Received: 31 December 2020 | Revised: 14 June 2021 | Accepted: 20 June 2021

DOI: 10.1111/lnc3.12432

ARTICLE

WILEY

Five sources of bias in natural language processing

Dirk Hovy[1] | **Shrimai Prabhumoye**[2]

[1]Marketing Department, Bocconi University, Milan, Italy

[2]School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Correspondence
Dirk Hovy, Marketing Department, Bocconi University, Via Roentgen 1-2, Milan 20136, Italy.
Email: dirk.hovy@unibocconi.it

Funding information
H2020 European Research Council: Grant/Award Number: 949944, INTEGRATOR
Open Access Funding provided by Universita Bocconi within the CRUI-CARE Agreement.

Abstract

Recently, there has been an increased interest in demographically grounded bias in natural language processing (NLP) applications. Much of the recent work has focused on describing bias and providing an overview of bias in a larger context. Here, we provide a simple, actionable summary of this recent work. We outline five sources where bias can occur in NLP systems: (1) the data, (2) the annotation process, (3) the input representations, (4) the models, and finally (5) the research design (or how we conceptualize our research). We explore each of the bias sources in detail in this article, including examples and links to related work, as well as potential counter-measures.

1 | INTRODUCTION

A visitor who arrived at night in the London of 1878 would not have seen very much: cities were dark, only illuminated by candles and gas lighting. The advent of electricity changed that. This technology lit up cities everywhere and conferred a whole host of other benefits. From household appliances to the internet, electricity brought in a new era. However, as with every new technology, electricity had some unintended consequences. To provide the energy necessary to light up cities, run the appliances and fuel the internet, we required more power plants. Those plants contributed to pollution and ultimately to the phenomenon of global warming. Presumably, those consequences were far from the minds of the people who had just wanted to illuminate their cities.

Dirk Hovy and Shrimai Prabhumoye contributed equally.

This dual nature of intended use and unintended consequences is common to all new technologies. And natural language processing (NLP) proves no exception. NLP has progressed relatively rapidly from a niche academic field to a topic of widespread industrial and political interest. Its economic impact is substantial: NLP-related companies are predicted to be valued at $26.4 billion by 2024.[1] We make daily use of machine translation (Wu et al., 2016), personal assistants like Siri or Alexa (Palakovich et al., 2017; Ram et al., 2018) and text-based search engines (Brin & Page, 1998). NLP is used in industrial decision-making processes for hiring, abusive language and threat detection on social media (Roberts et al., 2019), and mental health assessment and treatment (Benton et al., 2017; Coppersmith et al., 2015). An increasing volume of social science research uses NLP to generate insights into society and the human mind (Bhatia, 2017; Kozlowski et al., 2019).

However, the interest in and use of NLP have grown much faster than an understanding of the unintended consequences. Some researchers have pointed out how NLP technologies can be used for harmful purposes, such as suppressing dissenters (Bamman et al., 2012; Zhang et al., 2014), compromising privacy/anonymity (Coavoux et al., 2018; Grouin et al., 2015), or profiling (Jurgens et al., 2017; Wang et al., 2018). Those applications might be unintended outcomes of systems developed for other purposes but could be deliberately developed by malicious actors. A much more widespread unintended negative consequence is the unfairness caused by demographic biases, such as unequal performance for different user groups (Tatman, 2017), misidentification of speakers and their needs (Criado Perez, 2019) or the proliferation of harmful stereotypes (e.g., Agarwal et al., 2019; Koolen & van Cranenburgh, 2017; Kiritchenko & Mohammad, 2018). In this work, we follow the definition of Shah et al. (2020) for 'bias', which focuses on the mismatch of ideal and actual distributions of labels and user attributes in training and application of a system.

These biases are partially due to the rapid growth of the field and an inability to adapt to the new circumstances. Originally, machine learning and NLP were about solving toy problems on small data sets, promising to do it on more extensive data later. Any scepticism and worry about Artificial Intelligence (AI's) power were primarily theoretical. In essence, there was not enough data or computational power for these systems to impact people's lives. With the recent availability of large amounts of data and the universal application of NLP, this point has now arrived. However, even though we now have the possibility, many models are still trained without regard for demographic aspects. Moreover, many applications are focussed solely on information content, without awareness or concern for those texts' authors and the social meaning of the message (Hovy & Yang, 2021). But today, NLP's reach and ubiquity do have a real impact on people's lives (Hovy & Spruit, 2016). Our tools, for better or for worse, are used in everyday life. The age of academic innocence is over: we need to be aware that our models affect people's lives, yet not always in the way we imagine (Ehni, 2008). The most glaring reason for this disconnect is bias at various steps of the research pipeline.

The focus on applications has moved us away from models as a tool for understanding and towards predictive models. It has become clear that these tools produce excellent predictions but are much harder to analyse. They solve their intended task but also pick up on secondary aspects of language and potentially exploit them to fulfil the objective function. And language carries a lot of secondary information about the speaker, their self-identification, and membership in socio-demographic groups (Flek, 2020; Hovy & Spruit, 2016; Hovy & Yang, 2021). Whether I say 'I am totally pumped' or 'I am very excited' conveys information about me far beyond the actual meaning of the sentence. In a conversation, we actively use this information to pick up on a speaker's likely age, gender, regional origin or social class (Eckert, 2019; Labov, 1972). We know

that the same sentence ('That was a sick performance!') can express either approval or disgust, based on whether a teenager or an octogenarian says it.

In contrast, current NLP tools fail to incorporate demographic variation and instead expect all language to follow the 'standard' encoded in the training data. But the question is: whose standard (Eisenstein, 2013)? This approach is equivalent to expecting everyone to speak like the octogenarian from above: it leads to problems when encountering the teenager. As a consequence, NLP tools trained on one demographic sample perform worse on another sample (Garimella et al., 2019; Hovy & Søgaard, 2015; Jørgensen et al., 2015). This mismatch was known to affect text domains, but it also applies to socio-demographic domains: people of, say, different age groups are linguistically as diverse as a text from a web blog and a newspaper (Johannsen et al., 2015). Incidentally, demographics like age and text-domain can often be correlated (Hovy, 2015). Plank (2016) therefore suggests treating these aspects of language as parts of our understanding of 'domain'.

