

## **How can we develop more effective ways to mitigate bias in language models?**

### **Abstract**

Large Language models are a new technology which are predicted to have a significant impact on society. They are in use now and starting to transform society in a variety of ways. However, these systems still face many limitations, for example they learn historical biases from their training data and output discriminatory statements as a result. There is a growing body of research dedicated to bias in machine learning, but we are yet to develop effective ways to mitigate the harms that these systems pose, especially in the domain of natural language. In this paper I explore why this is by asking how language models threaten to harm certain demographics and why previous work on this topic has failed to result in conclusive answers. I find that many previous methods are disconnected from the way these systems are used in the real world. In addition, the concepts of bias and fairness are relatively new in machine learning which leads to them being treated very differently by different researchers. I argue that more research needs to be conducted on the effects these systems have when applied in social contexts, and doing this will require interdisciplinary collaboration with disciplines that have more experience examining bias in society. I use my findings to propose recommendations for future research to consider and advise ways to facilitate the necessary research. I also suggest how a framework from sociolinguistics can be adapted into a method to evaluate bias in language models.

## **Projected Influence of Language models**

### **Growing interest**

The start of 2023 has made it evident that Large Language Models (LLMs) are set to have a profound impact on society. They have demonstrated impressive capabilities which are highly applicable in the real world such as question answering, reading comprehension and translation. Eloundou et al. (2023) conducted a study on LLMs' projected impact on the U.S. economy and expect their impact to be widespread and significant. The value of such systems has been recognised by companies like Microsoft who have dedicated significant resources to exclusively licensing OpenAI's GPT series and integrating it in many of their applications (Microsoft, 2020). One of these applications is Bing search, which, with its new AI powered chatbot, is expected to compete with Google who have historically had a monopoly on search (The Economic Times, 2023). LLMs have not only garnered interest by being profitable. Kosinski (2023) discovered evidence of Theory-Of-Mind like behaviour in the latest iterations of these systems, and discussions of making them multimodal (combining text-processing with processing of other forms of data, e.g. (Alayrac et al., 2020)) have led to talk about the imminent advent of AGI (Artificial General Intelligence) (Fei et al., 2022). An open letter was published and signed by many AI experts calling for a six month pause on all AI research on a greater scale than current state of the art systems (Future of Life Institute, 2023). This letter has received much criticism (Gebru et al., 2023) and we are yet to see the full effects of language models (LMs) in industry settings, but these systems are expected to have a significant impact.

### **Technical requirements**

Additionally, these systems have been improving at an alarming rate. With GPT-3, Brown et al. (2020) showed that scaling the parameters of the training model continues to improve its accuracy at numerous tasks which are prominent in NLP research (NLP (Natural Language Processing) is the use of computers to interact with and generate human language). This is an important revelation because it indicates a straightforward path for making versions that better imitate humans. For those that intend to profit from these systems, this is good news. The developers could simply repeat the process but on a larger scale and with more data, and this is already being done, evidenced by the quick release of GPT-4. However, there are many other limitations that these systems have which need to be tackled before they can be scaled without fear of adverse effects. Currently, research into the safety of these systems is progressing slower than their capabilities are. There are now many of these systems in development and their usage rapidly increases the volume of text data available on the internet, facilitating the development of larger models in the future.

### **Potential for danger**

It can often be difficult to predict how influential new technologies will be but with LLMs there is little doubt. ChatGPT has quickly become the fastest growing application in history (Hu, 2023). So far analysis of these systems has focused on narrow assessments of their performance, i.e., how accurately they model human language. This is important for making these systems better at tasks, but more consideration needs to be given to the social dimensions in which they will operate.

Technology is often defined as the "scientific knowledge for practical purposes" (Britannica, 2018). The philosopher Martin Heidegger (1977) warned that viewing technology simply as 'a means to an end', leaves us blind to the ways in which it can shape society. This is especially true for LMs. Text data on the internet is viewed as a standing reserve from which we can develop models that generate language in a human-sounding way. Research is pursued in this area because of predictions

that this language generation will have many profitable uses by automating various everyday tasks that require processing of natural language. By viewing LMs with this approach, we fail to consider unintended effects they could have. LMs are particularly interesting because they output the same kind of data they are trained on, so these systems could be trained on their own output. This could even happen unintentionally because as of now, we have no reliable method to differentiate human text from AI generated text (Gebru et al., 2023). A potential result of this, is that the properties of their training data are slowly sealed into language as time passes (e.g., as they are primarily proficient in English, widespread usage could lead to a future where English is even more prevalent than it is now). To make these systems safe to use we must first seek to understand what constitutes safe behaviour from a language model. This cannot be done without knowing how they will interact with society. A slight bias could be negligible in one application but have serious long-term effects in another.

There has already been a case of man's suicide being attributed to his conversations with a chatbot. According to Xiang (2023), the chatbot "encouraged the man to kill himself and suggested methods to attempt suicide with very little prompting." To avoid scenarios like this, most commercially available chatbots have moderation in place that limits their ability to respond when discussing sensitive topics. However, these are far from perfect solutions. Despite being constantly monitored and improved, users are repeatedly finding relatively simple ways to circumvent the safeguards causing the models to output dangerous or biased content (Nast, 2023).

There is sufficient evidence to say that the content produced by these systems passes the Turing test (Mark, 2023), meaning that it is indistinguishable from text generated by humans. This means people are likely to believe the things said by these systems and users with malicious intent could use them to produce and disseminate misinformation or harmful rhetoric such as conspiracy theories or hate speech at a large scale.

Even if humans can easily tell whether the information has come from an AI system, there is still a high likelihood that they may believe and act on this information. This may be because AI-generated text is a relatively new phenomenon and the text produced is convincing enough to sound human. It is possible that many people believe these systems use search engines to find some of the information they talk about, because this is a phenomenon which they are more familiar with, when in fact, this is not normally the case. Systems like Bing Chat or Google Bard do make use of search engines and cite their sources, but we tend to put blind faith in decisions made by machines (Portilho, 2019), so we are unlikely to verify the validity of their statements. However, this is still a very novel technology and perhaps as society grows accustomed to these types of systems, we might become more sceptical of the statements that they make.

### **Aim and Objectives**

Many attempts to address the issue of bias in LMs have not led to conclusive results. The aim of this research is to provide a series of recommendations to future work on mitigating bias in LMs that will help to ensure it results in as much progress as possible. To achieve this goal, I will engage with the following questions. How do LMs become biased? How do these biases affect their behaviour? How can this behaviour be harmful? And why have previous strategies failed to mitigate these harms? Hopefully identifying the disconnects between existing research and these questions will highlight areas where future research can make improvements.

## **Methodology**

To investigate bias in LMs, I used multiple recent surveys that summarise the current research as my starting point (Binns, 2021), (Blodgett et al., 2020), (Sheng et al., 2021), (Mehrabi et al., 2021), (Field et al., 2021).

These surveys directed me towards the most widely cited and recent research on this topic. Additionally, these surveys agreed that definitions of bias and fairness in ML (Machine Learning) differed significantly throughout research and that the definitions proposed and strategies to measure and mitigate bias were largely disconnected from the real impacts these systems were likely to have. This is probably because these concepts are relatively new to the field of AI so there is no standard for approaching them. A call for interdisciplinary research was a common theme. As a result, I decided to supplement the information I had gathered from ML papers with ideas from two main disciplines.

First, I considered philosophy and history of technology research because currently ML is lacking research that investigates how bias from LMs will affect society and these disciplines deal with the real-world impacts of technologies.

Second, I am specifically investigating bias in LMs, so I considered psycholinguistics and sociolinguistics because these disciplines have analysed social biases, particularly in relation to language.

In addition to academic research, as developments in this field are ongoing, I have used news articles to include the latest developments.

## **General Bias in AI**

There is a plethora of examples of bias in AI systems, referred to as algorithmic bias. This is because most modern AI systems are trained using machine learning algorithms that learn from vast quantities of data. Historical biases present in society are frequently reflected in this data and then picked up by the machine learning algorithms. Even if features that represent factors inherently linked to bias (protected characteristics e.g., race, gender, etc) are filtered from the training set, the algorithms can still learn these biases through proxy variables, which when combined, or in some cases individually, can make accurate predictors for protected characteristics. For example, area codes are often good predictors of race or class (Ritov, Sun and Zhao, 2017).

## **Bias in language AI**

In 2014, Amazon's recruitment team developed and began using an NLP machine learning tool to filter the CVs of potential employee candidates (Lewis, 2018). The tool was scrapped after researchers found that its decision making was biased against female candidates. The examples of good CVs it had been trained on were comprised of past successful applicants who were majority male. From this it learned to associate masculinity with success. It did not directly use candidates' sex, as these variables were removed for privacy concerns. Instead, it was discovered that language that indicated an applicant was female (e.g., the name of an all-female school) was used to penalise candidates and language more commonly found on the CVs of male applicants (verbs such as executed or captured) was used to reward them.

