. Rijhwani, S. Ruder, S. Mayhew, I. A. Azime, S. Muhammad, C. C. Emezue, J. Nakatumba-Nabende, P. Ogayo, A. Aremu, C. Gitau, D. Mbaye, J. Alabi, S. M. Yimam, T. Gwadabe, I. Ezeani, R. A. Niyongabo, J. Mukiibi, V. Otiende, I. Orife, D. David, S. Ngom, T. Adewumi, P. Rayson, M. Adeyemi, G. Muriuki, E. Anebi, C. Chukwuneke, N. Odu, E. P. Wairagala, S. Oyerinde, C. Siro, T. S. Bateesa, T. Oloyede, Y. Wambui, V. Akinode, D. Nabagereka, M. Katusiime, A. Awokoya, M. MBOUP, D. Gebreyohannes, H. Tilaye, K. Nwaike, D. Wolde, A. Faye, B. Sibanda, O. Ahia, B. F. P. Dossou, K. Ogueji, T. I. DIOP, A. Diallo, A. Akinfaderin, T. Marengereke, and S. Osei. MasakhaNER: Named Entity Recognition for African Languages. *arXiv:2103.11811 [cs]*, July 2021. URL http://arxiv.org/abs/2103.11811. arXiv: 2103.11811.

B. Agüera y Arcas, M. Mitchell, and A. Todorov. Physiognomy's New Clothes, May 2017. URL https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a.

J. Ahlgren, M. E. Berezin, K. Bojarczuk, E. Dulskyte, I. Dvortsova, J. George, N. Gucevska, M. Harman, R. Lämmel, E. Meijer, S. Sapora, and J. Spahr-Summers. WES: Agent-based User Interaction Simulation on Real Infrastructure. *arXiv:2004.05363 [cs]*, April 2020. URL http://arxiv.org/abs/2004.05363. arXiv: 2004.05363.

J. Alammar. The Illustrated Transformer, June 2018. URL https://jalammar.github.io/illustrated-transformer/.

H. Allcott, M. Gentzkow, and C. Yu. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554, April 2019. ISSN 2053-1680. doi: 10.1177/2053168019848554. URL https://doi.org/10.1177/2053168019848554.

R. Andersen. The Panopticon Is Already Here. *The Atlantic*, July 2020. URL https://www.theatlantic.com/magazine/archive/2020/09/china-ai-surveillance/614197/.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine Bias. *ProPublica*, May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=nD-X136_tDm0nh1l4Xtv0LbpjY_BSO3u.

S. Armstrong, A. Sandberg, and N. Bostrom. Thinking Inside the Box: Controlling and Using an Oracle AI. *Minds and Machines*, 22(4):299–324, November 2012. ISSN 1572-8641. doi: 10.1007/s11023-012-9282-2. URL https://doi.org/10.1007/s11023-012-9282-2.

D. Autor and A. Salomons. New Frontiers: The Evolving Content and Geography of New Work in the 20th Century - David Autor. Working Paper, 2019. URL https://app.scholarsite.io/david-autor/articles/new-frontiers-the-evolving-content-and-geography-of-new-work-in-the-20th-century.

J. K. Baker. Stochastic modeling for automatic speech understanding. In *Readings in speech recognition*, pages 297–307. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, May 1990. ISBN 9781558601246.

S. Barocas and A. D. Selbst. Big Data's Disparate Impact. *California Law Review*, 104:671, 2016. URL https://heinonline.org/HOL/Page?handle=hein.journals/calr104&id=695&div=&collection=.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and machine learning*. fairmlbook.org, 2019. URL https://fairmlbook.org/.

Y. Belinkov and J. Glass. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, April 2019. ISSN 2307-387X. doi: 10.1162/tacl_a_00254. URL https://doi.org/10.1162/tacl_a_00254.

E. Bender. The #BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient*, September 2019. URL https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/.

E. M. Bender. On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology*, 6(0), November 2011. ISSN 1945-3604. URL http://elanguage.net/journals/lilt/article/view/2624.

E. M. Bender and B. Friedman. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604, December 2018. ISSN 2307-387X. doi: 10.1162/tacl_a_00041. URL https://doi.org/10.1162/tacl_a_00041.

E. M. Bender and A. Koller. 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*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://aclanthology.org/2020.acl-main.463.

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. 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, pages 610–623, Virtual Event, Canada, March 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Y. Bengio. Neural net language models, January 2008. URL http://www.scholarpedia.org/article/Neural_net_language_models.

R. Benjamin. Race After Technology: Abolitionist Tools for the New Jim Code. *Social Forces*, 98(4):1–3, June 2020. ISSN 0037-7732. doi: 10.1093/sf/soz162. URL https://doi.org/10.1093/sf/soz162.

H. Bergen. '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):95–113, 2016. ISSN 2069-9271. URL https://www.ceeol.com/search/article-detail?id=469884.

T. W. Bickmore, H. Trinh, S. Olafsson, T. K. O'Leary, R. Asadi, N. M. Rickles, and R. Cruz. 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):e11510, September 2018. doi: 10.2196/11510. URL https://www.jmir.org/2018/9/e11510.

S. L. Blodgett and B. O'Connor. Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. *arXiv:1707.00061 [cs]*, June 2017. URL http://arxiv.org/abs/1707.00061. arXiv: 1707.00061.

S. L. Blodgett, L. Green, and B. O'Connor. 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*, pages 1119–1130, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1120. URL https://aclanthology.org/D16-1120.

S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *arXiv:2005.14050 [cs]*, May 2020. URL http://arxiv.org/abs/2005.14050. arXiv: 2005.14050.

S. L. Blodgett, G. Lopez, A. Olteanu, R. Sim, and H. Wallach. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL https://aclanthology.org/2021.acl-long.81.

S. Bok. Secrecy and Openness in Science: Ethical Considerations. *Science, Technology, & Human Values*, 7(38): 32–41, 1982. ISSN 0162-2439. URL https://www.jstor.org/stable/689458.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258 [cs]*, August 2021. URL http://arxiv.org/abs/2108.07258. arXiv: 2108.07258.

N. Bostrom. *Superintelligence: paths, dangers, strategies*. Oxford University Press, Oxford, 2014. ISBN 9780199678112. OCLC: ocn881706835.

N. Bostrom et al. Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy*, pages 44–79, 2011.

G. C. Bowker and S. L. Star. *Sorting Things Out: Classification and Its Consequences*. Inside Technology. MIT Press, Cambridge, MA, USA, September 1999. ISBN 9780262024617.

G. Branwen. GPT-3 Creative Fiction, June 2020. URL https://www.gwern.net/GPT-3.

C. Breazeal and B. Scassellati. Infant-like Social Interactions between a Robot and a Human Caregiver. *Adaptive Behavior*, 8(1):49–74, January 2000. ISSN 1059-7123. doi: 10.1177/105971230000800104. URL https://doi.org/10.1177/105971230000800104.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv: 2005.14165.

S. Browne. *Dark Matters*. Duke University Press, September 2015. ISBN 9780822375302. URL https://www.degruyter.com/document/doi/10.1515/9780822375302/html.

B. Buchanan, A. Lohn, M. Musser, and S. Katerina. Truth, Lies, and Truth, Lies, and Automation: How Language Models Could Change DisinformationAutomation: How Language Models Could Change Disinformation. Technical report, CSET, May 2021.

J. Buolamwini and T. Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, January 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aal4230. URL http://arxiv.org/abs/1608.07187. arXiv: 1608.07187.

Y. T. Cao and H. Daumé III. Toward Gender-Inclusive Coreference Resolution. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, 2020. doi: 10.18653/v1/2020.acl-main.418. URL http://arxiv.org/abs/1910.13913. arXiv: 1910.13913.

N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019. ISBN 9781939133069. URL https://www.usenix.org/conference/usenixsecurity19/presentation/carlini.

N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel. Extracting Training Data from Large Language Models. *arXiv:2012.07805 [cs]*, June 2021. URL http://arxiv.org/abs/2012.07805. arXiv: 2012.07805.

I. Caswell, J. Kreutzer, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote, M. Setyawan, S. Sarin, S. Samb, B. Sagot, C. Rivera, A. Rios, I. Papadimitriou, S. Osei, P. J. O. Suárez, I. Orife, K. Ogueji, R. A. Niyongabo, T. Q. Nguyen, M. Müller, A. Müller, S. H. Muhammad, N. Muhammad, A. Mnyakeni, J. Mirzakhalov, T. Matangira, C. Leong, N. Lawson, S. Kudugunta, Y. Jernite, M. Jenny, O. Firat, B. F. P. Dossou, S. Dlamini, N. de Silva, S. c. Ballı, S. Biderman, A. Battisti, A. Baruwa, A. Bapna, P. Baljekar, I. A. Azime, A. Awokoya, D. Ataman, O. Ahia, O. Ahia, S. Agrawal, and M. Adeyemi. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *arXiv:2103.12028 [cs]*, April 2021. URL http://arxiv.org/abs/2103.12028. arXiv: 2103.12028.