The consequences of these shortfalls range from an inconvenience to something much more insidious. In the most straightforward cases, systems fail and produce no output. This outcome is annoying and harms the user who cannot benefit from the service, but at least it is obvious enough for the user to see and respond to. In many cases, though, the effect is much less easy to notice: the performance degrades, producing sub-par output for some users. This difference will become only evident in comparison but is not apparent to the individual user. This degradation is much harder to see but often systematic for a particular demographic group and creates a demographic bias in NLP applications.

The problem of bias introduced by socio-demographic differences in the target groups is not restricted to NLP, though, but occurs in all data sciences (O'Neil, 2016). For example, in speech recognition, there is a strong bias towards native speakers of any language (Lawson et al., 2003). But even for native speakers, there are barriers: dialect speakers or children can struggle to make themselves understood by a smart assistant (Harwell, 2018). Moreover, women and children—who speak in a higher register than the speakers in the predominantly male training sample—might not be processed correctly (or at all) in voice-to-text systems (Criado Perez, 2019). There have been several examples of computer vision bias, from an image captioning system labelling pictures of black people as 'gorillas' to cameras designed to detect whether a subject blinked, which malfunctioned if Asian people were in the picture (Howard & Borenstein, 2018). In a more abstract form, the correlation of socio-demographics with variables of interest can cause problems, such as when ZIP code and income level can act as proxies for race (O'Neil, 2016). ProPublica reported that a machine learning system designed to predict bail decisions overfit on the defendants' skin colour: in this case, the social prejudices of the prior decisions became encoded in the data (Angwin et al., 2016).

With great (predictive) power comes great responsibility, and several ethical questions arise when working with language. There are no hard and fast rules for everything, and the topic is still evolving, but several issues have emerged so far. While the overview here is necessarily incomplete, it is a starting point on the issue. It is based on recent work by Hovy and Spruit (2016), Shah et al. (2020) as well as the two workshops by the Association for Computational Linguistics on Ethics in NLP (Alfano et al., 2018; Hovy et al., 2017). See those sources for further in-depth discussion.

2 | OVERVIEW

This article is not the first attempt to comprehensively address demographic factors in NLP. General bias frameworks in Artificial Intelligence (AI) exist that lay the necessary groundwork for our approach. For example, Friedler et al. (2021) defined bias as fairness in algorithms by capturing all the latent features (i.e., demographics) in the data. Suresh and Guttag (2019) suggested a qualitative framework for bias in machine learning, defining bias as a 'potential harmful property of the data', though they leave out demographic and modelling aspects. Hovy and Spruit (2016) noted three qualitative sources of bias: data, modelling and research design, related to demographic bias, overgeneralization and topic exposure. In Shah et al. (2020), these and other frameworks are combined under a joint mathematical approach. Blodgett et al. (2020) provide an extensive survey of the way bias is studied in NLP. It points out the weaknesses in the research design and recommends grounding work analysing 'bias' in NLP systems in the relevant literature outside of NLP, understanding why system behaviours can be harmful and to whom, and engaging in a conversation with the communities that are affected by the NLP systems.

One thing to stress is that 'bias' per se is neither good nor bad: in a Bayesian framework, the prior $P(X)$ serves as a bias: the expectation or base-rate we should have for something before we see any further evidence. In real life, many of our reactions to everyday situations are biases that make our lives easier. Biases as a preset are not necessarily an issue: they only become problematic when they are kept even in the face of contradictory evidence or when applied to areas they were not meant for.

Many of the biases we will discuss here can also represent a form of information: as a diagnostic tool about the state of society (Garg et al., 2018; Kozlowski et al., 2019), or as a way to regularize our models (Plank et al., 2014a; Uma et al., 2020). However, as input to predictive systems, these biases can have severe consequences and exacerbate existing inequalities between users.

Figure 1 shows the five sources of bias we discuss in this article. The first entry point for bias in the NLP pipeline is the choice of data for the experimentation. The labels chosen for training and the procedure used for annotating the labels introduces the annotation bias. Selection bias is introduced by the samples chosen for training or testing an NLP model. The third type of bias is introduced by the choice of representation used for the data. The choice of models or machine learning algorithms used also introduces the issue of bias amplification. Finally, the entire research design process can introduce bias if researchers are not careful with their choices in the NLP pipeline. In what follows, we discuss each of these biases in detail and provide insights into how they occur and how to mitigate them.

2.1 | Bias from data

NLP systems reflect biases in the language data used for training them. Many data sets are created from long-established news sources (e.g., Wall Street Journal, Frankfurter Rundschau from the 1980s through the 1990s), a very codified domain predominantly produced by a small, homogeneous sample: typically white, middle-aged, educated, upper-middle-class men (Garimella et al., 2019; Hovy & Søgaard, 2015). However, many syntactic analysis tools (taggers and parsers) are still trained on the newswire data from the 1980s and 1990s. Modern syntactic tools, therefore, expect everyone to speak like journalists from the 1980s. It should come as no surprise that most people today do not: language has evolved since then, and expressions that were ungrammatical

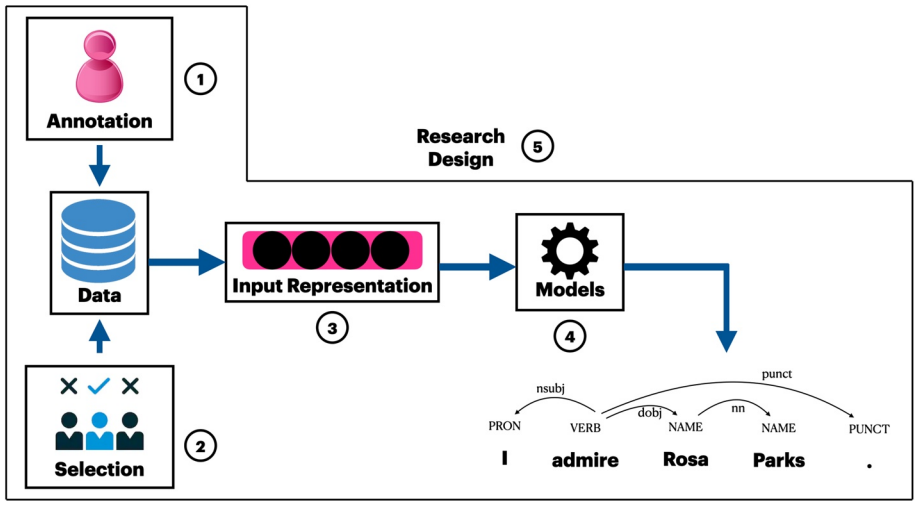

FIGURE 1 Schematic of the five bias sources in the general natural language processing pipeline

then are acceptable today, 'because internet' (McCulloch, 2020). NLP is, therefore, unprepared to cope with this demographic variation.