Amazon's tool was eventually rid of this bias, but it was not used again, for fear that it could have learned other kinds of biases that the developers had not even considered. Amazon's system made

use of labelled data and was trained using supervised learning. Mitigating bias becomes even less straightforward when applied to unsupervised learning, the approach used for LLMs.

### **How do Language models work and where are they going?**

To understand how LMs output biased content, I will first give a brief explanation of how they work.

Early LMs made use of knowledge-based AI (Rosenfeld, 2000). Programmers working in conjunction with domain experts, such as linguists or grammarians would write algorithms that enact specific functions or apply logical rules to carry out tasks, while making use of knowledge bases constructed by humans. The behaviour of LMs created by this method is easy to interpret as each decision can be traced to an instruction, coded by a human programmer. However, where these models fell short was their ability to 'generalise' (handle multiple tasks or multiple cases of one task). Each subset of instructions is designed for a specific language task and a new set of instructions would be required for each new task as well as definitions for each new word that could be involved in the task. For example, in sentiment analysis, a model receives a section of text and outputs a score to reflect the attitude associated with the text, in simple cases this is a linear score where greater values correlate with more positive text or in more complex cases, the model will output a vector where different dimensions are associated with different emotions. Using knowledge-based AI to achieve this task would require humans to assign sentiment scores to each word that could appear in the text as well as define rules to deal with the nuances of human language e.g., how to interpret words preceded by negative modifiers such as "not happy". This method proved inefficient as it required much human labour. Advancements since then have dealt with LMs ability to generalise.

In the 1980s Statistical LMs were introduced (Rosenfeld, 2000). These systems are trained on vast amounts of text data to create a function that estimates the probability of words following a sequence of text. For example, an  $n$ -gram model uses the last  $n$  words in a sequence of text to predict what word is most likely to follow. These systems made significant improvements on knowledge-based approaches, trading accuracy at individual tasks for generality. Despite this they still suffered from several issues. The main problem that limited these systems is referred to as "the curse of dimensionality". It relates to how statistical properties change in high dimensional spaces and is a common issue in ML. In the case of LMs, it arises because as the size of the models' vocabulary grows, the training data required grows exponentially. This problem is particularly relevant to statistical LMs because each word is represented discretely, so adding a single new word means a major increase in training data is needed. Improving statistical LMs would require storage and computation along with collection and preparation of data at an exponential rate so they did not seem like a viable long-term solution.

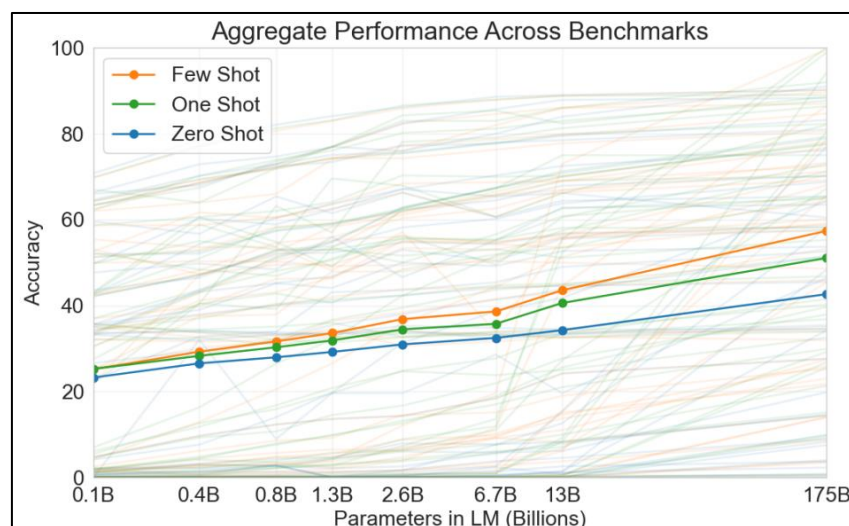
The curse of dimensionality was overcome by the introduction of Neural Networks (NNs) (Jing and Xu, 2019). The key difference lies in how they store the words in their vocabulary. Statistical LMs store each word individually with information about the probability of other words appearing after it. These models would be just as effective at predicting the continuation of a sequence of meaningless symbols, they capture no semantic information about the words. Meanwhile, with NN architectures, the words are stored as low dimensional vectors called word embeddings. Each dimension of the vector is continuous and represents some aspect of the word. This method uses the semantic relationship between the words to represent them more efficiently meaning less training data about each word is required. If the model finds that the word "cat" is used in a similar way to the word "dog", it can make inferences about how to use the word "cat" based on how "dog"

is used. This means the model can generalise very well to words, for which it has seen few examples. This is demonstrated by some interesting properties that researchers have found within word vectors. For example, adding the vector for the word “Germany” with the vector for “capital” results in a vector very close to that of the word “Berlin” (Mikolov et al., 2013).

Since their introduction, many modifications have been made to NNs to improve their performance, most notably attention mechanisms, which highlight key parts of the text that the LM is using to make decisions (Mikolov et al., 2013). This is still an active area of research and there is no consensus on what kind of NN architecture is the best. However, the fundamental elements of the architecture, especially the word embeddings which they use, are ubiquitous in modern LMs and some form of them will likely remain relevant for the foreseeable future.

Brown et al. (2020) found evidence to support the hypothesis that LMs improve at various natural language tasks as the size of the model scales.

*Figure 1 (Brown et al., 2020)*



*Figure 1 shows that as the size of the model (measured by number of parameters) increases, the model’s aggregate accuracy at a range of NLP tasks improves. The lines in the background represent the model’s performance at a range of tasks and the three in bold show aggregate performance across three environments: few shot (the model is fine-tuned on a few examples first); one shot (the model has seen one example); and zero shot (the model blindly attempts a new task).*

LMs designed in this way purely aim to predict what will follow next in a sequence of words but they have developed emergent properties that allow them to perform well at a variety of NLP tasks. The focus of NLP research over the last 40 years, from knowledge-based approaches to Statistical LMs and now to NN based LMs, has been improving their generality. This graph indicates that creating larger LMs will continue to improve their general ability at NLP tasks so a brand-new architecture might not be necessary, and research will likely continue to focus on improving the current architecture. Although, simply increasing the size of these systems is not the ideal way forward. This is because bigger models require much more computation to train, and concerns have been raised on the effect that their energy consumption has on the environment (Bender et al., 2021). In addition, statistical LMs showed a similar promise at their conception but as they increased in scale

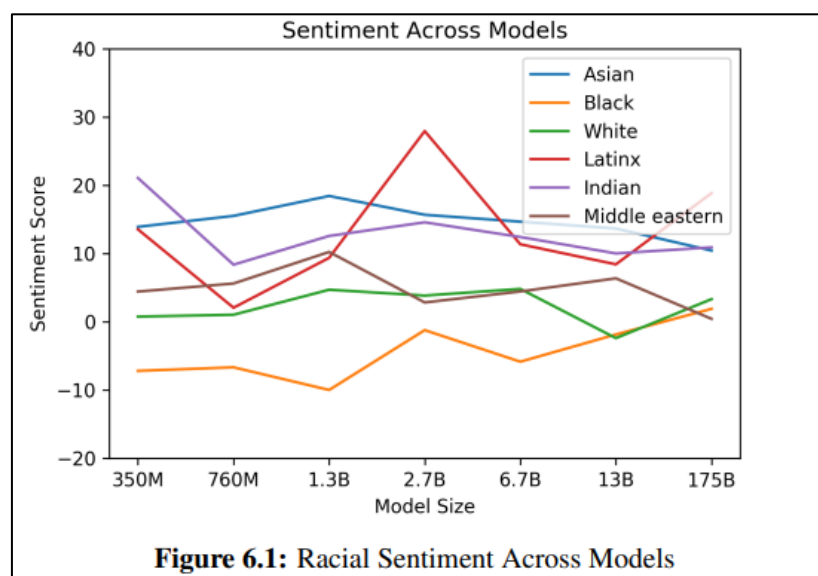
their performance improvement began to plateau (Rosenfeld, 2000) and a similar thing could eventually happen to the NN architecture.

### Why and how do word embeddings become biased?

Word embeddings require training on large volumes of text to accurately encode semantic information. Therefore, they are trained using large text datasets available on the internet, such as the common crawl dataset, which has been updated regularly since 2008 and currently contains data from 3.1 billion web pages (Nagel, 2023). Roughly 46% of this text is in English and the second most prevalent language makes up less than 6% of the data, so it is expected that any model trained on this dataset is likely to be far more proficient in English and possibly exhibit some western values or ways of thinking.

