Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition

Nina Markl
The University of Edinburgh
Edinburgh, Scotland
nina.markl@ed.ac.uk

ABSTRACT

All language is characterised by variation which language users employ to construct complex social identities and express social meaning. Like other machine learning technologies, speech and language technologies (re)produce structural oppression when they perform worse for marginalised language communities. Using knowledge and theories from sociolinguistics, I explore why commercial automatic speech recognition systems and other language technologies perform significantly worse for already marginalised populations, such as second-language speakers and speakers of stigmatised varieties of English in the British Isles. Situating language technologies within the broader scholarship around algorithmic bias, consider the allocative and representational harms they can cause even (and perhaps especially) in systems which do not exhibit predictive bias, narrowly defined as differential performance between groups. This raises the question whether addressing or "fixing" this "bias" is actually always equivalent to mitigating the harms algorithmic systems can cause, in particular to marginalised communities.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; **Speech recognition**.

KEYWORDS

algorithmic bias, speech and language technologies, language variation, speech recognition

ACM Reference Format:

Nina Markl. 2022. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3531146.3533117

1 INTRODUCTION

As has been pointed out in recent years in particular by Black, queer and feminist scholars (e.g. [19, 20, 34, 40, 56, 97]), "algorithmic bias", or as Hampton [56] and Noble [97] put it, "algorithmic oppression" (re)produces existing structures of oppression in a society.[1] Tools frequently discussed in the context of algorithmic oppression often uphold oppressive systems in very direct ways: technologies used in carceral and border systems ("predictive policing and sentencing", facial recognition) or the (uneven) distribution of housing, capital and services (credit allocation, hiring, education, healthcare) [19, 48, 97, 98]. In this paper, I argue that speech and language technologies (SLTs) are an increasingly important site of algorithmic oppression. They are embedded in high-stakes contexts such as hiring [101] and healthcare [83] and ubiquitous in daily technology use (e.g., voice assistants, language models embedded in web search). Their harms are in some ways more pernicious than those of other machine learning tools, especially where they reinforce existing cultural discourses and ideologies about, in particular, marginalised groups and their ways of using language.

Drawing on knowledge from sociolinguistics, I show that marginalised populations are disproportionately affected because language variation, power and social identity are deeply intertwined. Against this background (and exemplifying this dynamic), I evaluate two British English automatic speech recognition (ASR) systems developed by Google and Amazon. Both systems perform substantially worse on second language speakers of English and speakers of some (stigmatised) regional varieties of British English. I explore potential reasons and consequences of this and other types of SLT bias, as well as ways to detect and mitigate it. Then I turn to the limits of discourses of fairness and bias, and the harms even "unbiased" systems can cause, ending on an open question, pertinent to all discussions around algorithmic oppression – when should we attempt to "fix" biased systems and when should we avoid their use altogether?

2 ALGORITHMIC BIAS AND SPEECH AND LANGUAGE TECHNOLOGIES

2.1 Understanding algorithmic bias

"Algorithmic bias" and "algorithmic oppression" are valuable concepts because they highlight that it is the same systems of oppression and socio-technical contexts, and specifically, the same dominant groups within those contexts, which create and facilitate a wide range of technologies harming the same marginalised communities (albeit in different ways). The overarching frame allows us (or forces us) to recognise that origins and consequences of "biased" sociotechnical systems are systemic [56]. To disentangle the different ways in which these underlying structures "show up"

[1] In this paper, I often use the language of "(algorithmic) bias" rather than "(algorithmic) oppression" as this is the term used by most researchers whose work I draw on. Hampton's critique [56] is, however, a crucial intervention in this field and, as will be evident, informs this paper.

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

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533117

in a sociotechnical system, more fine-grained terminology does, however, help. In recent years, several taxonomies to account for the origins [116], types [109] and consequences [13] of algorithmic bias have been proposed.