Models trained on these data sets treat language as if it resembles this restricted training data, creating demographic bias. For example, Hovy (2015) and Jørgensen et al. (2015) have shown that this bias leads to significantly decreased performance for people under 35 and ethnic minorities, even in simple NLP tasks like finding verbs and nouns (i.e., part-of-speech tagging). The results are ageist, racist or sexist models that are biased against the respective user groups. This is the issue of ***selection bias***, which is rooted in data.

When choosing a text data set to work with, we are also making decisions about the demographic groups represented in the data. As a result of the demographic signal present in language, any data set carries a demographic bias, that is, latent information about the demographic groups present in it. As humans, we would not be surprised if someone who grew up hearing only their dialect would have trouble understanding other people. If our data set is dominated by the 'dialect' of a specific demographic group, we should not be surprised that our models have problems understanding others. Most data sets have some built-in bias, and in many cases, it is benign. It becomes problematic when this bias negatively affects certain groups or disproportionately advantages others. On biased data sets, statistical models will overfit to the presence of specific linguistic signals that are particular to the dominant group. As a result, the model will work less well for other groups, that is, it excludes demographic groups. Hovy (2015) and Jørgensen et al. (2015) have shown the consequences of exclusion for various groups, for example, people under 35 and speakers of African-American vernacular English. Part-of-speech(POS) tagging models have a significantly lower accuracy for young people and ethnic minorities, vis-á-vis the dominant demographics in the training data. Apart from exclusion, these models will pose a problem for future research. Given that a large part of the world's population is currently under 30, such models will degrade even more over time and ultimately not meet their users' needs. This issue also has severe ramifications for the general applicability of any findings using these tools. In psychology, most studies are based on college students, a very specific demographic: western, educated, industrialized, rich and democratic research participants (so-called WEIRD; Henrich et al., 2010). The assumption that findings from this group would generalize to all other

demographics has proven wrong and led to a heavily biased corpus of psychological data and research.

2.1.1 | Counter-measures

Potential counter-measures to demographic selection bias can be simple. The most salient is undoubtedly to pay more attention to how data is collected and clarify what went into the construction of the data set. Bender and Friedman (2018) proposed a framework to document these decisions in a Data Statement. This statement includes various aspects of the data collection process and the underlying demographics. It provides future researchers with a way to assess the effect of any bias they might notice when using the data. As a beneficial side effect, it also forces us to consider how our data is made up. For already existing data sets, post-stratification is the down-sampling of over-represented groups in the training data to even out the distribution until it reflects the actual distribution. Mohammady and Culotta (2014) have shown how existing demographic statistics can be used as supervision. In general, we can use measures to address overfitting or imbalanced data to correct for demographic bias in data. However, as various papers have pointed out (Bender et al., 2021; Hutchinson et al., 2021), addressing data bias is not a 'one-and-done' exercise but requires continual monitoring throughout a data sets lifecycle.

Alternatively, we can also collect additional data to balance existing data sets to account for exclusions or misrepresentations. Webster et al. (2018) released a gender-balanced data set for co-reference resolution task. Zhao et al. (2017) also explore balancing a data set with gender confounds for multi-label object classification and visual semantic role labelling tasks. Data augmentation by controlling the gender attribute is an effective technique in mitigating gender bias in NLP processes (Dinan et al., 2020; Sun et al., 2019). Wei and Zou (2019) explore data augmentation techniques that improve performance on various text classification tasks.

2.2 | Bias from annotations

Annotation can introduce bias in various forms through a mismatch of the annotator population with the data. This is the issue of ***label bias***. Label and selection bias can—and most often do—interact, so it can be challenging to distinguish them. It does, however, underscore how important it is to address them jointly. There are several ways in which annotations introduce bias.

In its simplest form, bias arises because annotators are distracted, uninterested, or lazy about the annotation task. As a result, they choose the 'wrong' labels. More problematic is label bias from informed and well-meaning annotators that systematically disagree. Plank et al. (2014b) have shown that this type of bias arises when there is more than one possible correct label. For example, the term 'social media' can be validly analysed as either a noun phrase composed of an adjective and a noun, or a noun compound, composed of two nouns. Which label an annotator chooses depends on their interpretation of how lexicalized the term 'social media' is. If they perceive it as fully lexicalized, they will choose a noun compound. If they believe the process is still ongoing, that is, the phrase is analytical, they will choose an 'adjective plus noun' construct. Two annotators with these opposing views will systematically label 'social' as an adjective or a noun, respectively. While we can spot the disagreement, we cannot discount either of them as wrong or malicious.

Finally, label bias can result from a mismatch between authors' and annotators' linguistic and social norms. Sap et al. (2019) showed that they reflect social and demographic differences, for example, that annotators rate the utterances of different ethnic groups differently and that they mistake innocuous banter as hate speech because they are unfamiliar with communication norms of the original speakers.