One might not expect this data to be predisposed to a particular kind of bias but, Bender et al. (2021) “found several factors which narrow Internet participation” suggesting that Internet data is over representative of the hegemonic viewpoint. This hypothesis is supported by the training of GPT-3, which, despite being trained on a curated version of the Common Crawl dataset with added data to “increase diversity” exhibits behaviour indicating racial and gender biases.

*Figure 2 (Brown et al., 2020)*



**Figure 6.1: Racial Sentiment Across Models**

*Figure 2 was generated by prompting GPT-3 to describe members of a specified race and measuring the positivity of the description using sentiment analysis (average sentiment for each race is displayed on the y-axis). This test was performed on different sized versions of GPT-3 (measured on the x axis) and while the relationship between model size and racial sentiment is unclear, the difference in sentiment across races clearly indicates racial bias.*

Similar unintended biases have been shown to exist in two of the most popular pre-trained word embeddings, Glove and Word2vec, which are widely used commercially and academically (Caliskan, Bryson and Narayanan, 2017), (Bolukbasi et al., 2016).

The biases that these embeddings develop are not random and in fact have been shown to reflect stereotypical human biases (Caliskan, Bryson and Narayanan, 2017). An example is in the task of coreference resolution (identifying which expressions refer to the same entity in a section of text) where it has been shown that NLP models struggle to associate women with stereotypically male

occupations and vice versa. This is also evident in commercial tools like Google Translate where, when translating from a gender-neutral language, the pronouns chosen for certain occupations reflect stereotypical gender biases (Prates, Avelar and Lamb, 2019). However, this is just one of the many ways in which a language model can exhibit biased behaviour. After solving the issue of occupational gender bias, other gender biases might remain and beyond gender many other forms of discrimination exist and are present in LMs.

The prevalence of bias in NLP systems suggests that careful data selection alone will not mitigate bias. The volume of data required to train a LLM means that training sets will invariably contain bias. Using human workers to filter out biased documents would be very inefficient, and humans are unlikely to be able to tell how significant of an influence on bias, a single document would have.

### **Previous work quantifying the harms of biased algorithms**

Algorithmic bias is a growing topic of debate in machine learning research and a variety of methods to debias algorithms as well as metrics to evaluate their bias levels have been proposed. There are many ways to formalise fairness in an algorithm's performance. Demographic parity is satisfied if the algorithm has the same chance to classify an outcome as positive across all demographics, while equal opportunity is satisfied if the algorithm correctly classifies positive examples at the same rate across all demographics (Mehrabi et al., 2021). Deciding which metric or combination of metrics to use is no trivial task as sometimes they can contradict each other. For example, Kleinberg, Mullainathan and Raghavan (2016) proved that three popular metrics for measuring bias could only all be satisfied at once in rare edge cases that were unlikely to arise.

A recent survey on fairness in machine learning concluded that incorporating knowledge from disciplines outside of computer science is necessary to make progress (Binns, 2021). Machine learning researchers approach fairness as a set of mathematical constraints to be met but the constraints they choose are not clearly connected to the potential negative impacts of their systems. While fairness is a relatively new concept in computer science, moral and political philosophers have grappled with notions of fairness for a long time, so leveraging some literature from these disciplines could prove useful. The survey also raises the point that what constitutes fairness is context dependent and so it is necessary to use the scenarios in which the algorithms will act as a starting point for defining fairness. This will differ across ML and is particularly challenging in the case of LMs. Objective measures of an algorithms performance in relation to a certain group cannot be applied, because the output of these systems, natural language, harbours bias in a more nuanced way and so has a much higher degree of subjectivity.

A recent survey of research on fairness, specifically in NLP systems makes this clear. It finds that research typically has "vague and inconsistent motivations" and lacks grounding in the real-world impacts of the systems (Blodgett et al., 2020). This points to a clear need for collaboration with researchers from disciplines that do engage with these real-world impacts, namely sociolinguistics and psycholinguistics in the case of NLP.



### **How will biased LMs cause harm?**

To ensure that efforts to mitigate bias are effective, we must analyse the potential ways that biased LMs can negatively impact society.

### **Who are the target audience for LMs?**

At present the text data available on the internet is predominantly in English (Richter, 2022). This means current LMs perform the best in English, they struggle with languages with very different grammatical structures such as Japanese (Shimizu, 2022), and they are unable to communicate in languages which are not widely spoken or from which there is very little text data. Research is being done to further the multilingual capabilities of these systems (Ritchie et al., 2022) but much more is needed before the benefits of these systems can be evenly distributed (Joshi et al., 2020). Even within English, accessibility to these models is not evenly distributed. The text outputted by LMs is best classified as Standard American English (Groenwold et al., 2020). Groenwold et al. (2020) have shown that GPT-2 struggles to understand prompts written in African American Vernacular English and is more likely to associate a negative sentiment with them. Bias against dialects, referred to as linguistic discrimination is a well-documented phenomenon in linguistics and has been shown to restrict access in several areas of everyday life including education, employment, and healthcare (Terry et al., 2010), (Baker-Bell, 2019), (Grogger, 2011), (Leech, Irby-Shasanmi and Mitchell, 2018), (Kugelmass, 2016). By automating the roles that perpetuate this bias, we risk exacerbating this form of discrimination.

As LMs pervade industries worldwide, the advantages they create will be less accessible to those who are less proficient in the dialects they use. In 2007, Woodhouse and Sarewitz (2007) argued that “New technoscientific capacities introduced into a non-egalitarian civilization will tend disproportionately to benefit the affluent and powerful” and found much evidence to support this statement. For example, despite an overall decline in death rates in the United States since 1960, the difference in death rates between affluent and poorer citizens increased between 1960 and 1986. The medical advances made over this period, which lead to the decline in death rates, were designed for the health problems of concern to the wealthy, rather than all of society. LMs are designed to serve the hegemonic group that have historically contributed the most to their training data and in doing so, like many other new technologies, will disproportionately benefit this group.

Accessibility to LMs will not be evenly distributed and this is not an issue that can be solved by standard, task-agnostic definitions of fairness. The behaviour of LMs needs to be examined in the context of individual use cases to know how certain groups will be disadvantaged once these technologies are widely implemented. This could range from: gatekeeping entry to positions that require the use of LMs (e.g., if they are used to write financial reports); to making it more difficult to use LMs for everyday tasks (e.g., if they are used to order products online).

### **Representational Harms**

The main concern that stems from bias in LMs is representational harms. Representational harms occur when a group is portrayed more negatively than others or its existence is not acknowledged at all. When a social group is portrayed with a false or misleading generalization, this is referred to as a stereotype. It is easy to see how negative representation can harm a social group (e.g., subconsciously affecting the decision of an employer) but moral philosophers argue that seemingly positive stereotypes are also harmful because they make social groups seem more alien as well as stripping away the individuality of that group’s members in the perspective of others (Blum, 2004).

In social psychology, “Language is considered as the major means by which stereotypes are communicated” (Maass, 1999). From private discussions to mass media, language is the conduit through which stereotypes pass on to future generations. This can occur in explicit ways, when a group is overtly described as having a certain property, but also in subtle ways that are much harder to detect. In 1989, a group of social psychology researchers developed a framework for measuring “intergroup bias” by the language used to describe behaviours (Maass, 1999). After running experiments with the LIB (Linguistic Intergroup Bias) framework for a decade, they found evidence to support the hypothesis that, when describing an action of a member of a certain social group, the abstractness of the description reflects the speaker’s bias towards that group. When describing the behaviour of members of a social group, behaviour that matches the stereotypes about that group held by the speaker, will be described in relatively abstract terms implying that it reflects the general behaviour of that group, while behaviour that does not match the held stereotypes will be described concretely implying that it is an anomalous instance of such behaviour.

For example, Gorham (2006) conducted a study where participants were shown identical news stories about a murder investigation except in half of them the suspect was described as a black male and in the other half, as a white male. Only the results of white participants were collected, and the study found that they used more abstract terms to describe the event when the suspect was black (e.g., he is probably violent), and more concrete terms when the suspect was white (e.g., he probably hurt the victims). The researchers concluded that the results indicated an unconscious racial bias in their participants.

Although subtle, these biases are important because they contribute to a cycle that establishes stereotypes and shields them from critique (Maass, 1999). Once in place, these biases are difficult to undo. This body of research shows that, in addition to overt bias, subtle properties of language are worth considering when analysing bias in LMs because there is strong evidence linking them to bias and they have the potential to solidify stereotypical beliefs in society.

### **Long-term implications**

At present, existing LMs are biased artefacts. Technologists have examined the political attributes of artefacts and found that they can support certain distributions of power in society and undermine others (Winner, 1980). For example, it has been argued that the close control and safety measures required to operate nuclear power stations make them an inherently totalitarian technology, meanwhile solar power inherently lends itself to distributed power generation and storage, making it more democratic.