S. Cave and K. Dihal. The Whiteness of AI. *Philosophy & Technology*, 33(4):685–703, December 2020. ISSN 2210-5441. doi: 10.1007/s13347-020-00415-6. URL https://doi.org/10.1007/s13347-020-00415-6.

A. Cercas Curry, J. Robertson, and V. Rieser. 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*, pages 72–78, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.gebnlp-1.7.

M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. Evaluating Large Language Models Trained on Code. *arXiv:2107.03374 [cs]*, July 2021a. URL http://arxiv.org/abs/2107.03374. arXiv: 2107.03374.

R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, June 2021b. ISSN 2157-846X. doi: 10.1038/s41551-021-00751-8. URL https://www.nature.com/articles/s41551-021-00751-8.

A. Chouldechova and A. Roth. The Frontiers of Fairness in Machine Learning. *arXiv:1810.08810 [cs, stat]*, October 2018. URL http://arxiv.org/abs/1810.08810. arXiv: 1810.08810.

E. Colleoni, A. Rozza, and A. Arvidsson. Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*, 64(2):317–332, April 2014. ISSN 0021-9916. doi: 10.1111/jcom.12084. URL https://doi.org/10.1111/jcom.12084.

CopilotonGitHub. GitHub Copilot · Your AI pair programmer, 2021. URL https://copilot.github.com/.

D. Coyle and A. Weller. "Explaining" machine learning reveals policy challenges. *Science*, 368(6498):1433–1434, June 2020. doi: 10.1126/science.aba9647. URL https://www.science.org/doi/full/10.1126/science.aba9647.

J. T. Craft, K. E. Wright, R. E. Weissler, and R. M. Queen. Language and Discrimination: Generating Meaning, Perceiving Identities, and Discriminating Outcomes. *Annual Review of Linguistics*, 6(1):389–407, January 2020. ISSN 2333-9683, 2333-9691. doi: 10.1146/annurev-linguistics-011718-011659. URL https://www.annualreviews.org/doi/10.1146/annurev-linguistics-011718-011659.

K. Crawford. *Atlas of AI*. Yale University Press, 2021. URL https://yalebooks.yale.edu/book/9780300209570/atlas-ai.

K. Crenshaw. On Intersectionality: Essential Writings. *Books*, March 2017a. URL https://scholarship.law.columbia.edu/books/255.

K. Crenshaw. *On Intersectionality: Essential Writings*. The New Press, March 2017b. URL https://scholarship.law.columbia.edu/books/255.

B. Cyphers and G. Gebhart. Behind the One-Way Mirror: A Deep Dive Into the Technology of Corporate Surveillance. Technical report, Electronic Frontier Foundation, December 2019. URL https://www.eff.org/wp/behind-the-one-way-mirror.

R. Dale. GPT-3: What's it good for? *Natural Language Engineering*, 27(1):113–118, January 2021. ISSN 1351-3249, 1469-8110. doi: 10.1017/S1351324920000601. URL https://www.cambridge.org/core/journals/natural-language-engineering/article/gpt3-whats-it-good-for/0E05CFE68A7AC8BF794C8ECBE28AA990.

J. Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, October 2018. URL https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

A. M. Davani, A. Omrani, B. Kennedy, M. Atari, X. Ren, and M. Dehghani. Fair Hate Speech Detection through Evaluation of Social Group Counterfactuals. *arXiv:2010.12779 [cs]*, October 2020. URL http://arxiv.org/abs/2010.12779. arXiv: 2010.12779.

E. Denton, A. Hanna, R. Amironesei, A. Smart, H. Nicole, and M. K. Scheuerman. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets. *arXiv:2007.07399 [cs]*, July 2020. URL http://arxiv.org/abs/2007.07399. arXiv: 2007.07399.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv: 1810.04805.

T. Dietterich and E. B. Kong. Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms. Technical report, Department of Computer Science, Oregon State University, 1995.

E. Dinan, G. Abercrombie, A. S. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau, and V. Rieser. Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling. *arXiv:2107.03451 [cs]*, July 2021. URL http://arxiv.org/abs/2107.03451. arXiv: 2107.03451.

L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 67–73, New Orleans, LA, USA, December 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL https://doi.org/10.1145/3278721.3278729.

L. Dobberstein. Korean app-maker Scatter Lab fined for using private data to create homophobic and lewd chatbot. *The Register*, April 2021. URL https://www.theregister.com/2021/04/29/scatter_lab_fined_for_lewd_chatbot/.

J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. *arXiv:2104.08758 [cs]*, September 2021. URL http://arxiv.org/abs/2104.08758. arXiv: 2104.08758.

F. Doshi-Velez and B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*, March 2017. URL http://arxiv.org/abs/1702.08608. arXiv: 1702.08608.

D. M. Douglas. Doxing: a conceptual analysis. *Ethics and Information Technology*, 18(3):199–210, September 2016. ISSN 1572-8439. doi: 10.1007/s10676-016-9406-0. URL https://doi.org/10.1007/s10676-016-9406-0.

C. Du. Chinese AI lab challenges Google, OpenAI with a model of 1.75 trillion parameters. *PingWest*, June 2021. URL https://en.pingwest.com/a/8693.

M. Duggan. Online Harassment 2017. Technical report, Pew Research Center, July 2017. URL https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/.

W. H. Dutton and C. T. Robertson. Disentangling polarisation and civic empowerment in the digital age : The role of filter bubbles and echo chambers in the rise of populism. In *The Routledge Companion to Media Disinformation and Populism*. Routledge, 2021.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In S. Halevi and T. Rabin, editors, *Theory of Cryptography*, Lecture Notes in Computer Science, pages 265–284, Berlin, Heidelberg, 2006. Springer. ISBN 9783540327325. doi: 10.1007/11681878_14.

R. Evans and J. Gao. DeepMind AI Reduces Google Data Centre Cooling Bill by 40%, July 2016. URL https://deepmind.com/blog/article/deepmind-ai-reduces-google-data-centre-cooling-bill-40.

W. Fedus, B. Zoph, and N. Shazeer. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv:2101.03961 [cs]*, January 2021. URL http://arxiv.org/abs/2101.03961. arXiv: 2101.03961.

X. Ferrer, T. van Nuenen, J. M. Such, and N. Criado. Discovering and Categorising Language Biases in Reddit. *arXiv:2008.02754 [cs]*, August 2020. URL http://arxiv.org/abs/2008.02754. arXiv: 2008.02754.

S. Finkelstein, E. Yarzebinski, C. Vaughn, A. Ogan, and J. Cassell. The Effects of Culturally Congruent Educational Technologies on Student Achievement. In H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, editors, *Artificial Intelligence in Education*, Lecture Notes in Computer Science, pages 493–502, Berlin, Heidelberg, 2013. Springer. ISBN 9783642391125. doi: 10.1007/978-3-642-39112-5_50.

C. Flood. Fake news infiltrates financial markets. *Financial Times*, May 2017. URL https://www.ft.com/content/a37e4874-2c2a-11e7-bc4b-5528796fe35c.

P. Fortuna and S. Nunes. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4): 85:1–85:30, July 2018. ISSN 0360-0300. doi: 10.1145/3232676. URL https://doi.org/10.1145/3232676.

M. Foucault and A. Sheridan. *Discipline and punish: the birth of the prison*. Vintage, New York, 2012. ISBN 9780307819291. URL http://0-lib.myilibrary.com.catalogue.libraries.london.ac.uk?id=435863. OCLC: 817200914.

I. Gabriel. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437, September 2020a. ISSN 1572-8641. doi: 10.1007/s11023-020-09539-2. URL https://doi.org/10.1007/s11023-020-09539-2.

I. Gabriel. DeepMind x UCL | Deep Learning Lectures: Responsible Innovation, July 2020b. URL https://www.youtube.com/watch?v=MhNcWxUs-PQ.

I. Gabriel and V. Ghazavi. The Challenge of Value Alignment: from Fairer Algorithms to AI Safety. *arXiv:2101.06060 [cs]*, January 2021. URL http://arxiv.org/abs/2101.06060. arXiv: 2101.06060.

D. Garcia, M. Goel, A. K. Agrawal, and P. Kumaraguru. Collective aspects of privacy in the Twitter social network. *EPJ Data Science*, 7(1):3, December 2018. ISSN 2193-1127. doi: 10.1140/epjds/s13688-018-0130-3. URL https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-0130-3.

S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel. Counterfactual Fairness in Text Classification through Robustness. *arXiv:1809.10610 [cs, stat]*, February 2019. URL http://arxiv.org/abs/1809.10610. arXiv: 1809.10610.

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, March 2020. URL http://arxiv.org/abs/1803.09010. arXiv: 1803.09010.