There has been a movement towards increasingly using annotations from crowdsourcing rather than trained expert annotators. While it is cheaper and (in theory) equivalent to the quality of trained annotators (Snow et al., 2008), it does introduce a range of biases. For example, various works have shown that crowdsourced annotators' demographic makeup is not as representative as one might hope (Pavlick et al., 2014). On the one hand, crowdsourcing is easier to scale, potentially covering more diverse backgrounds than we would find in expert annotator groups. On the other hand, it is much harder to train and communicate with crowdsourced annotators, and their incentives might not align with the projects we care about. For example, suppose we ask crowd workers to annotate concepts like dogmatism, hate speech, or microaggressions. Their answers will inherently include their societal perspective of these concepts. This bias can be good or bad, depending on the sample of annotators: we may get multiple perspectives that approximate the population as a whole, or annotations may get skewed results due to the selection. However, we might also not want various perspectives if there is a theoretically motivated and well-defined way in which we plan to annotate. Crowdsourcing and its costs raise several other ethical questions about worker payment and fairness (Fort et al., 2011).

2.2.1 | Counter-measures

Malicious annotators are luckily relatively easy to spot and can be remedied by using multiple annotations per item and aggregating with an annotation model (Hovy et al., 2013; Passonneau & Carpenter, 2014; Paun et al., 2018). These models help us find biased annotators and let us account for the human disagreement between labels. A free online version of such a tool is available at https://mace.unibocconi.it/. They presuppose, however, that there is a single correct gold label for each data point and that annotations are simply corruptions of it.

If there is more than one possible correct answer, we can use disagreement information in the update process of our models (Fornaciari et al., 2021; Plank et al., 2014a; Uma et al., 2020). That is, we can encourage the models to make more minor updates if human annotators easily confuse the categories with each other (say, adjectives and nouns in noun compounds like 'social media'). We make regular updates if they are mutually exclusive categories (such as verbs and nouns).

The only way to address mismatched linguistic norms is to pay attention to selecting annotators (i.e., matching them to the author population in terms of linguistic norms) or provide them with dedicated training. The latter should be generally considered. While annotator training is time-intensive and potentially costly, it can be worth the effort in terms of better and less biased labels.

2.3 | Bias from input representations

Even balanced, well-labelled data sets contain bias: the most common text inputs representing in NLP systems, word embeddings (Mikolov et al., 2013), have been shown to pick up on racial

and gender biases in the training data (Bolukbasi et al., 2016; Manzini et al., 2019). For example, 'woman' is associated with 'homemaker' in the same way 'man' is associated with 'programmer'. There has been some justified scepticism over whether these analogy tasks are the best way to evaluate embedding models (Nissim et al., 2020), but there is plenty of evidence that (1) embeddings do capture societal attitudes (Bhatia, 2017; Garg et al., 2018; Kozlowski et al., 2019), and that (2) these societal biases are resistant to many correction methods (Gonen & Goldberg, 2019). This is the issue of ***semantic bias***.

These biases hold not just for word embeddings but also for the contextual representations of big pre-trained language models that are now widely used in different NLP systems. As they are pre-trained on almost the entire available internet, they are even more prone to societal biases. Several papers have shown that these models reproduce and thereby perpetuate these biases and stereotypes (Kurita et al., 2019; Tan and Celis, 2019).

There exist a plethora of efforts for debiasing embeddings (Bolukbasi et al., 2016; Sun et al., 2019; Zhao et al., 2017, 2019). The impact and applicability of debiased embeddings are unclear on a wide range of downstream tasks. As stated above, biases are usually masked, not entirely removed, by these methods. Even if it was possible to remove biases in the embeddings, it is not always clear whether it is useful (bias might carry information).

A central issue is the language models' training objective: to predict the most likely next term, given the previous context (*n*-grams). While this objective captures distributional semantic properties, it may itself not contribute to building unbiased embeddings, as it represents the world as we find it, rather than as we would like to have it (descriptive vs. normative view).

2.3.1 | Counter-measures

In general, when using embeddings for downstream applications, it is good practice to be aware of their biases. This awareness helps to identify the applicability of such embeddings to your specific domains and tasks. For example, these models are not directly applicable to data sets that contain scientific articles or medical terminologies.

Recent work has focussed on debiasing embeddings for specific downstream applications and groups of the population. For example debiasing embeddings for reducing gender bias in text classification (Prost et al., 2019), dialogue generation (Dinan et al., 2020; Liu et al., 2020), and machine translation (Font & Costa-jussà, 2019). Such efforts are more conscious of the effects of debiasing on the target application. Additional metrics, approaches and data sets have been proposed to measure the bias inherent in large language models and their sentence completions (Nangia et al., 2020; Nozza et al., 2021; Sheng et al., 2019).

2.4 | Bias from models

Simply using 'better' training data is not a feasible long-term solution: languages evolve continuously, so even a representative sample can only capture a snapshot—at best a short-lived solution (see Fromreide et al., 2014). These biases compound to create severe performance differences for different user groups. Zhao et al. (2017) demonstrated that systems trained on biased data exacerbate that bias even further when applied to new data, and Kiritchenko and Mohammad (2018) have shown that sentiment analysis tools pick up on societal prejudices, leading to different outcomes for different demographic groups. For example, by merely changing the gender of a

pronoun, the systems classified the sentence differently. Hovy et al. (2020) found that machine translation systems changed the perceived user demographics to make samples sound older and more male in translation. This issue is ***bias overamplification***, which is rooted in the models themselves.

One of the sources of bias overamplification is the choice of loss objective used in training the models. These objectives usually correspond to improving the precision of the predictions. Models might exploit spurious correlations (e.g., all positive examples in the training data happened to come from female authors so that gender can be used as a discriminative feature) or statistical irregularities in the data set to achieve higher precision (Gururangan et al., 2018; Poliak et al., 2018). In other words, they might give the correct answers for the wrong reasons. This behaviour is hard to track until we find a consistent case of bias.

Another issue with the design of machine learning models is that they *always* make a prediction, even when they are unsure or when they cannot know the answer. The latter could be due to the test data point lying outside the training data distribution or the model's representation space. Prabhumoye et al. (2021) discuss this briefly in a case study for machine translation systems. If a machine translation tool translates the gender-neutral Turkish 'O bir doktor, o bir hemşire' into 'He is a doctor, she is a nurse', it might provide us with an insight into societal expectations (Garg et al., 2018). Still, it also induces an incorrect result the user did not intend. Ideally, models should report to the user that they could not translate rather than produce a wrong translation.