An architect intentionally designed bridges in New York to be low enough that buses could not pass under them, but cars could. By doing this he restricted poorer minority groups, who typically could not afford cars and were dependent on public transport, from accessing certain areas of the city. After his death, his architectural designs, now embedded in the foundation of the city, continued to affect the public (Winner, 1980). This is an example of how a subtle feature can contribute to long term bias. The deployment and usage of LMs that output biased text will have a similar effect. As automation becomes more commonplace, certain key services will become less accessible to minority groups and AI generated content will gradually shape public opinion of them in a negative way.

Technology is often defined as the “application of scientific knowledge for practical purposes” (Britannica, 2018). The philosopher Martin Heidegger (1977) warned that viewing technology simply as ‘a means to an end’, leaves us blind to the ways in which it can shape society. This is especially

true for LMs. Text data on the internet is viewed as a standing reserve from which we can develop models that generate language in a human-sounding way. Research is pursued in this area because of predictions that this language generation will create profit by automating various everyday tasks that require processing of natural language. By viewing LMs with this approach, we fail to consider potential unintended. LMs are particularly interesting because they output the same kind of data they are trained on, so these systems could be trained on their own output. This could even happen unintentionally because as of now, we have no reliable method to differentiate human text from AI generated text (Gebru et al., 2023). A potential result of this, is that the properties of their training data are slowly sealed into language as time passes (e.g., as they are primarily proficient in English, widespread usage could lead to a future where English is even more prevalent than it is now).

To understand how to evaluate the performance of these systems first we must better understand their place in society. The need for this is made evident by the limitations of previous attempts at mitigating bias.

### **Methods of debiasing word embeddings**

#### **Geometric Manipulation**

One of the first methods proposed to tackle bias in LMs was to alter the position of vectors in the word embedding space to remove unwanted associations (Bolukbasi et al., 2016a). Each dimension in the vector space represents some property that a word could have. The researchers hypothesised that some combination of dimensions represents the direction of gender and identified this by averaging the difference of pairs of gendered words (e.g., 'he' – 'she', 'man' – 'woman' etc). Next, they set all the genderless words (as opposed to words like actress which inherently refer to one gender and so are gendered) to be zero in the direction of gender. This meant all genderless words were now equidistant from pairs of gendered words. The researchers tested the results of their method by prompting the model, trained on these now debiased word embeddings, to generate gender analogies, asking questions such as "He is to Doctor as she is to X?". Before debiasing the word embeddings, the model would output "Nurse", demonstrating a clear gender bias but after debiasing, the output was "Physician". This may still be demonstrating a form of bias, but overall crowdsourced reviewers found the models outputs to be significantly less reflective of gender stereotypes after debiasing.

Subsequent research found this method was largely ineffective at removing gender bias from the model. In the task of generating analogies, it performed well but Gonen and Goldberg (2019) found that genderless words that reflected stereotypes were still proximal to each other in the vector space and that this was sufficient for machine learning algorithms to exhibit gender bias from these vectors. So, although the vectors for the words "nurse" and "receptionist" were moved sufficiently far away enough from the vector for "she" to remove their association, the two vectors were still close to each other. Bolukbasi et al. (2016a) failed to detect this because the bias metric that they used (a method of evaluating how biased the output of a language model is) only captured a narrow part of the spectrum of ways language can display gender bias. Other researchers have proposed similar methods since then (Zhao et al., 2018). Despite making improvements they fail at evaluating their models in sufficient ways to confirm they are free of bias. Gonen and Goldberg (2019) describe these attempts as "mostly superficial", and state that "The bias... is ingrained much more deeply in the embeddings space".

In such approaches, although not majorly impacted, the extent to which the debiasing technique affects the "desirable properties" of the word embedding is a concern (Bolukbasi et al., 2016a).

Desirable properties refer to any semantic information in the vector outside of bias. This is important because the system relies on these vectors to know the meaning of the words. In attempts to neutralise bias, one might lose important properties for example the difference between the words: “man” and “woman”. Blum (2004) notes that “In criticizing stereotypes, one should not fall into the trap of denying often regrettable but sound comparative statistics because these are vital to measuring the social wellbeing of certain groups” which is a possible consequence of such techniques. Therefore, I think it is unwise to continue researching geometric approaches because any manipulation of the vector space that is sufficient to remove bias would likely interfere with the vectors’ desirable properties.

### **Dictionary-based debiasing**

An alternative method was proposed where a decoder is used to extract the key semantic information from a word embedding and then an encoder is used to recreate it such that its meaning is similar to a given dictionary definition of the word and the vector is orthogonal to any subspaces that represent a biased direction (Kaneko and Bollegala, 2021). Here, orthogonality is intended to be analogous to neutrality. The technique proposed can be applied to any set of word embeddings without access to the training data because it leverages the words in the intersection of the model’s vocabulary and the words used in the dictionary definition. A minor limitation of this method is that it relies on the existence of dictionary definitions for each word in the corpus. Although, dictionaries may have to be manually reviewed because, in rare cases, definitions can amplify the bias of words. Kaneko and Bollegala (2021) note that the definition for “homemaker” in their dictionary was “a wife who manages a household while her husband earns the family income” and this is almost certainly responsible for its outlier status after debiasing.

The method proved effective at debiasing the vectors for four popular natural language bias evaluation metrics, however this still suffers from the same issue as the geometric approach because there is little evidence to suggest that satisfying these metrics will result in no biased outputs further downstream. In geometric approaches, retaining semantic properties is a concern but an unintended consequence of this approach is that performance on natural language tasks improved after debiasing. The researchers speculate that the information learned from the dictionary not only helped to reduce bias but also improved the word embeddings representativeness of the words that they code for. This result means that dictionary-based debiasing is a worthwhile avenue of future research because the intactness of the vectors’ desirable properties is not a concern.

### **Data Augmentation**

A series of pre-processing debiasing approaches have been proposed. This means instead of operating on the vectors after or during their learning, these methods augment the training data before the learning takes place. Basta and Costa-jussà (2023) showed that the ratio of male to female entities referred to in the training data can have a significant impact on the bias of the model. Zhao et al. (2019) demonstrated this concept by applying two similar techniques to the training data. Both methods effectively create a duplicate training set in which all gendered entities are swapped to the opposite gender, and then train the word embeddings on the union of these two sets. This method proved effective at satisfying popular bias metrics.

Another example of how augmenting the training data can contribute to debiasing vectors was proposed by Brunet et al. (2019). Certain documents in a training data set contribute to bias more than others, but to empirically measure this would be highly costly in terms of computation. Brunet

et al. (2019) developed a method to approximate how changes to the training data would influence the bias of the model. They tested this empirically and found it to be effective for the given bias metrics. The main advantage of this method is that it can be used to identify small subsections of the training data that have a disproportionate impact on bias, and this can be used to curate the training data instead of relying on human intuition. Qualitative analysis revealed that the most bias increasing documents were typically about scientific work conducted by men, e.g., the article “60 New Members Elected to Academy of Sciences” which listed awards given to predominantly male scientists. This is consistent with the conclusion drawn from the other data augmentation approach, namely that the difference in number of mentions of a particular social group plays a significant role in the bias of the word embeddings. However, the strong influence of articles specifically about science is likely because the bias metric used, WEAT 1, is based on the stereotype that men are more often associated with science and women with art (Caliskan, Bryson and Narayanan, 2017).

Data augmentation methods are promising in that they do not alter already trained word embeddings so there are no concerns of removing useful semantic properties, and of the methods proposed they require the least human intervention so the probability of human biases affecting the results is lowest. These approaches face the same limitation however, in that their performance is measured using metrics with no clear connection to real world use cases of LMs and that do not comprehensively measure all the ways in which bias can persist through language.

A clear insight from these methods is that representation in the dataset has a large influence on bias. This is the case in other branches of ML, for example in image classification where underrepresentation of a demographic in the training data has a significant impact on error rate for classification of that demographic (Buolamwini et al., 2018). However, in this case, and others that are similar, the bias is the result of an ML system underperforming in classifying examples with which it is unfamiliar resulting in allocational harms (e.g., minorities being wrongly identified by facial recognition systems). In LMs, this not only seems to contribute to underperformance (e.g., a lower likelihood of mentioning members of that group) but also contributes to negative representations of these groups. It is not obvious why this is the case and is a potential area for future research. I expect that it is related to the fact that the number of operations performed on a word’s vector is proportional to its frequency in the training data. Fewer operations could result in the vector being associated with fewer words meaning a stronger connection to the words it is already associated with, in the case of minority groups, are more likely to relate to bias. i.e., less representation means a higher weighting to negative representations.

### **Bias mitigation summary**

Throughout the various papers that propose these approaches, it is common for researchers to use their methods in combination with others that have been proposed which often leads to better results. The effectiveness of these methods is still in question because there is no conclusive evidence showing that satisfying the proposed bias metrics prevents biased behaviour. From this work we can see that before we start comparing methods of debiasing LMs, a system of evaluating the success of these methods which accounts for all possible biased behaviours of the model is necessary.