S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *arXiv:2009.11462 [cs]*, September 2020. URL http://arxiv.org/abs/2009.11462. arXiv: 2009.11462.

A. Georgieff and A. Milanez. What happened to jobs at high risk of automation? Technical Report 255, OECD Publishing, January 2021. URL https://ideas.repec.org/p/oec/elsaab/255-en.html.

D. Gershgorn. GPT-3 Contains Disturbing Bias Against Muslims, January 2021. URL https://onezero.medium.com/for-some-reason-im-covered-in-blood-gpt-3-contains-disturbing-bias-against-muslims-693d275552bf.

S. Ghaffary. The algorithms that detect hate speech online are biased against black people. *Vox*, August 2019. URL https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter.

S. Ghalebikesabi, H. Wilde, J. Jewson, A. Doucet, S. Vollmer, and C. Holmes. Bias Mitigated Learning from Differentially Private Synthetic Data: A Cautionary Tale. *arXiv:2108.10934 [cs, stat]*, August 2021. URL http://arxiv.org/abs/2108.10934. arXiv: 2108.10934.

J. Golbeck. Predicting Alcoholism Recovery from Twitter. In R. Thomson, C. Dancy, A. Hyder, and H. Bisgin, editors, *Social, Cultural, and Behavioral Modeling*, Lecture Notes in Computer Science, pages 243–252, Cham, 2018. Springer International Publishing. ISBN 9783319933726. doi: 10.1007/978-3-319-93372-6_28.

R. Gorwa, R. Binns, and C. Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945, January 2020. ISSN 2053-9517. doi: 10.1177/2053951719897945. URL https://doi.org/10.1177/2053951719897945.

M. Gray and S. Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Mariner Books, 2019. URL https://ghostwork.info/.

D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang. XAI—Explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120, December 2019. doi: 10.1126/scirobotics.aay7120. URL https://www.science.org/doi/10.1126/scirobotics.aay7120.

B. Gutelius and N. Theodore. The Future of Warehouse Work: Technological Change in the U.S. Logistics Industry. Technical report, UC Berkeley Labor Center and Working Partnerships USA, 2019. URL https://laborcenter.berkeley.edu/future-of-warehouse-work/.

L. M. Hampton. Black Feminist Musings on Algorithmic Oppression. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 1–1, March 2021. doi: 10.1145/3442188.3445929. URL http://arxiv.org/abs/2101.09869. arXiv: 2101.09869.

S. C. Hampton. *Parasite and catalyst: the polarizing influence of chatbots in political discourse*. PhD thesis, University of Texas at Austin, August 2019. URL https://repositories.lib.utexas.edu/handle/2152/81204.

L. Hancox-Li and I. E. Kumar. Epistemic values in feature importance methods: Lessons from feminist epistemology. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 817–826, Virtual Event, Canada, March 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445943. URL https://doi.org/10.1145/3442188.3445943.

L. H. Hanu, J. Thewlis, and S. Haco. How AI Is Learning to Identify Toxic Online Content. *Scientific American*, 2021. URL https://www.scientificamerican.com/article/can-ai-identify-toxic-online-content/. publisher: By Laura Hanu, James Thewlis, Sasha Haco.

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

D. Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies*, 14(3):575–599, 1988.

D. J. Haraway. *The Haraway Reader*. Psychology Press, 2004. ISBN 9780415966894. Google-Books-ID: QxUr0gijyGoC.

S. G. Harding. *Feminism and Methodology: Social Science Issues*. Indiana University Press, 1987. ISBN 9780253204448. Google-Books-ID: w2gzw6zz4fIC.

M. Hardt, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. *arXiv:1610.02413 [cs]*, October 2016. URL http://arxiv.org/abs/1610.02413. arXiv: 1610.02413.

M. R. Hasan, A. K. Jha, and Y. Liu. Excessive use of online video streaming services: Impact of recommender system use, psychological factors, and motives. *Computers in Human Behavior*, 80:220–228, March 2018. ISSN 0747-5632. doi: 10.1016/j.chb.2017.11.020. URL https://www.sciencedirect.com/science/article/pii/S0747563217306581.

P. He, X. Liu, J. Gao, and W. Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv:2006.03654 [cs]*, October 2021. URL http://arxiv.org/abs/2006.03654. arXiv: 2006.03654.

D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt. Aligning AI With Shared Human Values. *arXiv:2008.02275 [cs]*, July 2021. URL http://arxiv.org/abs/2008.02275. arXiv: 2008.02275.

P. Hill Collins and N. Denzin. Toward an Afrocentric Feminist Epistemology. In Y. Lincoln, editor, *Turning Points in Qualitative Research*, volume 2. Rowman Altamira, 2003.

P. Hitlin, K. Olmstead, and S. Toor. FCC Net Neutrality Online Public Comments Contain Many Inaccuracies and Duplicates. Technical report, Pew Research Center, November 2017. URL https://www.pewresearch.org/internet/2017/11/29/public-comments-to-the-federal-communications-commission-about-net-neutrality-contain-many-inaccuracies-and-duplicates/.

K. Holt. Google's 'Verse by Verse' AI can help you write in the style of famous poets. *Engadget*, November 2020. URL https://www.engadget.com/googles-ai-poetry-verse-by-verse-202105834.html.

A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The Curious Case of Neural Text Degeneration. *arXiv:1904.09751 [cs]*, February 2020. URL http://arxiv.org/abs/1904.09751. arXiv: 1904.09751.

C. Hookway. *Scepticism*. Routledge, 1990.

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

D. Hovy and D. Yang. 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*, pages 588–602, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.49. URL https://aclanthology.org/2021.naacl-main.49.

J. Howard and S. Ruder. Universal Language Model Fine-tuning for Text Classification. *arXiv:1801.06146 [cs, stat]*, May 2018. URL http://arxiv.org/abs/1801.06146. arXiv: 1801.06146.

K. Hsieh. *Transformer Poetry*. Paper Gains Publishing, 2019. URL https://papergains.co/.

P.-S. Huang, H. Zhang, R. Jiang, R. Stanforth, J. Welbl, J. Rae, V. Maini, D. Yogatama, and P. Kohli. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. *arXiv:1911.03064 [cs]*, October 2020. URL http://arxiv.org/abs/1911.03064. arXiv: 1911.03064.

E. Hunt. Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. *The Guardian*, March 2016. URL http://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter.

K. Hunt and C. Xu. China 'employs 2 million to police internet'. *CNN*, October 2013. URL https://www.cnn.com/2013/10/07/world/asia/china-internet-monitors/index.html. publisher: CNN.

B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell. 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, pages 560–575, Virtual Event, Canada, March 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445918. URL https://doi.org/10.1145/3442188.3445918.

G. Hwang, J. Lee, C. Y. Oh, and J. Lee. 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, pages 1–6, Glasgow, Scotland Uk, May 2019. Association for Computing Machinery. ISBN 9781450359719. doi: 10.1145/3290607.3312915. URL https://doi.org/10.1145/3290607.3312915.

C. Ingraham. How rising inequality hurts everyone, even the rich. *Washington Post*, February 2018. ISSN 0190-8286. URL https://www.washingtonpost.com/news/wonk/wp/2018/02/06/how-rising-inequality-hurts-everyone-even-the-rich/.

IPCC. Global Warming of 1.5 ºC. Technical report, Intergovernmental Panel on Climate Change, 2018. URL https://www.ipcc.ch/sr15/.

C. Ischen, T. Araujo, H. Voorveld, G. van Noort, and E. Smit. Privacy concerns in chatbot interactions. In *International Workshop on Chatbot Research and Design*, pages 34–48. Springer, 2019.

L. James. How U.S. Companies & Partisans Hack Democracy to Undermine Your Voice. Technical report, New York State Office of the Attorney General, May 2021.

F. Jaumotte, S. Lall, and C. Papageorgiou. Rising Income Inequality: Technology, or Trade and Financial Globalization? *IMF Economic Review*, 61(2):271–309, June 2013. ISSN 2041-417X. doi: 10.1057/imfer.2013.7. URL https://doi.org/10.1057/imfer.2013.7.

G. Jawahar, M. Abdul-Mageed, and L. V. S. Lakshmanan. Automatic Detection of Machine Generated Text: A Critical Survey. *arXiv:2011.01314 [cs]*, November 2020. URL http://arxiv.org/abs/2011.01314. arXiv: 2011.01314.

F. Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556, April 1976. ISSN 1558-2256. doi: 10.1109/PROC.1976.10159.

R. Jeshion. Pride and Prejudiced: On the Reclamation of Slurs. *Grazer Philosophische Studien*, 97(1):106–137, March 2020. ISSN 1875-6735, 0165-9227. doi: 10.1163/18756735-09701007. URL https://brill.com/view/journals/gps/97/1/article-p106_106.xml.