2.4.1 | Counter-measures

The susceptibility of models to all aspects of the training data makes it so important to test our systems on various held-out data sets rather than a single, designated test set. Recent work has explored objectives other than recall, F1 and so on, for example, the performance stratified by subgroup present in the data. These metrics can lead to fairer predictions across subgroups (Chouldechova, 2017; Corbett-Davies & Goel, 2018; Dixon et al., 2018), for example, if the metrics show that the performance for a specific group is much lower than for the rest. Moving away from pure performance metrics and looking at the robustness and behaviour of the model in suites of specially designed cases can add further insights (Ribeiro et al., 2020).

Card and Smith (2020) explore constraints to be specified on outcomes of models. Specifically, these constraints ensure that the proportion of predicted labels should be the same or approximately the same for each user group.

More generally, methods designed to probe and analyse the model can help us understand how it reached decisions. Neural features like attention (Bahdanau et al., 2015) can provide visualizations. Kennedy et al. (2020) propose a sampling-based algorithm to explore the impact of individual words on classification. As policy changes put an increased focus on explainable AI (EU High-Level Expert Group on AI, 2019), such methods will likely become useful for both bias spotting and legal recourse.

Systems that explicitly model user demographics will help produce both more personalized and less biased translations (Font & Costa-jussà, 2019; Mirkin et al., 2015; Mirkin & Meunier, 2015; Saunders & Byrne, 2020; Stanovsky et al., 2019).

2.5 | Bias from research design

Despite a growing interest in multi- and cross-lingual work, most NLP research is still in and on English. It generally focuses on Indo-European data/text sources, rather than other language groups or smaller languages, for example, in Asia or Africa (Joshi et al., 2020). Even if there is a potential wealth of data available from other languages, most NLP tools skew towards English (Munro, 2013; Schnoebelen, 2013).

This underexposure is a self-fulfilling prophecy: researchers are less likely to work on those languages for which there are not many resources. Instead, they work on languages and tasks for which data is readily available, potentially generating more data in the process. Consequently, there is a severe shortage for some languages but an overabundance for others. In a random sample of Tweets from 2013, there were 31 different languages (Plank, 2016), but no treebanks for about two-thirds of them and even fewer semantically annotated resources like WordNets. Note that the number of language *speakers* does not necessarily correlate with the number of available *resources*. These were not obscure languages with few speakers, but often languages with millions of speakers. The shortage of syntactic resources has since been addressed by the Universal Dependency Project (Nivre et al., 2020). However, a recent paper (Joshi et al., 2020) found that most conferences still focus on the well-resourced languages and are less inclusive of less-resourced ones.

This dynamic makes new research on smaller languages more complicated, and it naturally directs new researchers towards the existing languages, first among them English. The existence of off-the-shelf tools for English makes it easy to try new ideas in English. The focus on English may therefore be self-reinforcing and has created an overexposure of this variety. The overexposure to English (as well as to particular research areas or methods) creates a bias described by the availability heuristic (Tversky & Kahneman, 1973). If we are exposed to something more often, we can recall it more efficiently, and if we can recall things quickly, we infer that they must be more important, bigger, better, more dangerous and so on. For instance, people estimate the size of cities they recognize to be larger than that of unknown cities (Goldstein & Gigerenzer, 2002). It requires a much higher start-up cost to explore other languages in terms of data annotation, basic analysis models and other resources. The same holds for languages, methods and topics we research.

Overexposure can also create or feed into existing biases, for example, that English is the 'default' language, even though both morphology and syntax of English are global outliers. It is questionable whether NLP would have focused on *n*-gram models to the same extent if it had instead been developed on a morphologically complex language (e.g., Finnish, German). However, because of the unique structure of English, *n*-gram approaches worked well, spread to become the default approach and only encountered problems when faced with different languages. Lately, there has been a renewed interest beyond English, as there are economic incentives for NLP groups to work on and in other languages. Concurrently, new neural methods have made more multi-lingual and cross-lingual approaches possible. These methods include, for example, multi-lingual representations (Devlin et al., 2019; Nozza et al., 2020) and the zero-shot learning they enable (e.g., Bianchi et al., 2021; Jebbara & Cimiano, 2019; Liu et al., 2019, inter alia). However, English is still one of the most widely spoken languages and by far the biggest market for NLP tools. So there are still more commercial incentives to work on English than other languages, perpetuating the overexposure.

One of the reasons for the linguistic and cultural skew in research is the makeup of research groups themselves. In many cases, these groups do not necessarily reflect the demographic

composition of the user base. Hence, marginalized communities or speaker groups do not have their voice represented proportionally. Initiatives like Widening NLP[2] are beginning to address this problem, but the issue still leaves a lot of room for improvement.

Finally, not analysing the behaviour of models sufficiently, or not fully disclosing it can be harmful (Bianchi & Hovy, 2021). These omissions are not necessarily due to ill will, but are often the result of a relentless pressure to publish. An example of the resulting bias is not fully understanding the intended use of the trained models and how they can be misused (i.e., its dual use). The introduction of ethical consideration sections and an ethics reviews in NLP venues is a step to give these aspects more attention and encourage reflection.

An interesting framework to think about these issues is the suggestion by Van de Poel (2016) to think of new technology (such as NLP) as a large-scale social experiment. An experiment we are all engaged in at a massive scale. As an experiment, however, we need to make sure we respect specific guidelines and ground rules. There are detailed requirements for social and medical sciences experiments to get the approval of an ethics committee or IRB (internal review board). These revolve around the safety of the subjects and involve *beneficence* (no harm to subjects, maximize benefits, minimize risk), respect for *subjects' autonomy* (informed consent), and justice (weighing of benefits vs. harms, protection of vulnerable subjects). Not all of these categories are easily translated into NLP as a large-scale experiment. However, it can help us frame our decisions within specific philosophical schools of thought, as outlined by Prabhumoye et al. (2021).

2.5.1 | Counter-measures

There are no easy solutions to design bias, which might only become apparent in hindsight. However, any activity or measure that increases the chance of reflection on the project can help to counter inherent biases. For example, Emily Bender has suggested making overexposure bias more apparent by stating explicitly which language we work on 'even if it is English' (Bender, 2019). There is, of course, no issue with research on English, but it should be made explicit that the results might not automatically hold for all languages.