## Bias metrics

### Intrinsic Metrics

Intrinsic bias metrics measure bias based on the position of the vectors in the word embeddings space.

*Figure 3 (Goldfarb-Tarrant et al., 2021)*



*Figure 3 visualises a set of word embeddings in 2 dimensions. Each point is a word and the distance from one word to another represents semantic nearness.*

WEAT (Word Embedding Association Test (Caliskan, Bryson and Narayanan, 2017)) is a set of popular intrinsic bias metrics based on the IAT (Implicit Association Test) first proposed by Greenwald, McGhee and Schwartz (1998). The IAT is a popular framework for measuring bias in humans and has demonstrated successful results in the field of social psychology. It measures reaction times in a person's ability to associate positive words with different demographics. The WEAT metrics try to apply this concept to word embeddings under the assumption that distance in the word embedding space (a measure of "semantic nearness") is "analogous to reaction time" in humans (Caliskan, Bryson and Narayanan, 2017)). This is probably the core fault of WEAT because there is no evidence to support this assumption.

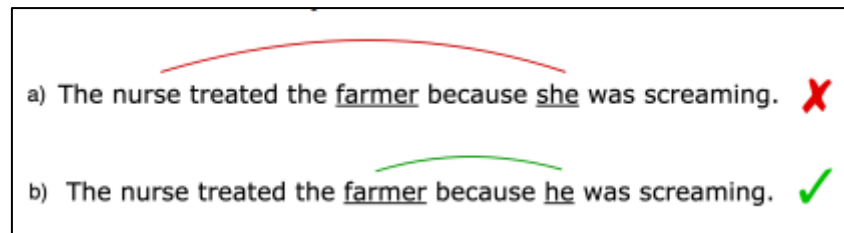
Goldfarb-Tarrant et al. (2021) showed that there is no correlation between WEAT metrics and the bias in the outputs of LMs and extended this to intrinsic bias metrics in general. The logic behind the creation of WEAT, applying a framework from disciplines with bias measures that have proven successful, is sound but the lack of connection between distance in a vector space and reaction time effectively renders the metric useless.

Intrinsic bias measures do not present a promising area of future research because the structure of the vector space and its connection to the behaviour of LMs is too poorly understood. Advances in explainability, particularly in unsupervised learning, may build credence in this direction of research, but until then extrinsic measures should be the focus of research on fairness in LMs.

## Extrinsic Metrics

Extrinsic bias metrics measure bias in the models' external behaviour i.e., the text that it outputs.

Figure 4 (Goldfarb-Tarrant et al., 2021)



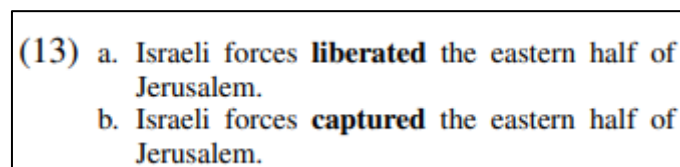
*An example of a model exhibiting gender bias: based on the context it is evident that the farmer is the one screaming. The model correctly identifies this when the farmer is male but fails to when female because it struggles to associate female entities with farmers.*

These metrics typically measure properties of the text and apply standard notions of fairness in ML. For example, to show a lack of correlation between intrinsic debiasing and model behaviour, Goldfarb-Tarrant et al. (2021) measured the rate of false negatives in coreference resolution. The model was tasked with identifying which entities pronouns referred to and false negatives was chosen as the metric because this seemed to be the most harmful way it could represent minority groups i.e., failing to associate minority groups with positions of power.

Other extrinsic bias metrics leverage earlier NLP techniques such as textual sentiment analysis or hate speech detection which have been used since before LMs became prominent (Sheng et al., 2021). However, Hate Speech detection has been known to have problematic consequences, for example it often identifies the usage of reclaimed slurs as hateful which leads to the censoring of minority demographics in online spaces (Bender et al., 2021). Sentiment analysis is similar in that it is often more likely to associate a negative sentiment with minority groups and, when measuring emotions, it can reflect stereotypical views (Kiritchenko and Mohammad, 2018). Therefore, using these techniques in evaluating bias could lead to an increase in bias.

The design of these metrics may be such that they are more suited to capturing overt biases, but subtle biases might slip through undetected. Recasens, Danescu-Niculescu-Mizil and Jurafsky (2013) proposed training to evaluate bias in natural language using a logistic regression model trained on NPOV (Neutral Point of View) Wikipedia edits. Wikipedia (2020) operate under an NPOV policy which guides editors to use an objective standpoint.

Figure 5 (Wikipedia, 2020)



*An edit where the word "liberated" was changed to "captured" to give the description a neutral point of view.*

By training a model on examples of edits that have been made to Wikipedia articles with NPOV as the reason, Recasens, Danescu-Niculescu-Mizil and Jurafsky (2013) created a system that detects subtle subjective biases in language. A similar method could be used to design evaluation systems

for LMs. However, as a model of this kind is trained on man-made examples, it could be susceptible to human biases e.g., any biases commonly held by Wikipedia editors.

While these metrics provide a good start to detecting bias extrinsically, none provide a comprehensive measure of the broad ways language can express bias, but it is possible that a combination of them could achieve this. Researchers have been able to produce better results by combining various methods. Some development has already begun on methods to easily implement multiple bias metrics simultaneously (Czarnowska, Vyas and Shah, 2021). If an effective way to measure bias is found, it will likely be a result of satisfying multiple metrics at once. Which combination is chosen will likely be highly dependent on the use case of the LM because satisfying too many at the same time might be infeasible.

### **Is there a solution?**

Algorithmic bias is more of a social issue than an engineering challenge. A long historical context has shaped the training data that influence these algorithms, and they will operate in a complex environment where social, political, and economic factors influence and are influenced by their actions. In addition, the arena of natural language is particularly hard to navigate because of the nuance and versatility in the way we communicate. Words have differing definitions based on their context and no dictionary definition captures all the connotations a word might have. The scale at which these systems operate makes it difficult for humans to alter their behaviour, datasets are too large to read through and they generate language too quickly for us to monitor them. But designing algorithms to evaluate and mitigate their biases, technological fixes, are not promising avenues either.

After exploring many successful and unsuccessful attempts that use technology to fix social problems, Sarewitz and Nelson (2008) proposed three criteria necessary for a technological fix to work. The first criterion for a technological fix requires that the fix directly link to the cause-effect relationship creating the problem. Collectively, we have a good understanding of how bias exists in society and how it could potentially cause harm, but machine learning passes these biases through an uninterpretable blackbox and they manifest in different forms as they come out the other side. The effects of a biased LM have not been fully explored and what causes LMs to be biased is only understood in a limited capacity. The second criterion is that the effectiveness of the fix can be measured in an unambiguous manner. As we have seen, in measuring the bias of ML systems this could not be farther from the case, not only are there a range of different metrics but some even contradict each other. Finally, the knowledge being applied to fix the problem should be embodied in the technology itself. It should not rely on training individuals and organisations to apply the required knowledge. Vaccinations are a good example of this. Those administering the vaccine need not understand how it works, once it is injected the scientific knowledge is applied by the vaccine itself. Meanwhile, measures of bias in LMs will likely be context dependent. Without a clear agreed upon framework of how to evaluate bias in any given context, those implementing the model in a specific use case, and the regulators evaluating the safety of this implementation, will have to understand how the bias metric links to the real-world effects of the system.

The lesson from the rules for success of technological fixes is a clear path of research that needs to be completed before a debiasing strategy can be developed. If the cause-effect relationship is better understood, a bias metric can be developed for a specific context without debate about its aptness and a framework can be developed to assist those using and governing these systems.



## **What to consider going forward**

In this essay, I have provided a review of the ways in which bias in LMs can cause harm and explained why existing bias mitigation strategies fail to prevent this. I shall now provide a series of recommendations to consider in the development of future strategies.

### **Ghost Work**

Ghost work is a term used to refer to the use of human labour in the development of AI systems that generally goes unnoticed (Gray, 2019). It typically involves repetitive online tasks such as labelling data and is often used in the development of LMs for example in evaluating the accuracy of bias metrics. Some have described this practice as unethical because of the general poor treatment of the workers involved and, in some cases, workers can be traumatised by the content that they must view, for example if they must filter out graphic or violent media (Perrigo, 2023).

If ways are found to do it that treat the workers in a more ethical way, it still runs the risk of further incorporating human bias into the systems. For example, if a non-diverse set of human labellers is used, their “societal norms and linguistic variations” may “influence the evaluation standards for generated text” (Sheng et al., 2021). Human generated data is always prone to bias, (take the biased dictionary definition for homemaker as an example) and so for these reasons it is best to avoid the use of ghost work.