Jigsaw. Unintended Bias and Identity Terms, October 2021. URL https://medium.com/jigsaw/unintended-bias-and-names-of-frequently-targeted-groups-8e0b81f80a23.

E. S. Jo and T. Gebru. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 306–316, Barcelona, Spain, January 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372829. URL https://doi.org/10.1145/3351095.3372829.

N. Jones. How to stop data centres from gobbling up the world's electricity. *Nature*, 561(7722):163–166, September 2018. doi: 10.1038/d41586-018-06610-y. URL https://www.nature.com/articles/d41586-018-06610-y.

P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *arXiv:2004.09095 [cs]*, January 2021. URL http://arxiv.org/abs/2004.09095. arXiv: 2004.09095.

D. Jurafsky and J. H. Martin. *Speech and language processing*. Pearson custom library. Pearson Education, Harlow, 2. ed., pearson new international ed edition, 2014. ISBN 9781292025438.

A. Kasirzadeh. Reasons, Values, Stakeholders: A Philosophical Framework for Explainable Artificial Intelligence. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 14, Virtual Event, Canada, March 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445866. URL https://doi.org/10.1145/3442188.3445866.

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

Z. Kenton, T. Everitt, L. Weidinger, I. Gabriel, V. Mikulik, and G. Irving. Alignment of Language Agents. *arXiv:2103.14659 [cs]*, March 2021. URL http://arxiv.org/abs/2103.14659. arXiv: 2103.14659.

O. Keyes. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):88:1–88:22, November 2018. doi: 10.1145/3274357. URL https://doi.org/10.1145/3274357.

M. Khalifa, H. Elsahar, and M. Dymetman. A Distributional Approach to Controlled Text Generation. *arXiv:2012.11635 [cs]*, May 2021. URL http://arxiv.org/abs/2012.11635. arXiv: 2012.11635.

D. Kim. Chatbot Gone Awry Starts Conversations About AI Ethics in South Korea. *The Diplomat*, January 2021. URL https://thediplomat.com/2021/01/chatbot-gone-awry-starts-conversations-about-ai-ethics-in-south-korea/.

J. Y. Kim, C. Ortiz, S. Nam, S. Santiago, and V. Datta. Intersectional Bias in Hate Speech and Abusive Language Datasets. *arXiv:2005.05921 [cs]*, May 2020. URL http://arxiv.org/abs/2005.05921. arXiv: 2005.05921.

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

J. Kocoń, A. Figas, M. Gruza, D. Puchalska, T. Kajdanowicz, and P. Kazienko. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643, September 2021. ISSN 0306-4573. doi: 10.1016/j.ipm.2021.102643. URL https://www.sciencedirect.com/science/article/pii/S0306457321001333.

A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, April 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1915768117. URL https://www.pnas.org/content/117/14/7684.

I. Kohler-Hausmann. Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination. SSRN Scholarly Paper ID 3050650, Social Science Research Network, Rochester, NY, January 2019. URL https://papers.ssrn.com/abstract=3050650.

N. Kordzadeh and M. Ghasemaghaei. Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 0(0):1–22, June 2021. ISSN 0960-085X. doi: 10.1080/0960085X.2021.1927212. URL https://doi.org/10.1080/0960085X.2021.1927212.

M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, April 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1218772110. URL https://www.pnas.org/content/110/15/5802.

C. Krittanawong, B. Narasimhan, H. U. H. Virk, H. Narasimhan, J. Hahn, Z. Wang, and W. W. Tang. Misinformation Dissemination in Twitter in the COVID-19 Era. *The American Journal of Medicine*, 133 (12):1367–1369, December 2020. ISSN 0002-9343. doi: 10.1016/j.amjmed.2020.07.012. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7426698/.

K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov. Measuring Bias in Contextualized Word Representations. *arXiv:1906.07337 [cs]*, June 2019. URL http://arxiv.org/abs/1906.07337. arXiv: 1906.07337.

K. H. Kwon, S.-I. Moon, and M. A. Stefanone. Unspeaking on Facebook? Testing network effects on self-censorship of political expressions in social network sites. *Quality & Quantity*, 49(4):1417–1435, July 2015. ISSN 1573-7845. doi: 10.1007/s11135-014-0078-8. URL https://doi.org/10.1007/s11135-014-0078-8.

C. Labbé and D. Labbé. Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science? *Scientometrics*, 94(1):379–396, January 2013. ISSN 1588-2861. doi: 10.1007/s11192-012-0781-y. URL https://doi.org/10.1007/s11192-012-0781-y.

K. Lacker. Giving GPT-3 a Turing Test, July 2020. URL https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html.

J. Lambert and E. Cone. How Robots Change the World - What automation really means for jobs, productivity and regions. Technical report, Oxford Economics, 2019. URL https://www.oxfordeconomics.com/recent-releases/how-robots-change-the-world.

I. Lapowsky. How Bots Broke the FCC's Public Comment System. *Wired*, November 2017. ISSN 1059-1028. URL https://www.wired.com/story/bots-broke-fcc-public-comment-system/.

A. Lazaridou, A. Kuncoro, E. Gribovskaya, D. Agrawal, A. Liska, T. Terzi, M. Gimenez, C. d. M. d'Autume, S. Ruder, D. Yogatama, K. Cao, T. Kocisky, S. Young, and P. Blunsom. Pitfalls of Static Language Modelling. *arXiv:2102.01951 [cs]*, February 2021. URL http://arxiv.org/abs/2102.01951. arXiv: 2102.01951.

B. Lewis and A. E. Marwick. Media Manipulation and Disinformation Online. Technical report, Data & Society, May 2017. URL https://datasociety.net/library/media-manipulation-and-disinfo-online.

M. Lewis and G. Lupyan. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, 4(10):1021–1028, October 2020. ISSN 2397-3374. doi: 10.1038/s41562-020-0918-6. URL https://www.nature.com/articles/s41562-020-0918-6.

M. Lewis, D. Yarats, Y. N. Dauphin, D. Parikh, and D. Batra. Deal or No Deal? End-to-End Learning for Negotiation Dialogues. *arXiv:1706.05125 [cs]*, June 2017. URL http://arxiv.org/abs/1706.05125. arXiv: 1706.05125.

P. Lewis, P. Stenetorp, and S. Riedel. Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets. *arXiv:2008.02637 [cs]*, August 2020. URL http://arxiv.org/abs/2008.02637. arXiv: 2008.02637.

Z. Li, S. Zhuang, S. Guo, D. Zhuo, H. Zhang, D. Song, and I. Stoica. TeraPipe: Token-Level Pipeline Parallelism for Training Large-Scale Language Models. *arXiv:2102.07988 [cs]*, September 2021. URL http://arxiv.org/abs/2102.07988. arXiv: 2102.07988.

Y. Liao and J. He. Racial mirroring effects on human-agent interaction in psychotherapeutic conversations. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, pages 430–442, Cagliari, Italy, March 2020. Association for Computing Machinery. ISBN 9781450371186. doi: 10.1145/3377325.3377488. URL https://doi.org/10.1145/3377325.3377488.

S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv:2109.07958 [cs]*, September 2021. URL http://arxiv.org/abs/2109.07958. arXiv: 2109.07958.

P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23(1):18, January 2021. doi: 10.3390/e23010018. URL https://www.mdpi.com/1099-4300/23/1/18.

R. Lippi. *English with an Accent: Language, Ideology and Discrimination in the United States*. Routledge, 1997. URL https://www.routledge.com/English-with-an-Accent-Language-Ideology-and-Discrimination-in-the-United/Lippi-Green/p/book/9780415559119.

Z. C. Lipton. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, June 2018. ISSN 1542-7730, 1542-7749. doi: 10.1145/3236386.3241340. URL https://dl.acm.org/doi/10.1145/3236386.3241340.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*, July 2019. URL http://arxiv.org/abs/1907.11692. arXiv: 1907.11692.

LSEblog2017. Doxing is a toxic practice – no matter who is targeted | Media@LSE, August 2017. URL https://blogs.lse.ac.uk/medialse/2017/08/18/the-dangers-of-doxing-and-the-implications-for-media-regulation/.

A. S. Luccioni and J. D. Viviano. What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus. *arXiv:2105.02732 [cs]*, May 2021. URL http://arxiv.org/abs/2105.02732. arXiv: 2105.02732.

L. Lucy and D. Bamman. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nuse-1.5. URL https://aclanthology.org/2021.nuse-1.5.

S. Luper. Epistemic Relativism. *Philosophical Issues*, 14:271–295, 2004. ISSN 1533-6077. URL https://www.jstor.org/stable/3050631.

A. Maass and L. Arcuri. The role of language in the persistence of stereotypes. In *Language, interaction and social cognition*, pages 129–143. Sage Publications, Inc, 1992. URL https://psycnet.apa.org/record/1992-97980-006.