It can help to ask ourselves counterfactuals: '*Would I research this is if the data wasn't as easily available? Would my finding still hold on another language?'* We can also try to assess whether the research direction of a project feeds into existing biases or whether it overexposes certain groups.

A way forward is to use various evaluation settings and metrics (Ribeiro et al., 2020). Some conferences have also started suggesting guidelines to assess the potential for ethical issues with a system (e.g., the NAACL 2021 Ethics FAQ and guidelines).[3] Human intervention and thought are required at every stage of the NLP application design lifecycle to prioritize equity and stakeholders from marginalized groups (Costanza-Chock, 2020). Recent work by Bird (2020) suggests new ways of collaborating with Indigenous communities in the form of open discussions and proposes a postcolonial approach to computational methods for supporting language vitality. Finally, Havens et al. (2020) discuss the need for a bias-aware methodology in NLP and present a case study in executing it. Researchers have to be mindful of the entire research design: data sets they choose, the annotation schemes or labelling procedures they follow, how they decide to represent the data, the algorithms they choose for the task and how they evaluate the automated systems. Researchers need to be aware of the real-world applications of their work and consciously decide to choose to help marginalized communities via technology (Asad et al., 2019).

3 | CONCLUSION

This article outlined five of the most common sources of bias in NLP models: data selection, annotation, representations, models and our own research design. However, we are not merely at the mercy of these biases: there exists a growing arsenal of algorithmic and methodological approaches to mitigate biases from all sources. The most difficult might be bias from research design, which requires introspection and systematic analysis of our own preconceived notions and blind spots.

ACKNOWLEDGEMENT

Open Access Funding provided by Universita Bocconi within the CRUI-CARE Agreement.

ENDNOTES

[1] http://www.marketsandmarkets.com/Market-Reports/natural-language-processing-nlp-825.html.

[2] http://www.winlp.org/.

[3] https://2021.naacl.org/ethics/faq/.

REFERENCES

Agarwal, O., Durupınar, F., Badler, N. I., & Nenkova, A. (2019). Word embeddings (also) encode human personality stereotypes. *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, 205–211. *Minneapolis, Minnesota*: Association for Computational Linguistics. https://www.aclweb.org/anthology/S19-1023

Alfano, M., Hovy, D., Mitchell, M., & Strube, M. (Eds.). (2018). *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing. New Orleans, Louisiana, USA*: Association for Computational Linguistics. https://www.aclweb.org/anthology/W18-0800

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica, May*, 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Asad, M., Dombrowski, L., Costanza-Chock, S., Erete, S., & Harrington, C. (2019). Academic accomplices: Practical strategies for research justice. *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion* (pp. 353–356).

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Y. Bengio, & Y. LeCun (Eds.), *3rd International Conference on Learning Representations*, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. http://arxiv.org/abs/1409.0473

Bamman, D., O'Connor, B., & Smith, N. (2012). Censorship and deletion practices in Chinese social media. *First Monday*, 17.

Bender, E. (2019). *The BenderRule: On naming the languages we study and why it matters*. The Gradient. https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623).

Benton, A., Mitchell, M., & Hovy, D. (2017). Multitask learning for mental health conditions with limited social media data. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 152–162. Valencia, Spain: Association for Computational Linguistics. https://www.aclweb.org/anthology/E17-1015

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*, 1–20.

Bianchi, F., & Hovy, D. (2021). On the gap between adoption and understanding in nlp. *Findings of the Association for Computational Linguistics: ACL 2021*. Association for Computational Linguistics.

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E. (2021). Cross-lingual contextualized topic models with zero-shot learning. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1676–1683. Association for Computational Linguistics. https://www.aclweb.org/anthology/2021.eacl-main.143

Bird, S. (2020). Decolonising speech and language technology. *Proceedings of the 28th International Conference on Computational Linguistics*, 3504–3519. Barcelona, Spain (Online): International Committee on Computational Linguistics. https://www.aclweb.org/anthology/2020.coling-main.313

Blodgett, S. L., Barocas, S., Daumé, H., III, & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.acl-main.485

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 4349–4357.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, *30*(1–7), 107–117.

Card, D., & Smith, N. A. (2020). On consequentialism and fairness. *Frontiers in Artificial Intelligence*, *3*, 34. https://www.frontiersin.org/article/10.3389/frai.2020.00034

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*, 153–163.

Coavoux, M., Narayan, S., & Cohen, S. B. (2018). Privacy-preserving neural representations of text. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1–10. Brussels, Belgium: Association for Computational Linguistics. https://www.aclweb.org/anthology/D18-1001

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). Clpsych 2015 shared task: Depression and ptsd on twitter. *CLPsych@ HLT-NAACL*, 31–39.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Criado Perez, C. (2019). *Invisible women: Exposing data bias in a world designed for men*. Random House.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1423

Dinan, E., Fan, A., Williams, A., Urbanek, J., Kiela, D., & Weston, J. (2020). Queens are powerful too: Mitigating gender bias in dialogue generation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8173–8188. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.emnlp-main.656

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES* '18, 67 –73. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3278721.3278729

Eckert, P. (2019). The limits of meaning: Social indexicality, variation, and the cline of interiority. *Language*, *95*, 751–776.

Ehni, H.-J. (2008). Dual use and the ethical responsibility of scientists. *Archivum Immunologiae et Therapiae Experimentalis*, *56*, 147–152.

Eisenstein, J. (2013). What to do about bad language on the internet. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 359–369. Atlanta, Georgia: Association for Computational Linguistics. https://www.aclweb.org/anthology/N13-1037

EU High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

Flek, L. (2020). Returning the N to NLP: Towards contextually personalized classification models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7828–7838. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.acl-main.700

Font, J. E., & Costa-jussà, M. R. (2019). Equalizing gender bias in neural machine translation with word embeddings techniques. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 147–154. Florence, Italy: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-3821

Fornaciari, T., Uma, A., Paun, S., Plank, B., Hovy, D., & Poesio, M. (2021). Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2591–2597. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2021.naacl-main.204

Fort, K., Adda, G., & Cohen, K. B. (2011). Last words: Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, *37*, 413–420. https://www.aclweb.org/anthology/J11-2010

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, *64*, 136–143.