### **Intersectionality**

Although much of NLP bias research has focused on gender, developers are starting to extend their methods to other forms of bias in society (Field et al., 2021).

However little consideration has been given to intersectionality, the phenomenon where multiple forms of bias combine leading to different forms of discrimination. This has been shown to affect machine learning in other fields for example in image processing (Buolamwini et al., 2018), and research is beginning to show that our current techniques for evaluating bias in language are insufficient to capture intersectional biases (Tan and Celis, 2019).

Overall, the research conducted on intersectional biases has been insufficient (Devinney, Björklund and Björklund, 2022), so there is more work to be done in this area. For example, if mitigating one form of discrimination exacerbates another.

### **Static Data**

Bender et al. (2021) have noted that while documents that form the training data of LLMs are static in time, societal views change. Older documents from less politically correct time periods could be responsible for some of the severe cases of bias we see now. As time passes, views will continue to change and bias metrics must be regularly updated to reflect this.

Technological lock-in is the theory that once a technology is widely implemented, it is difficult to alter it regardless of its flaws (Cowan, 1990). There are several strong examples that support this idea, and a similar problem may occur with LMs. Any views that LMs hold could be embedded into society so it is important that their views change to reflect ours if we become locked into this type of technology. Any policies that regulate the bias of LMs will need to be updated regularly to reflect this. A potential solution could be to limit the earliest date from which documents can be used in training of LMs. For example, only documents from the last 10 years. However, modern documents can contain biases and older documents could be bias free. Additionally, these views differ widely across the world and even in different parts of the same country.

The Overton Window is a term that describes acceptable social views at a given point in time. Kahmann and Heyer (2019) conducted a study to measure changes in the Overton Window by comparing text on the internet considering the date it was uploaded. Ideas from this approach could be used to adjust bias metrics over time or in the filtration of datasets and ensure that the views expressed by LMs are confined towards the middle of the Overton Window (the section which is agreed upon by the most people). In the paper that proposed this method, the researchers observed the shift in a German newspaper over the span of roughly a decade. Restricting the data used in this method to a set of reputable sources such as newspapers would likely be a necessity. As we have discussed, internet access is not evenly distributed so an Overton Window constructed from unfiltered data will not be representative of society's views.

### **Interdisciplinarity**

Bias and fairness are concepts that have been extensively studied by other academic disciplines, such as psychology and philosophy. They are relatively new concepts in ML which often leads to key terminology surrounding these concepts being loosely defined resulting in researchers and developers taking them to mean different things. Knowledge from these disciplines must be used to support ML research into bias and fairness.

The IAT (Greenwald, McGhee and Schwartz, 1998), research from social psychology, being the inspiration for the WEAT metrics is a good sign, but that particular test and its implementation on word embeddings proved ineffective. I would propose investigating similar metrics using other frameworks such as the LIB (Maass, 1999). The LIB is like the IAT in that it can reveal human biases at a subconscious level, but it differs in that it is designed to measure this through language used by humans.

The ways in which the LIB framework can be implemented are broad. In addition to proving successful across a variety of languages, the researchers noticed that it can be detected through a range of parts of speech including verbs (and tense of verb), nouns, adjectives, and adverbs. This indicates that it is applicable to language in general and does not require specific types of language. They also used it to find evidence of bias in "journalistic language" (Maass, 1999). A bias metric that implements this framework could be designed using knowledge from linguistics experts or in a similar way to in (Recasens, Danescu-Niculescu-Mizil and Jurafsky, 2013), where a model is trained to detect it using labelled data.

Therefore, I think it would serve as a good base for an extrinsic bias metric. This is just one example of the many ways in which other disciplines could contribute to the foundation of mitigating bias in NLP systems.

### **Conclusion**

Language models face a serious issue in terms of bias and a better understanding of how these systems work is needed before they can be implemented safely. Flaws in existing research can mostly be explained by inadequate ways of measuring bias. This indicates that future research should focus on studying the ways these systems are likely to cause harm in the real world before attempts at debiasing them are made. Doing this will require knowledge and collaboration with experts from disciplines that are more familiar with these problems.

Woodhouse and Sarewitz (2007) found that new innovations tend to be less beneficial to minority groups and LMs are on course to follow this pattern. Profit driven companies are currently leading AI

research, and academic research relies on funding, typically from industries seeking to profit from their insights (Brennen et al., 2019). The power of market forces has had several worrying consequences. One is that research is geared towards meeting the needs of the wealthier in society, from who more money can be made. Progress is focused on business applications and little thought has been given to these systems' potential as educational tools, and other applications aimed at increasing equity in society. Another consequence is that research in fairness and responsible use of AI are much lower priorities as they do not make systems more profitable. In fact, these systems are being deployed before they are shown to be safe, because generating media hype about their capabilities is needed to secure investments (De Vynck, Lerman and Tiku, 2023).

These systems are in use now, and the effects they will have on society are unknown. The public is effectively subject to an experiment and van de Poel (2015) proposed a framework to evaluate the ethics of new technologies as if they were an experiment. By his standard, this is an unethical experiment because of the lack of thought given to the harms and risks LMs could pose. One requirement is that "The experiment is approved by democratically legitimized bodies." While countries are beginning to design policy to govern the use of AI, these policies are lagging due to the speed at which progress is being made in this field.

Future of Life Institute (2023) published an open letter asking the AI community to pause development of AI systems larger than GPT-4 and many notable experts in AI have co-signed this letter. The motivations for this letter are questionable. The main concern discussed is the dangers of AGI and as a result it may have an effect opposite to the one intended because it generates further hype about the capabilities of these systems and may lead to more investment and faster training of larger systems. Alternatively, the research during the pause might focus on the long-term threat of AGI, and therefore neglect near-term threats of bias and misuse. However, the idea of pausing AI research is not without merit. Woodhouse and Sarewitz (2007) proposed slowing down research as a potential way to reduce inequalities that new technologies bring. Although, garnering support from all the major organisations around the world involved in AI development seems challenging, history has shown that humanity can reach a common ground on these kinds of topics. For example, the 1975 Asilomar Conference where scientists and regulators successfully reached an agreement to avoid the risks of gene-editing while still being able to make use of its benefits (Jasanoff and Kim, 2015). A pause on development of LLMs could be a way to let the necessary research and regulation catch up and ensure the safety of these systems.

One of the reasons to justify experimental use of new technologies, proposed by van de Poel (2015) is a lack of an alternative way to learn about their effects. This could be an important reason for continued use of LLMs. Worthington and Collingridge (1982) argued that despite the dangers posed by emerging technologies whose societal effects are poorly understood, usage is one of the best ways to learn more about these effects so that we can regulate them effectively. A middle ground could be found for LLMs. A pause on models above a certain size is enacted, and research is then dedicated to studying the effects of their use at a smaller scale and testing potential solutions to bias among other concerns. This would be one of the most effective ways to assess their potential societal ramifications and in doing so develop effective strategies to mitigate issues like bias.

# References

- Alayrac, J.-B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S. and Zisserman, A. (2020). Self-Supervised MultiModal Versatile Networks. *arXiv:2006.16228 [cs]*. [online] Available at: <https://arxiv.org/abs/2006.16228>.
- Baker-Bell, A. (2019). DISMANTLING ANTI-BLACK LINGUISTIC RACISM IN ENGLISH LANGUAGE ARTS CLASSROOMS: Towards an Anti-Racist Black Language Pedagogy. *Theory Into Practice*, 59(1). doi:<https://doi.org/10.1080/00405841.2019.1665415>.
- Basta, C. and Costa-jussà, M.R. (2023). Impact of Gender Debaised Word Embeddings in Language Modeling. *Computational Linguistics and Intelligent Text Processing*, pp.342–350. doi:[https://doi.org/10.1007/978-3-031-24337-0\\_25](https://doi.org/10.1007/978-3-031-24337-0_25).
- Bender, E., Mcmillan-Major, A., Shmitchell, S., Gebru, T. and Shmitchell, S.-G. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Timnit Gebru \* [timnit@blackinai.org](mailto:timnit@blackinai.org) Black in AI Palo Alto, CA, USA CCS CONCEPTS • Computing Methodologies → Natural Language processing. ACM Reference Format. [online] doi:<https://doi.org/10.1145/3442188.3445922>.
- Binns, R. (2021). Fairness in Machine Learning: Lessons from Political Philosophy. *arXiv:1712.03586 [cs]*, [online] 81(3). Available at: <https://arxiv.org/abs/1712.03586>.
- Blodgett, S.L., Barocas, S., Daumé III, H. and Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of ‘Bias’ in NLP. *arXiv:2005.14050 [cs]*. [online] Available at: <https://arxiv.org/abs/2005.14050>.
- Blum, L. (2004). Stereotypes And Stereotyping: A Moral Analysis. *Philosophical Papers*, [online] 33(3), pp.251–289. doi:<https://doi.org/10.1080/05568640409485143>.
- Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. and Kalai, A. (2016a). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv:1607.06520 [cs, stat]*. [online] Available at: <https://arxiv.org/abs/1607.06520>.
- Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. and Kalai, A. (2016b). Quantifying and Reducing Stereotypes in Word Embeddings. *arXiv:1606.06121 [cs, stat]*. [online] Available at: <https://arxiv.org/abs/1606.06121>.