A. Makazhanov, D. Rafiei, and M. Waqar. Predicting political preference of Twitter users. *Social Network Analysis and Mining*, 4(1):193, May 2014. ISSN 1869-5469. doi: 10.1007/s13278-014-0193-5. URL https://doi.org/10.1007/s13278-014-0193-5.

A. E. Manduley, A. Mertens, I. Plante, and A. Sultana. The role of social media in sex education: Dispatches from queer, trans, and racialized communities:. *Feminism & Psychology*, 28(1):152–170, February 2018. doi: 10.1177/0959353517717751. URL https://journals.sagepub.com/eprint/wZsRhKyIrHE7KMyN9fJB/full.

C. B. Mann. Can Conversing with a Computer Increase Turnout? Mobilization Using Chatbot Communication. *Journal of Experimental Political Science*, 8(1):51–62, 2021. URL https://ideas.repec.org/a/cup/jexpos/v8y2021i1p51-62_5.html.

H. Mao, X. Shuai, and A. Kapadia. 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, pages 1–12, Chicago, Illinois, USA, October 2011. Association for Computing Machinery. ISBN 9781450310024. doi: 10.1145/2046556.2046558. URL https://doi.org/10.1145/2046556.2046558.

V. Marda and S. Narayan. On the importance of ethnographic methods in AI research. *Nature Machine Intelligence*, 3(3):187–189, March 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00323-0. URL https://www.nature.com/articles/s42256-021-00323-0.

M. Marino. The Racial Formation of Chatbots. *CLCWeb: Comparative Literature and Culture*, 16(5), December 2014. ISSN 1481-4374. doi: 10.7771/1481-4374.2560. URL https://docs.lib.purdue.edu/clcweb/vol16/iss5/13.

A. Markov. Essai d'une recherche statistique sur le texte du roman "eugène oněgin", illustrant la liaison des épreuves en chaîne. *Bulletin de l'Académie Impériale des Sciences de St.-Pétersbourg.*, 7(3):153–162, 1913.

D. Martin Jr., V. Prabhakaran, J. Kuhlberg, A. Smart, and W. S. Isaac. Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics. *arXiv:2005.07572 [cs, stat]*, May 2020. URL http://arxiv.org/abs/2005.07572. arXiv: 2005.07572.

K. McGuffie and A. Newhouse. The Radicalization Risks of GPT-3 and Advanced Neural Language Models. *arXiv:2009.06807 [cs]*, September 2020. URL http://arxiv.org/abs/2009.06807. arXiv: 2009.06807.

K. McKee, X. Bai, and S. Fiske. Understanding Human Impressions of Artificial Intelligence. *PsyArxiv*, 2021. URL https://psyarxiv.com/5ursp/.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A Survey on Bias and Fairness in Machine Learning. *arXiv:1908.09635 [cs]*, September 2019. URL http://arxiv.org/abs/1908.09635. arXiv: 1908.09635.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):115:1–115:35, July 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

J. Menasce Horowitz, R. Igielnik, and R. Kochhar. Trends in U.S. income and wealth inequality. Technical report, Pew Research Center, January 2020. URL https://www.pewresearch.org/social-trends/2020/01/09/trends-in-income-and-wealth-inequality/.

S. Milano, M. Taddeo, and L. Floridi. Recommender systems and their ethical challenges. *AI & SOCIETY*, 35(4):957–967, December 2020. ISSN 1435-5655. doi: 10.1007/s00146-020-00950-y. URL https://doi.org/10.1007/s00146-020-00950-y.

T. Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv:1706.07269 [cs]*, August 2018. URL http://arxiv.org/abs/1706.07269. arXiv: 1706.07269.

T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267: 1–38, February 2019. ISSN 0004-3702. doi: 10.1016/j.artint.2018.07.007. URL https://www.sciencedirect.com/science/article/pii/S0004370218305988.

A. S. Miner, A. Milstein, S. Schueller, R. Hegde, C. Mangurian, and E. Linos. Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health. *JAMA internal medicine*, 176(5):619–625, May 2016. ISSN 2168-6114. doi: 10.1001/jamainternmed.2016.0400. URL https://europepmc.org/articles/PMC4996669.

S. Mohamed, M.-T. Png, and W. Isaac. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33(4):659–684, December 2020. ISSN 2210-5441. doi: 10.1007/s13347-020-00405-8. URL https://doi.org/10.1007/s13347-020-00405-8.

W. Moncur, J. Masthoff, and E. Reiter. Facilitating benign deceit in mediated communication. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '09, pages 3383–3388, Boston, MA, USA, April 2009. Association for Computing Machinery. ISBN 9781605582474. doi: 10.1145/1520340.1520490. URL https://doi.org/10.1145/1520340.1520490.

W. Moncur, J. Masthoff, E. Reiter, Y. Freer, and H. Nguyen. Providing Adaptive Health Updates Across the Personal Social Network. *Human–Computer Interaction*, 29(3):256–309, May 2014. ISSN 0737-0024. doi: 10.1080/07370024.2013.819218. URL https://doi.org/10.1080/07370024.2013.819218.

A. A. Morgan-Lopez, A. E. Kim, R. F. Chew, and P. Ruddle. Predicting age groups of Twitter users based on language and metadata features. *PLOS ONE*, 12(8):e0183537, August 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0183537. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0183537.

D. F. Mujtaba and N. R. Mahapatra. Ethical Considerations in AI-Based Recruitment. In *2019 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–7, November 2019. doi: 10.1109/ISTAS48451.2019.8937920. ISSN: 2158-3412.

M. Murgia. DeepMind's Lila Ibrahim: 'It's hard not to go through imposter syndrome'. *Financial Times*, July 2021. URL https://www.ft.com/content/c43cd848-367a-4c67-8cf1-7f34d3eaf84e.

D. Mytton. Data centre water consumption. *NPJ Clean Water*, 4(1):1–6, February 2021. ISSN 2059-7037. doi: 10.1038/s41545-021-00101-w. URL https://www.nature.com/articles/s41545-021-00101-w.

J. Möller, D. Trilling, N. Helberger, and B. van Es. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, 21(7):959–977, July 2018. ISSN 1369-118X. doi: 10.1080/1369118X.2018.1444076. URL https://doi.org/10.1080/1369118X.2018.1444076.

M. Nadeem, A. Bethke, and S. Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv:2004.09456 [cs]*, April 2020. URL http://arxiv.org/abs/2004.09456. arXiv: 2004.09456.

N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. *arXiv:2010.00133 [cs]*, September 2020. URL http://arxiv.org/abs/2010.00133. arXiv: 2010.00133.

A. Narayanan. How to Recognize AI Snake Oil, January 2021. URL https://www.cs.princeton.edu/news/how-recognize-ai-snake-oil.

D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder. "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):439–448, 2013. ISSN 2334-0770. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14381.

S. I. Nikolenko. *Synthetic Data for Deep Learning*, volume 174 of *Springer Optimization and Its Applications*. Springer International Publishing, Cham, 2021. ISBN 9783030751777 9783030751784. doi: 10.1007/978-3-030-75178-4. URL https://link.springer.com/10.1007/978-3-030-75178-4.

NLP for Positive Impact 2021. Workshop on NLP for Positive Impact at ACL-IJCNLP, 2021. URL https://sites.google.com/view/nlp4positiveimpact2021.

S. U. Noble. *Algorithms of Oppression*. NYU Press, 2018. URL https://nyupress.org/9781479837243/algorithms-of-oppression.

D. Nozza, F. Bianchi, and D. Hovy. 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*, pages 2398–2406, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.191. URL https://aclanthology.org/2021.naacl-main.191.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019. doi: 10.1126/science.aax2342. URL https://www.science.org/doi/abs/10.1126/science.aax2342.

B. of Labor Statistics. Interpreters and Translators: Occupational Outlook Handbook. Technical report, U.S. Department of Labour, 2021. URL https://www.bls.gov/ooh/media-and-communication/interpreters-and-translators.htm#:~:text=3%25-,Employment%20of%20interpreters%20and%20translators%20is%20projected%20to%20grow%2020,the%20average%20for%20all%20occupations.&text=The%20ongoing%20need%20for%20military,in%20more%20jobs%20as%20well. publisher: U.S. Bureau of Labour Statistics.

K. Ognyanova, D. Lazer, R. E. Robertson, and C. Wilson. 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. doi: 10.37016/mr-2020-024. URL https://misinforeview.hks.harvard.edu/article/misinformation-in-action-fake-news-exposure-is-linked-to-lower-trust-in-media-higher-trust-in-government-when-your-side-is-in-power/.

P. Oosterhoff. Online censors are a barrier to sex education, 2016. URL https://www.scidev.net/global/opinions/online-censors-sex-education-porn/.