Fromreide, H., Hovy, D., & Søgaard, A. (2014). Crowdsourcing and annotating NER for Twitter #drift. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2544–2547. Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/421_Paper.pdf

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*, E3635–E3644.

Garimella, A., Banea, C., Hovy, D., & Mihalcea, R. (2019). Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3493–3498. Florence, Italy: Association for Computational Linguistics. https://www.aclweb.org/anthology/P19-1339

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*, 75–90.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 609–614. https://www.aclweb.org/anthology/N19-1061

Grouin, C., Griffon, N., & Névéol, A. (2015). Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs? *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, 31–39. Lisbon, Portugal: Association for Computational Linguistics. https://www.aclweb.org/anthology/W15-2604

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *NAACL-HLT* (2).

Harwell, D. (2018). *The accent gap. Why some accents don't work on Alexa or Google Home*. The Washington Post. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/

Havens, L., Terras, M., Bach, B., & Alex, B. (2020). Situated data, situated systems: A methodology to engage with power relations in natural language processing research. *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, 107–124. Barcelona, Spain (Online): Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.gebnlp-1.10

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–83.

Hovy, D. (2015). Demographic factors improve classification performance. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 752–762. Beijing, China: Association for Computational Linguistics. https://www.aclweb.org/anthology/P15-1073

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., & Hovy, E. (2013). Learning whom to trust with mace. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1120–1130.

Hovy, D., Bianchi, F., & Fornaciari, T. (2020). "you sound just like your father" commercial machine translation systems include stylistic biases. *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, 1686–1690. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.acl-main.154

Hovy, D., & Søgaard, A. (2015). Tagging performance correlates with author age. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 483–488.

Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 591–598. Berlin, Germany: Association for Computational Linguistics. https://www.aclweb.org/anthology/P16-2096

Hovy, D., & Yang, D. (2021). The importance of modeling social factors of language: Theory and practice. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 588–602. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2021.naacl-main.49

Hovy, D., Spruit, S., Mitchell, M., Bender, E. M., Strube, M., & Wallach, H. (Eds.). (2017). *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics. https://www.aclweb.org/anthology/W17-1600

Howard, A., & Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics*, *24*, 1521–1536.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., & Mitchell, M. (2021). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 560–575.

Jebbara, S., & Cimiano, P. (2019). Zero-shot cross-lingual opinion target extraction. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2486–2495. Minneapolis, Minnesota: Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1257

Johannsen, A., Hovy, D., & Søgaard, A. (2015). Cross-lingual syntactic variation over age and gender. *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, 103–112. Beijing, China: Association for Computational Linguistics. https://www.aclweb.org/anthology/K15-1011

Jørgensen, A., Hovy, D., & Søgaard, A. (2015). Challenges of studying and processing dialects in social media. *Proceedings of the Workshop on Noisy User-generated Text*, 9–18.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & 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*, 6282–6293. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.acl-main.560

Jurgens, D., Tsvetkov, Y., & Jurafsky, D. (2017). Writer profiling without the writer's text. In G. L. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.), *Social informatics* (pp. 537–558). Springer International Publishing.

Kennedy, B., Jin, X., Mostafazadeh Davani, A., Dehghani, M., & Ren, X. (2020). Contextualizing hate speech classifiers with post-hoc explanation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5435–5442. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.acl-main.483

Kiritchenko, S., & Mohammad, S. (2018). Examining gender and race bias in two hundred sentiment analysis systems. *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 43–53.

Koolen, C., & van Cranenburgh, A. (2017). These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 12–22. Valencia, Spain: Association for Computational Linguistics. https://www.aclweb.org/anthology/W17-1602

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*, 905–949. https://doi.org/10.1177/0003122419877135

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–172. Florence, Italy: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-3823

Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press.

Lawson, A. D., Harris, D. M., & Grieco, J. J. (2003). Effect of foreign accent on speech recognition in the nato n-4 corpus. *Eighth European Conference on Speech Communication and Technology*.

Liu, H., Wang, W., Wang, Y., Liu, H., Liu, Z., & Tang, J. (2020). Mitigating gender bias for neural dialogue generation with adversarial learning. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 893–903. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.emnlp-main.64

Liu, Z., Shin, J., Xu, Y., Winata, G. I., Xu, P., Madotto, A., & Fung, P. (2019). Zero-shot cross-lingual dialogue systems with transferable latent variables. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1297–1303. Hong Kong, China: Association for Computational Linguistics. https://www.aclweb.org/anthology/D19-1129

Manzini, T., Yao Chong, L., Black, A. W., & Tsvetkov, Y. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 615–621. Minneapolis, Minnesota: Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1062

McCulloch, G. (2020). *Because internet: Understanding the new rules of language*. Riverhead Books.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119.

Mirkin, S., & Meunier, J.-L. (2015). Personalized machine translation: Predicting translational preferences. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2019–2025.

Mirkin, S., Nowson, S., Brun, C., & Perez, J. (2015). Motivating personality-aware machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1102–1108.

Mohammady, E., & Culotta, A. (2014). Using county demographics to infer attributes of twitter users. *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, 7–16.

Munro, R. (2013, May 22). *NLP for all languages*. Idibon Blog. http://idibon.com/nlp-for-all

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.emnlp-main.154

Nissim, M., van Noord, R., & van der Goot, R. (2020). Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, *46*, 487–497. https://www.aclweb.org/anthology/2020.cl-2.7

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. *Proceedings of the 12th Language Resources and Evaluation Conference*, 4034–4043. Marseille, France: European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.497

Nozza, D., Bianchi, F. and Hovy, D. (2021). Measuring hurtful sentence completion in language models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2398–2406. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2021.naacl-main.191

Nozza, D., Bianchi, F. and Hovy, D. (2020) What the [MASK]? Making sense of language-specific BERT models. *arXiv preprint arXiv:2003.02912*.