Brennen, J., Schulz, A., Howard, P. and Nielsen, R. (2019). Industry, Experts, or Industry Experts? Academic Sourcing in News Coverage of AI. *ora.ox.ac.uk*. [online] Available at: <https://ora.ox.ac.uk/objects/uuid:c25bdf8c-9c47-453b-bb9f-8e1a06dbb4ca> [Accessed 23 Apr. 2023].

Britannica (2018). technology | Definition & Examples. In: *Encyclopædia Britannica*. [online] Available at: <https://www.britannica.com/technology/technology>.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C. and Hesse, C. (2020). Language Models are Few-Shot Learners. *arxiv.org*. [online] Available at: <https://arxiv.org/abs/2005.14165>.

Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A. and Zemel, R. (2019). Understanding the Origins of Bias in Word Embeddings. *arXiv:1810.03611 [cs, stat]*. [online] Available at: <https://arxiv.org/abs/1810.03611> [Accessed 22 May 2020].

Buolamwini, J., Gebru, T., Friedler, S. and Wilson, C. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification \*. *Proceedings of Machine Learning Research*, [online] 81, pp.1–15. Available at: <https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>.

Caliskan, A., Bryson, J.J. and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, [online] 356(6334), pp.183–186. doi:<https://doi.org/10.1126/science.aal4230>.

Chen, A. (2019). *How Silicon Valley's successes are fueled by an underclass of 'ghost workers'*. [online] The Verge. Available at: <https://www.theverge.com/2019/5/13/18563284/mary-gray-ghost-work-microwork-labor-silicon-valley-automation-employment-interview>.

Cowan, R. (1990). Nuclear Power Reactors: A Study in Technological Lock-in. *The Journal of Economic History*, [online] 50(3), pp.541–567. Available at: [https://www.jstor.org/stable/2122817?saml\\_data=eyJzYW1sVG9rZW4iOiI4ZjFhYTlkMC01MWNhLTQ1YWYtYTkwNi05NDkxYjYxYzk3NDMiLCJpbmN0aXR1dGlvbklkcyI6WyIxOGVIZTJmYS1mODcxLTQwYTktODI4NS1mNTRlYzdhMDM4MjciXX0&seq=1](https://www.jstor.org/stable/2122817?saml_data=eyJzYW1sVG9rZW4iOiI4ZjFhYTlkMC01MWNhLTQ1YWYtYTkwNi05NDkxYjYxYzk3NDMiLCJpbmN0aXR1dGlvbklkcyI6WyIxOGVIZTJmYS1mODcxLTQwYTktODI4NS1mNTRlYzdhMDM4MjciXX0&seq=1) [Accessed 18 Apr. 2023].

Czarnowska, P., Vyas, Y. and Shah, K. (2021). Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics*, 9, pp.1249–1267. doi:[https://doi.org/10.1162/tacl\\_a\\_00425](https://doi.org/10.1162/tacl_a_00425).

De Vynck, G., Lerman, R. and Tiku, N. (2023). Microsoft's AI chatbot is going off the rails. *Washington Post*. [online] 16 Feb. Available at: <https://www.washingtonpost.com/technology/2023/02/16/microsoft-bing-ai-chatbot-sydney/>.

Devinney, H., Björklund, J. and Björklund, H. (2022). Theories of 'Gender' in NLP Bias Research. *arXiv:2205.02526 [cs]*. [online] Available at: <https://arxiv.org/abs/2205.02526> [Accessed 25 Mar. 2023].

Eloundou, T., Manning, S., Mishkin, P. and Rock, D. (2023). *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*. [online] Available at: <https://arxiv.org/pdf/2303.10130.pdf>.

Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., Lu, H., Song, R., Gao, X., Xiang, T., Sun, H. and Wen, J.-R. (2022). Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1). doi:<https://doi.org/10.1038/s41467-022-30761-2>.

Field, A., Blodgett, S.L., Waseem, Z. and Tsvetkov, Y. (2021). A Survey of Race, Racism, and Anti-Racism in NLP. *arXiv:2106.11410 [cs]*. [online] Available at: <https://arxiv.org/abs/2106.11410>.

Future of Life Institute (2023). *Pause Giant AI Experiments: An Open Letter*. [online] Future of Life Institute. Available at: <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.

Geburu, T., Bender, E.M., McMillan-Major, A. and Mitchell, M. (2023). *The DAIR Institute*. [online] [www.dair-institute.org](http://www.dair-institute.org). Available at: <https://www.dair-institute.org/blog/letter-statement-March2023>.

Goldfarb-Tarrant, S., Marchant, R., Sanchez, R.M., Pandya, M. and Lopez, A. (2021). Intrinsic Bias Metrics Do Not Correlate with Application Bias. *arXiv:2012.15859 [cs]*. [online] Available at: <https://arxiv.org/abs/2012.15859> [Accessed 18 Apr. 2023].

Gonen, H. and Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *arXiv:1903.03862 [cs]*. [online] Available at: <https://arxiv.org/abs/1903.03862>.


Gorham, B.W. (2006). News Media's Relationship with Stereotyping: The Linguistic Intergroup Bias in Response to Crime News. *Journal of Communication*, 56(2), pp.289–308. doi:<https://doi.org/10.1111/j.1460-2466.2006.00020.x>.

Gray, M. (2019). *GHOST WORK : how amazon, google, and uber are creating a new global underclass*. Harper Business.

Greenwald, A.G., McGhee, D.E. and Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6), pp.1464–1480. doi:<https://doi.org/10.1037//0022-3514.74.6.1464>.

Groenwold, S., Ou, L., Parekh, A., Honnavalli, S., Levy, S., Mirza, D. and Wang, W. (2020). *Investigating African-American Vernacular English in Transformer-Based Text Generation*. [online] Association for Computational Linguistics, pp.5877–5883. Available at: <https://aclanthology.org/2020.emnlp-main.473.pdf>.

Grogger, J. (2011). Speech Patterns and Racial Wage Inequality. *The Journal of Human Resources*, [online] 46(1), pp.1–25. Available at: <https://www.jstor.org/stable/25764802>.

Heidegger, M. (1977). *The Question Concerning Technology*  and Other Essays X. [online] Available at: [https://monoskop.org/images/4/44/Heidegger\\_Martin\\_The\\_Question\\_Concerning\\_Technology\\_and\\_Other\\_Essays.pdf](https://monoskop.org/images/4/44/Heidegger_Martin_The_Question_Concerning_Technology_and_Other_Essays.pdf).

Hu, K. (2023). ChatGPT sets record for fastest-growing user base - analyst note. *Reuters*. [online] 2 Feb. Available at: <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>.

Jasanoff, S. and Kim, S.-H. (2015). *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*. [online] *Google Books*. University of Chicago Press. Available at: [https://books.google.co.uk/books?hl=en&lr=&id=zNcpCwAAQBAJ&oi=fnd&pg=PA126&dq=succes+of+Asilomar&ots=Cx9sz--Jl3&sig=Hduka\\_JkXZTb0dU85HYhXnFDSd8#v=onepage&q=success%20of%20Asilomar&f=false](https://books.google.co.uk/books?hl=en&lr=&id=zNcpCwAAQBAJ&oi=fnd&pg=PA126&dq=succes+of+Asilomar&ots=Cx9sz--Jl3&sig=Hduka_JkXZTb0dU85HYhXnFDSd8#v=onepage&q=success%20of%20Asilomar&f=false) [Accessed 23 Apr. 2023].

Jing, K. and Xu, J. (2019). *A Survey on Neural Network Language Models*. [online] Available at: <https://arxiv.org/pdf/1906.03591.pdf>.

Joshi, P., Santy, S., Budhiraja, A., Bali, K. and Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. doi:<https://doi.org/10.18653/v1/2020.acl-main.560>.

Kahmann, C. and Heyer, G. (2019). Measuring Context Change to Detect Statements Violating the Overton Window. *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. [online] doi:<https://doi.org/10.5220/0008191803920396>.

Kaneko, M. and Bollegala, D. (2021). Dictionary-based Debiasing of Pre-trained Word Embeddings. *arXiv:2101.09525 [cs]*. [online] Available at: <https://arxiv.org/abs/2101.09525> [Accessed 18 Apr. 2023].

Kiritchenko, S. and Mohammad, S.M. (2018). *Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems*. [online] arXiv.org. Available at: <https://arxiv.org/abs/1805.04508>.