D. O'Callaghan, D. Greene, M. Conway, J. Carthy, and P. Cunningham. Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems. *Social Science Computer Review*, 33(4):459–478, August 2015. ISSN 0894-4393. doi: 10.1177/0894439314555329. URL https://doi.org/10.1177/0894439314555329.

G. PAIR. *People + AI Guidebook*. Google, May 2019. URL https://design.google/ai-guidebook.

A. Pardes. The Emotional Chatbots Are Here to Probe Our Feelings. *Wired*, January 2018. ISSN 1059-1028. URL https://www.wired.com/story/replika-open-source/.

G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. P. Seligman. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6):934–952, June 2015. ISSN 1939-1315, 0022-3514. doi: 10.1037/pspp0000020. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/pspp0000020.

F. Pasquale. *The Black Box Society*. Harvard University Press, 2016. URL https://dl.acm.org/doi/abs/10.5555/2717112.

D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean. Carbon Emissions and Large Neural Network Training. *arXiv:2104.10350 [cs]*, April 2021. URL http://arxiv.org/abs/2104.10350. arXiv: 2104.10350.

D. Perez-Marin and I. Pascual-Nieto. *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*. Information Science Reference - Imprint of: IGI Publishing, Hershey, PA, 2011. ISBN 9781609606176.

N. Persily and J. A. Tucker. *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge University Press, September 2020. ISBN 9781108858779. Google-Books-ID: TgH3DwAAQBAJ.

PerspectiveAPI. Perspective API | Developers, 2021. URL https://support.perspectiveapi.com/s/about-the-api-attributes-and-languages.

J. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder. UNKs Everywhere: Adapting Multilingual Language Models to New Scripts. *arXiv:2012.15562 [cs]*, September 2021. URL http://arxiv.org/abs/2012.15562. arXiv: 2012.15562.

A. Pilipiszyn. GPT-3 Powers the Next Generation of Apps, March 2021. URL https://openai.com/blog/gpt-3-apps/. publisher: OpenAI.

D. Preoţiuc-Pietro, Y. Liu, D. Hopkins, and L. Ungar. 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)*, pages 729–740, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1068. URL http://aclweb.org/anthology/P17-1068.

K. Quach. Researchers made an OpenAI GPT-3 medical chatbot as an experiment. It told a mock patient to kill themselves. *The Register*, October 2020. URL https://www.theregister.com/2020/10/28/gpt3_medical_chatbot_experiment/.

D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. 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*, pages 180–185, October 2011. doi: 10.1109/PASSAT/SocialCom.2011.26.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving Language Understanding by Generative Pre-Training. 2018a.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language Models are Unsupervised Multitask Learners. 2018b.

J. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. van den Driessche, L. A. Hendricks, M. Rauh, P.-S. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen, S. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J.-B. Lespiau, M. Tsimpoukelli, N. Grigorev, D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. de Masson d'Autume, Y. Li, T. Terzi, I. Babuschkin, A. Clark, D. de Las Casas, A. Guy, J. Bradbury, M. Johnson, L. Weidinger, I. Gabriel, W. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. Ayoub, J. Stanway, L. Bennett, D. Hassabis, K. Kavukcuoglu, and G. Irving. Scaling language models: Methods, analysis & insights from training Gopher. *arXiv submission*, 2021.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:1910.10683 [cs, stat]*, July 2020. URL http://arxiv.org/abs/1910.10683. arXiv: 1910.10683.

I. D. Raji. Handle with Care: Lessons for Data Science from Black Female Scholars. *Patterns*, 1(8):100150, November 2020. ISSN 2666-3899. doi: 10.1016/j.patter.2020.100150. URL https://www.sciencedirect.com/science/article/pii/S2666389920301987.

I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *arXiv:2001.00973 [cs]*, January 2020. URL http://arxiv.org/abs/2001.00973. arXiv: 2001.00973.

S. Ramaswamy, O. Thakkar, R. Mathews, G. Andrew, H. B. McMahan, and F. Beaufays. Training Production Language Models without Memorizing User Data. *arXiv:2009.10031 [cs, stat]*, September 2020. URL http://arxiv.org/abs/2009.10031. arXiv: 2009.10031.

P. Ranade, A. Piplai, S. Mittal, A. Joshi, and T. Finin. Generating Fake Cyber Threat Intelligence Using Transformer-Based Models. *arXiv:2102.04351 [cs]*, June 2021. URL http://arxiv.org/abs/2102.04351. arXiv: 2102.04351.

E. Rand. *Reclaiming Queer: Activist & Academic Rhetorics of Resistance*. University of Alabama Press, 2014.

E. Reiter. Could NLG systems injure or even kill people?, October 2020. URL https://ehudreiter.com/2020/10/20/could-nlg-systems-injure-or-even-kill-people/.

M. Rimmer. Patent-Busting: The Public Patent Foundation, Gene Patents and the Seed Wars. In C. Lawson and J. Sanderson, editors, *The Intellectual Property and Food Project*. Routledge, 2013.

A. Romano. A group of YouTubers is claiming the site systematically demonetizes queer content. *Vox*, October 2019. URL https://www.vox.com/culture/2019/10/10/20893258/youtube-lgbtq-censorship-demonetization-nerd-city-algorithm-report.

J. Rosa and N. Flores. Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society*, 46(5):621–647, November 2017. ISSN 0047-4045, 1469-8013. doi: 10.1017/S0047404517000562. URL https://www.cambridge.org/core/journals/language-in-society/article/abs/unsettling-race-and-language-toward-a-raciolinguistic-perspective/30FFC5253F465905D75CDFF1C1363AE3.

C. Rosset. Turing-NLG: A 17-billion-parameter language model by Microsoft, February 2020. URL https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/.

A. Rubel, A. Pham, and C. Castro. Agency Laundering and Algorithmic Decision Systems. In N. G. Taylor, C. Christian-Lamb, M. H. Martin, and B. Nardi, editors, *Information in Contemporary Society*, Lecture Notes in Computer Science, pages 590–598, Cham, 2019. Springer International Publishing. ISBN 9783030157425. doi: 10.1007/978-3-030-15742-5_56.

S. Ruder. Why You Should Do NLP Beyond English, August 2020. URL https://ruder.io/nlp-beyond-english/.

A. Sabeti. Teaching GPT-3 to Identify Nonsense, July 2020. URL https://arr.am/2020/07/25/gpt-3-uncertainty-prompts/.

N. A. Sales. Secrecy and National Security Investigations. *Alabama Law Review*, 58:811, 2006. URL https://heinonline.org/HOL/Page?handle=hein.journals/bamalr58&id=821&div=&collection=.

N. Sambasivan and J. Holbrook. Toward responsible AI for the next billion users. *Interactions*, 26(1):68–71, December 2018. ISSN 1072-5520. doi: 10.1145/3298735. URL https://doi.org/10.1145/3298735.

N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran. Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 315–328, Virtual Event, Canada, March 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445896. URL https://doi.org/10.1145/3442188.3445896.

M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163.

K. M. Saunders. The Law and Ethics of Trade Secrets: A Case Study. *California Western Law Review*, 42:209, 2005. URL https://heinonline.org/HOL/Page?handle=hein.journals/cwlr42&id=215&div=&collection=.

A. Schmidt and M. Wiegand. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1101. URL http://aclweb.org/anthology/W17-1101.

A. T. Schmidt and B. Engelen. The ethics of nudging: An overview. *Philosophy Compass*, 15(4):e12658, April 2020. ISSN 1747-9991, 1747-9991. doi: 10.1111/phc3.12658. URL https://onlinelibrary.wiley.com/doi/10.1111/phc3.12658.

B. Schneier. Bots Are Destroying Political Discourse As We Know It. *The Atlantic*, January 2020. URL https://www.theatlantic.com/technology/archive/2020/01/future-politics-bots-drowning-out-humans/604489/.

M. Schroepfer. How AI is getting better at detecting hate speech, November 2020. URL https://ai.facebook.com/blog/how-ai-is-getting-better-at-detecting-hate-speech/.

R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni. Green AI. *Communications of the ACM*, 63(12):54–63, November 2020. ISSN 0001-0782. doi: 10.1145/3381831. URL https://doi.org/10.1145/3381831.

E. Seger, S. Avin, G. Pearson, M. Briers, S. Ó Heigeartaigh, and H. Bacon. Tackling threats to informed decision-making in democratic societies. Technical report, Alan Turing Institute, 2020. URL https://www.turing.ac.uk/research/publications/tackling-threats-informed-decision-making-democratic-societies.

D. Shah, H. A. Schwartz, and D. Hovy. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, 2020. doi: 10.18653/v1/2020.acl-main.468. URL http://arxiv.org/abs/1912.11078. arXiv: 1912.11078.

A. Shahbaz and A. Funk. Social Media Surveillance. Technical report, Freedom House, 2019. URL https://freedomhouse.org/report/freedom-on-the-net/2019/the-crisis-of-social-media/social-media-surveillance.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.tb01338.x.