O'Neil, C. (2016, February 4). *The ethical data scientist*. Slate. http://www.slate.com/articles/technology/future_tense/2016/02/how_to_bring_better_ethics_to_data_science.html

Palakovich, J., Eigeman, J., McDaniel, C. E., Maringas, M., Chodavarapu, S., et al. (2017). Virtual agent proxy in a real-time chat service. *US Patent*, *9*(559), 993.

Passonneau, R. J., & Carpenter, B. (2014). The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, *2*, 311–326.

Paun, S., Carpenter, B., Chamberlain, J., Hovy, D., Kruschwitz, U., & Poesio, M. (2018). Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, *6*, 571–585.

Pavlick, E., Post, M., Irvine, A., Kachaev, D., & Callison-Burch, C. (2014). The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, *2*, 79–92. https://www.aclweb.org/anthology/Q14-1007

Plank, B. (2016). What to do about non-standard (or non-canonical) language in NLP. *Proceedings of the Conference on Natural Language Processing (KONVENS)*, 13–20. Bochumer Linguistische Arbeitsberichte.

Plank, B., Hovy, D., & Søgaard, A. (2014a). Learning part-of-speech taggers with inter-annotator agreement loss. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 742–751.

Plank, B., Hovy, D., & Søgaard, A. (2014b). Linguistically debatable or just plain wrong? *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2:* Short Papers), 507–511. Baltimore, Maryland: Association for Computational Linguistics. https://www.aclweb.org/anthology/P14-2083

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., & Van Durme, B. (2018). Hypothesis only baselines in natural language inference. *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 180–191. New Orleans, Louisiana: Association for Computational Linguistics. https://www.aclweb.org/anthology/S18-2023

Prabhumoye, S., Boldt, B., Salakhutdinov, R., & Black, A. W. (2021). Case study: Deontological ethics in NLP. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3784–3798. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2021.naacl-main.297

Prost, F., Thain, N., & Bolukbasi, T. (2019). Debiasing embeddings for reduced gender bias in text classification. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 69–75. Florence, Italy: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-3810

Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., Nunn, J., Hedayatnia, B., Cheng, M., Nagar, A., et al. (2018). Conversational AI: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4902–4912. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.acl-main.442

Roberts, S. T., Tetreault, J., Prabhakaran, V., & Waseem, Z. (Eds.). (2019). *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-3500

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 1668–1678. Florence, Italy: Association for Computational Linguistics. https://www.aclweb.org/anthology/P19-1163

Saunders, D., & Byrne, B. (2020). Reducing gender bias in neural machine translation as a domain adaptation problem. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7724–7736. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.acl-main.690

Schnoebelen, T. (2013, June 21). *The weirdest languages*. Idibon Blog. http://idibon.com/the-weirdest-languages

Shah, D. S., Schwartz, H. A., & Hovy, D. (2020). Predictive biases in natural language processing models: A conceptual framework and overview. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5248–5264. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.acl-main.468

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3407–3412. Hong Kong, China: Association for Computational Linguistics. https://www.aclweb.org/anthology/D19-1339

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. (2008). Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263. Honolulu, Hawaii: Association for Computational Linguistics. https://www.aclweb.org/anthology/D08-1027

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1679–1684. Florence, Italy: Association for Computational Linguistics.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1630–1640. Florence, Italy: Association for Computational Linguistics. https://www.aclweb.org/anthology/P19-1159

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.

Tan, Y. C., & Celis, L. E. (2019). Assessing social and intersectional biases in contextualized word representations. *Advances in Neural Information Processing Systems*, 13230–13241.

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 53–59. Valencia, Spain: Association for Computational Linguistics. https://www.aclweb.org/anthology/W17-1606

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B., & Poesio, M. (2020). A case for soft loss functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *8*, 173–177.

Van de Poel, I. (2016). An ethical framework for evaluating experimental technology. *Science and Engineering Ethics*, *22*, 667–686.

Wang, J., Li, S., Jiang, M., Wu, H., & Zhou, G. (2018). Cross-media user profiling with joint textual and social user embedding. *Proceedings of the 27th International Conference on Computational Linguistics*, 1410–1420. Santa Fe, New Mexico, USA: Association for Computational Linguistics. https://www.aclweb.org/anthology/C18-1119

Webster, K., Recasens, M., Axelrod, V., & Baldridge, J. (2018). Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, *6*, 605–617. https://www.aclweb.org/anthology/Q18-1042

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6382–6388. Hong Kong, China: Association for Computational Linguistics. https://www.aclweb.org/anthology/D19-1670

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., ... Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhang, B., Huang, H., Pan, X., Ji, H., Knight, K., Wen, Z., Sun, Y., Han, J., & Yener, B. (2014). Be appropriate and funny: Automatic entity morph encoding. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 706–711. Baltimore, Maryland: Association for Computational Linguistics. https://www.aclweb.org/anthology/P14-2115

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K.-W. (2019). Gender bias in contextualized word embeddings. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 629–634. Minneapolis, Minnesota: Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1064

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2979–2989. Copenhagen, Denmark: Association for Computational Linguistics. https://www.aclweb.org/anthology/D17-1323

AUTHOR BIOGRAPHIES

Dirk Hovy is an Associate Professor of Computer Science in the Department of Marketing, and the scientific director of the Data and Marketing Insights research unit at Bocconi University in Milan, Italy. His research focuses on what language can tell us about society, and what computers can tell us about language. He is interested in the interplay of social dimensions of language and NLP models, and the consequences for bias and fairness. His work explores how to integrate sociolinguistic knowledge into NLP models to counteract demographic bias, and was recently awarded a Starting Grant by the European Research Council (ERC) on this topic. Dirk co-founded and organized two editions of the Ethics in NLP workshops, and is a frequent invited speaker on panels on ethics. Website: http://www.dirkhovy.com/

Shrimai Prabhumoye is a PhD student at the Language Technologies Institute at the School of Computer Science, Carnegie Mellon University. Her work focuses on controllable text generation with focus on style, content and structure. She is also exploring the ethical considerations of controllable text generation. She co-designed the Computational Ethics for NLP course at CMU, which was offered for the first time in Spring 2018. Website: https://www.cs.cmu.edu/~sprabhum/

How to cite this article: Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, e12432. https://doi.org/10.1111/lnc3.12432