Kleinberg, J., Mullainathan, S. and Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]*. [online] Available at: <https://arxiv.org/abs/1609.05807>.

Kosinski, M. (2023). Theory of Mind May Have Spontaneously Emerged in Large Language Models. *arXiv:2302.02083 [cs]*. [online] Available at: <https://arxiv.org/abs/2302.02083>.

Kugelmass, H. (2016). ‘Sorry, I’m Not Accepting New Patients’. *Journal of Health and Social Behavior*, 57(2), pp.168–183. doi:<https://doi.org/10.1177/0022146516647098>.

Leech, T.G.J., Irby-Shasanmi, A. and Mitchell, A.L. (2018). ‘Are you accepting new patients?’ A pilot field experiment on telephone-based gatekeeping and Black patients’ access to pediatric care. *Health Services Research*, 54(54), pp.234–242. doi:<https://doi.org/10.1111/1475-6773.13089>.

Lewis, N. (2018). *Error*. [online] [www.shrm.org](http://www.shrm.org). Available at: [https://www.shrm.org/LearningAndCareer/learning/educational-program-materials/Documents/SHRM\\_Will%20AI%20Remove%20Hiring%20Bias.pdf](https://www.shrm.org/LearningAndCareer/learning/educational-program-materials/Documents/SHRM_Will%20AI%20Remove%20Hiring%20Bias.pdf).

Maass, A. (1999). *Linguistic Intergroup Bias: Stereotype Perpetuation Through Language*. [online] ScienceDirect. Available at: <https://www.sciencedirect.com/science/article/pii/S0065260108602725> [Accessed 18 Apr. 2023].

Mark (2023). *ChatGPT Passes Turing Test: A Turning Point for Language Models*. [online] MLYearning. Available at: <https://www.mlyearning.org/chatgpt-passes-turing-test/>.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), pp.1–35. doi:<https://doi.org/10.1145/3457607>.

Microsoft (2020). *Microsoft teams up with OpenAI to exclusively license GPT-3 language model*. [online] The Official Microsoft Blog. Available at: <https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/>.



Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality*. [online] Available at:

[https://proceedings.neurips.cc/paper\\_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf).

Nagel, S. (2023). *Blog – Common Crawl*. [online] Common Crawl. Available at:

<https://commoncrawl.org/connect/blog/>.

Nast, C. (2023). *The Hacking of ChatGPT Is Just Getting Started*. [online] Wired UK. Available at:

<https://www.wired.co.uk/article/chatgpt-jailbreak-generative-ai-hacking> [Accessed 18 Apr. 2023].

Perrigo, B. (2023). *Exclusive: The \$2 Per Hour Workers Who Made ChatGPT Safer*. [online] Time.

Available at: <https://time.com/6247678/openai-chatgpt-kenya-workers/>.

Portilho, T. (2019). *Opinion: The consequences of our blind faith in Artificial Intelligence are catching up to us*. [online] The Independent. Available at:

<https://www.independent.co.uk/voices/artificial-intelligence-ethics-police-bias-healthcare-a8837286.html> [Accessed 18 Apr. 2023].

Prates, M.O.R., Avelar, P.H. and Lamb, L.C. (2019). Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, 32.

doi:<https://doi.org/10.1007/s00521-019-04144-6>.

Recasens, M., Danescu-Niculescu-Mizil, C. and Jurafsky, D. (2013). *Linguistic Models for Analyzing and Detecting Biased Language*. [online] [www.semanticscholar.org](http://www.semanticscholar.org). Available at:

<https://www.semanticscholar.org/paper/Linguistic-Models-for-Analyzing-and-Detecting-Recasens-Danescu-Niculescu-Mizil/2a501b074261e81b9126e80a0a308cfa5e76f8c1> [Accessed 18 Apr. 2023].

Richter, F. (2022). *Infographic: English Is the Internet's Universal Language*. [online] Statista

Infographics. Available at: <https://www.statista.com/chart/26884/languages-on-the-internet/>.

Ritchie, S., Cheng, Y.-C., Chen, M., Mathews, R., van Esch, D., Li, B. and Sim, K.C. (2022). Large vocabulary speech recognition for languages of Africa: multilingual modeling and self-supervised learning. *arXiv:2208.03067 [cs, eess]*. [online] Available at: <https://arxiv.org/abs/2208.03067>

[Accessed 18 Apr. 2023].

Ritov, Y., Sun, Y. and Zhao, R. (2017). On conditional parity as a notion of non-discrimination in machine learning. *arXiv:1706.08519 [cs, stat]*. [online] Available at: <https://arxiv.org/abs/1706.08519>

[Accessed 18 Apr. 2023].

Rosenfeld, R. (2000). Two decades of statistical language modeling: where do we go from here? *Proceedings of the IEEE*, [online] 88(8), pp.1270–1278. doi:<https://doi.org/10.1109/5.880083>.

Sarewitz, D. and Nelson, R. (2008). Three rules for technological fixes. *Nature*, 456(7224), pp.871–872. doi:<https://doi.org/10.1038/456871a>.

Sheng, E., Chang, K.-W., Natarajan, P. and Peng, N. (2021). *Societal Biases in Language Generation: Progress and Challenges*. [online] arXiv.org. Available at: <https://arxiv.org/abs/2105.04054> [Accessed 18 Apr. 2023].

Shimizu 清水亮 / R. (2022). チャットできる AI、ChatGPT が「そこまですごくない」理由。見えてしまった限界. [online] BUSINESS INSIDER JAPAN. Available at: <https://www.businessinsider.jp/post-263042> [Accessed 18 Apr. 2023].

Tan, Y.C. and Celis, L.E. (2019). *Assessing Social and Intersectional Biases in Contextualized Word Representations*. [online] Neural Information Processing Systems. Available at: <https://proceedings.neurips.cc/paper/2019/hash/201d546992726352471cfea6b0df0a48-Abstract.html> [Accessed 18 Apr. 2023].

Terry, J.M., Hendrick, R., Evangelou, E. and Smith, R.L. (2010). Variable dialect switching among African American children: Inferences about working memory. *Lingua*, 120(10), pp.2463–2475. doi:<https://doi.org/10.1016/j.lingua.2010.04.013>.

The Economic Times (2023). Microsoft attempts to compete with Google with its AI-powered Bing search engine. *The Economic Times*. [online] 9 Feb. Available at: <https://economictimes.indiatimes.com/news/international/us/microsoft-attempts-to-compete-with-google-with-its-ai-powered-bing-search-engine/articleshow/97743265.cms?from=mdr>.

van de Poel, I. (2015). An Ethical Framework for Evaluating Experimental Technology. *Science and Engineering Ethics*, 22(3), pp.667–686. doi:<https://doi.org/10.1007/s11948-015-9724-3>.

Wikipedia (2020). *Wikipedia:Neutral point of view*. [online] Wikipedia. Available at: [https://en.wikipedia.org/wiki/Wikipedia:Neutral\\_point\\_of\\_view](https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view).

Winner, L. (1980). *Do Artifacts Have Politics?* [online] www.jstor.org. Available at: [https://www.jstor.org/stable/pdf/20024652.pdf?casa\\_token=OXJ9Si0Qu8oAAAAA:0UFjgWM-y\\_4RR2v\\_sZygaONVtkWTOL5mWW422n0TgksinSGwwOTWtmYsEg-jDnEZkCgBQfeLS0g9\\_WHMH0NpMtYfJqaTGNe3d95VMR3ikmsg1XoS-yTJ](https://www.jstor.org/stable/pdf/20024652.pdf?casa_token=OXJ9Si0Qu8oAAAAA:0UFjgWM-y_4RR2v_sZygaONVtkWTOL5mWW422n0TgksinSGwwOTWtmYsEg-jDnEZkCgBQfeLS0g9_WHMH0NpMtYfJqaTGNe3d95VMR3ikmsg1XoS-yTJ) [Accessed 18 Apr. 2023].

Woodhouse, E. and Sarewitz, D. (2007). Science policies for reducing societal inequities. *Science and Public Policy*, 34(2), pp.139–150. doi:<https://doi.org/10.3152/030234207x195158>.

Worthington, R. and Collingridge, D. (1982). The Social Control of Technology. *The American Political Science Review*, 76(1), p.134. doi:<https://doi.org/10.2307/1960465>.

Xiang, C. (2023). 'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says. [online] [www.vice.com](https://www.vice.com). Available at: <https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says>.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V. and Chang, K.-W. (2019). *Gender Bias in Contextualized Word Embeddings*. [online] ACLWeb. doi:<https://doi.org/10.18653/v1/N19-1064>.

Zhao, J., Zhou, Y., Li, Z., Wang, W. and Chang, K.-W. (2018). Learning Gender-Neutral Word Embeddings. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. doi:<https://doi.org/10.18653/v1/d18-1521>.