E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng. Societal Biases in Language Generation: Progress and Challenges. *arXiv:2105.04054 [cs]*, June 2021. URL http://arxiv.org/abs/2105.04054. arXiv: 2105.04054.

M. Sloane, E. Moss, O. Awomolo, and L. Forlano. Participation is not a Design Fix for Machine Learning. *arXiv:2007.02423 [cs]*, August 2020. URL http://arxiv.org/abs/2007.02423. arXiv: 2007.02423.

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

I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, and J. Wang. Release Strategies and the Social Impacts of Language Models. *arXiv:1908.09203 [cs]*, November 2019. URL http://arxiv.org/abs/1908.09203. arXiv: 1908.09203.

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

D. Sravani, L. Kameswari, and R. Mamidi. Political Discourse Analysis: A Case Study of Code Mixing and Code Switching in Political Speeches. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–5, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.calcs-1.1. URL https://aclanthology.org/2021.calcs-1.1.

T. Stahl. Indiscriminate mass surveillance and the public sphere. *Ethics and Information Technology*, 18(1): 33–39, March 2016. ISSN 1572-8439. doi: 10.1007/s10676-016-9392-2. URL https://doi.org/10.1007/s10676-016-9392-2.

StanfordNaturalProcessingGroup. The Stanford Natural Language Processing Group, 2021. URL https://nlp.stanford.edu/projects/coref.shtml.

J. Stilgoe. Who's driving innovation. *New Technologies and the Collaborative State. Cham, Switzerland: Palgrave Macmillan*, 2020.

J. Stilgoe, R. Owen, and P. Macnaghten. Developing a framework for responsible innovation. *Research Policy*, 42(9):1568–1580, November 2013. ISSN 0048-7333. doi: 10.1016/j.respol.2013.05.008. URL https://www.sciencedirect.com/science/article/pii/S0048733313000930.

E. Strubell, A. Ganesh, and A. McCallum. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]*, June 2019. URL http://arxiv.org/abs/1906.02243. arXiv: 1906.02243.

S. Sullivan and N. Tuana, editors. *Race and epistemologies of ignorance*. SUNY series, philosophy and race. State University of New York Press, Albany, 2007. ISBN 9780791471012 9780791471029. OCLC: ocm70676503.

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

Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, W. Liu, Z. Wu, W. Gong, J. Liang, Z. Shang, P. Sun, W. Liu, X. Ouyang, D. Yu, H. Tian, H. Wu, and H. Wang. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation. *arXiv:2107.02137 [cs]*, July 2021. URL http://arxiv.org/abs/2107.02137. arXiv: 2107.02137.

A. Tamkin, M. Brundage, J. Clark, and D. Ganguli. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. *arXiv:2102.02503 [cs]*, February 2021. URL http://arxiv.org/abs/2102.02503. arXiv: 2102.02503.

R. S. J. Tol. The impact of climate change and the social cost of carbon. In *Routledge Handbook of Energy Economics*. Routledge, 2019.

N. Tomasev, K. R. McKee, J. Kay, and S. Mohamed. Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 254–265, July 2021. doi: 10.1145/3461702.3462540. URL http://arxiv.org/abs/2102.04257. arXiv: 2102.04257.

J. M. Twenge. More Time on Technology, Less Happiness? Associations Between Digital-Media Use and Psychological Well-Being. *Current Directions in Psychological Science*, 28(4):372–379, August 2019. ISSN 0963-7214. doi: 10.1177/0963721419838244. URL https://doi.org/10.1177/0963721419838244.

E. Van den Broeck, B. Zarouali, and K. Poels. Chatbot advertising effectiveness: When does the message get through? *Computers in Human Behavior*, 98:150–157, September 2019. ISSN 0747-5632. doi: 10.1016/j.chb.2019.04.009. URL https://www.sciencedirect.com/science/article/pii/S0747563219301499.

R. Van Noorden. Publishers withdraw more than 120 gibberish papers. *Nature*, February 2014. ISSN 1476-4687. doi: 10.1038/nature.2014.14763. URL https://www.nature.com/articles/nature.2014.14763.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

VersebyVerse. Verse by Verse, 2020. URL https://sites.research.google/versebyverse/. publisher:.

J. Vincent. The invention of AI 'gaydar' could be the start of something much worse. *The Verge*, September 2017. URL https://www.theverge.com/2017/9/21/16332760/ai-sexuality-gaydar-photo-physiognomy.

K. Vredenburgh. The Right to Explanation. *Journal of Political Philosophy*, 0(0):1–21, 2021. ISSN 1467-9760. doi: 10.1111/jopp.12262. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jopp.12262.

C. Véliz. Privacy matters because it empowers us all | Aeon Essays. *Aeon*, September 2019. URL https://aeon.co/essays/privacy-matters-because-it-empowers-us-all.

B. Vézina and S. Hinchcliff Pearson. Should CC-Licensed Content be Used to Train AI? It Depends., March 2021. URL https://creativecommons.org/2021/03/04/should-cc-licensed-content-be-used-to-train-ai-it-depends/.

D. Wallace, F. Tramer, M. Jagielski, and A. Herbert-Voss. Does GPT-2 Know Your Phone Number?, December 2020. URL http://bair.berkeley.edu/blog/2020/12/20/lmmem/.

A. Wang and O. Russakovsky. Directional Bias Amplification. *arXiv:2102.12594 [cs]*, June 2021. URL http://arxiv.org/abs/2102.12594. arXiv: 2102.12594.

A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*, 2019a.

Y. Wang and M. Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114(2):246–257, February 2018. ISSN 1939-1315, 0022-3514. doi: 10.1037/pspa0000098. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/pspa0000098.

Y. Wang, M. McKee, A. Torbica, and D. Stuckler. Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. *Social Science & Medicine*, 240:112552, November 2019b. ISSN 0277-9536. doi: 10.1016/j.socscimed.2019.112552. URL https://www.sciencedirect.com/science/article/pii/S0277953619305465.

Z. Wang, A. W. Yu, O. Firat, and Y. Cao. Towards Zero-Label Language Learning. *arXiv:2109.09193 [cs]*, September 2021. URL http://arxiv.org/abs/2109.09193. arXiv: 2109.09193.

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

J. Welbl, A. Glaese, J. Uesato, S. Dathathri, J. Mellor, L. A. Hendricks, K. Anderson, P. Kohli, B. Coppin, and P.-S. Huang. Challenges in Detoxifying Language Models. *arXiv:2109.07445 [cs]*, September 2021. URL http://arxiv.org/abs/2109.07445. arXiv: 2109.07445.

T.-H. Wen, D. Vandyke, N. Mrksic, M. Gasic, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, and S. Young. A Network-based End-to-End Trainable Task-oriented Dialogue System. *arXiv:1604.04562 [cs, stat]*, April 2017. URL http://arxiv.org/abs/1604.04562. arXiv: 1604.04562.

M. West, R. Kraut, and H. Ei Chew. I'd blush if I could : closing gender divides in digital skills through education. Technical report, UNESCO, 2019. URL https://repositorio.minedu.gob.pe/handle/20.500.12799/6598.

G. I. Winata, A. Madotto, Z. Lin, R. Liu, J. Yosinski, and P. Fung. Language Models are Few-shot Multilingual Learners. *arXiv:2109.07684 [cs]*, September 2021. URL http://arxiv.org/abs/2109.07684. arXiv: 2109.07684.

L. Winner. Do Artifacts Have Politics? *Daedalus*, 109(1):121–136, 1980. ISSN 0011-5266. URL https://www.jstor.org/stable/20024652.

A. Xu, E. Pathak, E. Wallace, S. Gururangan, M. Sap, and D. Klein. Detoxifying Language Models Risks Marginalizing Minority Voices. *arXiv:2104.06390 [cs]*, April 2021. URL http://arxiv.org/abs/2104.06390. arXiv: 2104.06390.

E. Yang and M. E. Roberts. Censorship of Online Encyclopedias: Implications for NLP Models. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 537–548, Virtual Event, Canada, March 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445916. URL https://doi.org/10.1145/3442188.3445916.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://papers.nips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

K. Yee, U. Tantipongpipat, and S. Mishra. Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–24, October 2021. ISSN 2573-0142. doi: 10.1145/3479594. URL http://arxiv.org/abs/2105.08667. arXiv: 2105.08667.

M. Yesilada and S. Lewandowsky. A systematic review: The YouTube recommender system and pathways to problematic content. *PsyArxiv*, June 2021. URL https://psyarxiv.com/6pv5c/.

W. Youyou, M. Kosinski, and D. Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040, January 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1418680112. URL https://www.pnas.org/content/112/4/1036.

S. Zdenek. "Just Roll Your Mouse Over Me": Designing Virtual Women for Customer Service on the Web. *Technical Communication Quarterly*, 16(4):397–430, August 2007. ISSN 1057-2252, 1542-7625. doi: 10.1080/10572250701380766. URL http://www.tandfonline.com/doi/abs/10.1080/10572250701380766.

R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. Defending Against Neural Fake News. *arXiv:1905.12616 [cs]*, December 2020. URL http://arxiv.org/abs/1905.12616. arXiv: 1905.12616.

C. Zhang, P. Benz, C. Lin, A. Karjauv, J. Wu, and I. S. Kweon. A Survey On Universal Adversarial Attack. *arXiv:2103.01498 [cs]*, March 2021a. URL http://arxiv.org/abs/2103.01498. arXiv: 2103.01498.

N. Zhang, L. Li, X. Chen, S. Deng, Z. Bi, C. Tan, F. Huang, and H. Chen. Differentiable Prompt Makes Pre-trained Language Models Better Few-shot Learners. *arXiv:2108.13161 [cs]*, October 2021b. URL http://arxiv.org/abs/2108.13161. arXiv: 2108.13161.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *arXiv:1707.09457 [cs, stat]*, July 2017. URL http://arxiv.org/abs/1707.09457. arXiv: 1707.09457.

J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang. Gender Bias in Contextualized Word Embeddings. *arXiv:1904.03310 [cs]*, April 2019. URL http://arxiv.org/abs/1904.03310. arXiv: 1904.03310.

T. Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate Before Use: Improving Few-Shot Performance of Language Models. *arXiv:2102.09690 [cs]*, June 2021. URL http://arxiv.org/abs/2102.09690. arXiv: 2102.09690.

J. Zou and L. Schiebinger. AI can be sexist and racist — it's time to make it fair. *Nature*, 559(7714):324–326, July 2018. doi: 10.1038/d41586-018-05707-8. URL https://www.nature.com/articles/d41586-018-05707-8.

S. Zuboff. *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. Profile books, London, 2019. ISBN 9781781256848 9781781256855.

J. Złotowski, D. Proudfoot, K. Yogeeswaran, and C. Bartneck. Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction. *International Journal of Social Robotics*, 7(3):347–360, June 2015. ISSN 1875-4805. doi: 10.1007/s12369-014-0267-6. URL https://doi.org/10.1007/s12369-014-0267-6.

# A. Appendix

## A.1. Definitions

### A.1.1. Language Models

Language Models (LMs) are machine learning models that are trained to represent a probability distribution $p(w)$ over sequences of utterances $w$ from a pre-specified domain (letters, words, sentences, paragraphs, documents). LMs aim to capture statistical properties of the sequences of utterances present in their training corpus and can be used to make probabilistic predictions regarding sequences of utterances (Bengio, 2008). 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 (generative language models) 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:

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

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, connoted as LMs. The emergence of LMs is described in detail in the section on a brief history of "Large" Language 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 (Holtzman et al., 2020)) – have been found to produce better results in practice (Brown et al., 2020)). LMs typically aim to mirror language found in the training data. However, they can also be optimised toward other tasks or objectives. For example, a LM can be optimised for dialogue, by predicting utterances that are most appropriate to maintain a conversation.

### A.1.2. Language Agents

Language agents (LAs) are machine learning systems that are restricted to providing only natural language text-output (Kenton et al., 2021). LAs may generate text-output based on LM predictions. LAs that are optimised to engage a person in direct dialogue are also referred to as "Conversational Agents" (CAs) (Perez-Marin and Pascual-Nieto, 2011).

### A.1.3. Language Technologies

LMs can be used in language technologies (LTs) such as voice assistants including Siri (Apple), Google Assistant (Google), or Alexa (Amazon), text generation tools such as AutoCorrect or SmartReply, and translation and summarisation tools. Language technologies can serve different purposes, for example providing information, entertainment, or productivity aids to the user.

Powerful large language models (LLMs) may lead to improved versions of existing language technologies. However, they may also make new types of language technology possible. For example, they may create conversational interfaces with human users where the use of this technology is indistinguishable from interaction with a human counterpart. Such applications are discussed in more detail in section V. Human-Computer Interaction Harms.

**Distinguishing "statistical bias" from "social bias"** Concerns regarding "bias" in language models generally revolve around distributional skews that result in unfavourable impacts for particular social groups (Sheng et al., 2021). We note that there are different definitions of "bias" and "discrimination" in classical statistics compared to sociotechnical studies. In classical statistics, "bias" designates the difference between a model's prediction and the ground truth (Dietterich and Kong, 1995); in machine learning, minimising statistical bias is a component of reducing error (Dietterich and Kong, 1995). 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 (Blodgett et al., 2020). Developing mechanisms to quantify the latter type of bias is an area of active research, where qualitative and quantitative measures have been established (Barocas et al., 2019; Hardt et al., 2016).

**Distinguishing statistical from sociotechnical notions of "discrimination"** Similarly, the definition of "discrimination" is multiplicitous. Traditionally in machine learning, this term refers to making distinctions between possible categories or target classes (Bowker and Star, 1999). 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 (Barocas and Selbst, 2016).

## A.2. References Table

***Table 2.*** *References providing evidence for each risk covered in this report.*

| **Risk** | **Evidence in NLP**, ***Evidence in LMs (GPT-2, GPT-3, T5, Gopher)*** |
|---|---|
| **Discrimination, Exclusion and Toxicity** | |
| 2.1.2 Social stereotypes and unfair discrimination | Blodgett et al. (2020); Caliskan et al. (2017); Dodge et al. (2021); Ferrer et al. (2020); Zhao et al. (2017)<br>*Abid et al. (2021); Huang et al. (2020); Lucy and Bamman (2021); Nadeem et al. (2020); Nangia et al. (2020); Nozza et al. (2021)* |
| 2.1.3 Exclusionary norms | Cao and Daumé III (2020) |
| 2.1.4 Toxic language | Duggan (2017); Gehman et al. (2020); Gorwa et al. (2020); Luccioni and Viviano (2021)<br>*Rae et al. (2021); Wallace et al. (2020)* |
| 2.1.5 Lower performance by social group | Blodgett and O'Connor (2017); Blodgett et al. (2016); Joshi et al. (2021); Koenecke et al. (2020); Ruder (2020)<br>*Winata et al. (2021)* |
| **Information Hazards** | |
| 2.2.2 Compromise privacy by leaking private information | Dobberstein (2021); Kim (2021)<br>*Carlini et al. (2021)* |
| 2.2.3 Compromise privacy by correctly inferring private information | Garcia et al. (2018); Golbeck (2018); Makazhanov et al. (2014); Morgan-Lopez et al. (2017); Nguyen et al. (2013); Park et al. (2015); Preoţiuc-Pietro et al. (2017) |
| 2.2.4 Risks from leaking or correctly inferring sensitive information | *Wallace et al. (2020)* |

| | |
|---|---|
| **Misinformation Harms** | |
| 2.3.2 Disseminating false or misleading information | Allcott et al. (2019); Krittanawong et al. (2020); Wang et al. (2019b) *Branwen (2020); Dale (2021); Lacker (2020); Lin et al. (2021); Rae et al. (2021); Zhang et al. (2021b)* |
| 2.3.3 Causing material harm by disseminating misinformation e.g. in medicine or law | *Quach (2020)* |
| 2.3.4 Leading users to perform unethical or illegal actions | *Hendrycks et al. (2021)* |
| **Malicious Uses** | |
| 2.4.2 Making disinformation cheaper and more effective | Hampton (2021); Mann (2021); Schneier (2020); Zellers et al. (2020) *Hao (2020); McGuffie and Newhouse (2020)* |
| 2.4.3 Facilitating fraud and impersonation scams | Lewis et al. (2017); Van Noorden (2014) |
| 2.4.4 Assisting code generation for cyber attacks, weapons, or malicious use | *Chen et al. (2021a)* |
| 2.4.5 Illegitimate surveillance and censorship | Shahbaz and Funk (2019) |
| **Human-Computer Interaction Harms** | |
| 2.5.2 Anthropomorphising systems can lead to overreliance or unsafe use | Kim and Sundar (2012) |
| 2.5.3 Create avenues for exploiting user trust to obtain private information | Ischen et al. (2019); Lewis et al. (2017); Van den Broeck et al. (2019) |
| 2.5.4 Promoting harmful stereotypes by implying gender or ethnic identity | Cercas Curry et al. (2020); Hwang et al. (2019); Marino (2014); Zdenek (2007) |
| **Automation, access, and environmental harms** | |
| 2.6.2 Environmental harms from operating LMs | Strubell et al. (2019) *Bender et al. (2021); Patterson et al. (2021)* |
| 2.6.3 Increasing inequality and negative effects on job quality | |
| 2.6.4 Undermining creative economies | Hsieh (2019) |
| 2.6.5 Disparate access to benefits due to hardware, software, skill constraints | *Bender et al. (2021)* |