Speech and language technologies (SLTs), in particular machine-learning based systems designed to process or analyse text or speech, can harm language communities in different ways. As shown in the case study of British English ASR below, SLTs can exhibit "predictive bias" [109], producing systematically higher error rates for some, usually marginalised, groups (e.g., [39, 80]). Harmful outcomes of machine learning systems can be traced back to a variety of points, including sampling and measurement bias during (training) data generation and curation [116], aggregation and learning biases during model building [61, 66], evaluation biases which miss biased behaviours [116], and inappropriate deployment contexts [116]. The consequences for the people affected depends on the application context and the degree to which individual they rely on it, and include "degraded service", a higher risk of adverse decisions in high-stake contexts, or representational harms [13]. It is through these harms, that these technologies (intentionally or not) (re)produce structures of oppression as described by [19, 40, 97]: in addition to mirroring the racist, (cis)sexist, ableist and queerphobic context[2], they also further entrench and strengthen these structures.

2.2 (Social) meaning and context: inherent limitations to speech and language technologies

Like other machine learning systems, SLTs are often (unhelpfully, see [14]) framed, especially by technology corporations, as "solving" a wide range of (social or communicative) "problems". Many of those "language problems", from more abstract tasks like "automatic speech recognition" to concrete applications like "hate speech detection", are extremely challenging to "solve". Without detracting from impressive advances in SLTs in recent years, it is imperative that we not lose sight of the limitations inherent to these tools [26]. Sociolinguistics, the study of language in society, is a useful starting point to understand why "solving language" is so difficult.

Language, both in production and perception, is fundamentally social. All parties to any linguistic interaction are situated in a particular social context which they draw on when expressing and interpreting ideas. Indeed, we use language to convey and construct social meaning in addition to those ideas, both as speakers and as listeners [46]. The social context, social meaning (and, arguably all meaning [17]) are generally not available to SLTs [96]. Some tasks, such as hate speech detection, are very difficult for both algorithmic systems and humans because the specific social context (what is said, by whom, to whom) is crucial [96]. Harmful system behaviours in those cases are not (just) the result of insufficient or biased training data, but of the exceptionally difficult, and perhaps inappropriate, task.

Language is also fundamentally characterised by variation. This variation isn't uniformly distributed across members of a language community, but strongly tied to language users' identities. Since machine learning models generally improve performance with a higher number of training examples, they tend to perform worse for small (sub)populations in a training data set [116]. Even a system trained on a "perfectly representative" language dataset would be prone to make wrongful predictions for numeric minorities.[3] Minoritised and marginalised communities are further often mis- and underrepresented even if they aren't a numeric minority [16].[4]

2.3 Prior work on algorithmic bias in SLTs

[24] show in their survey of 146 papers on "bias in NLP" that discussions of "bias" are often divorced from social, historical and sociolinguistic context. They also often fail to critically engage with how existing power structures shape who does and does not have access to reliable SLTs, and who gets harmed in what ways as a result [24]. Here, I highlight some of the work on algorithmic bias in SLTs, and, following this critique I return to their origins and harms in 5.

2.3.1 Unequal access to SLTs. SLTs are extremely unevenly distributed both across and within languages. There are over 7000 languages in the world [45], only a small subset of which has been integrated in SLTs. [76] find that 88.38% of the 2679 languages whose typology is described in WALS ([44]) are essentially "no resource languages" (see also [21]). They argue that it is "probably impossible" to create SLTs for these languages which are spoken by more than one billion people globally [76, p. 6284]. The seven languages with the most "resources", on the other hand, make up only 0.28% of all languages [76]. The framing of this inequality deployed in [76] as a "race" for language resources with "winners", "left-behinds" and "hopefuls" obscures the (obvious) legacy of colonialism and the effect it has had both on which languages, and more importantly, which ethnic and national groups dominate the world to this day.[5] It is no accident (and certainly not the result of a fair "win") that English, Spanish, German, Japanese and French are five of the seven most "well-resourced" languages spoken by 2.5 billion people globally. The upshot of this distribution is that there are many language communities around the world who have no access to SLTs. Furthermore, because many "high-resource" languages for which SLT architectures are initially developed are typologically similar, they might generalise poorly to those which are currently "under-resourced" [76].

2.3.2 Unequal performance of SLTs. But even for the "high-resource" languages of the world, power shapes which language communities can use SLTs successfully, and which ones may even be harmed by them. Here I focus on English, in part because of my own research background, and in part because it is the most well-researched

[2]To name just a few prevalent structures of oppression. Many people are marginalised in multiple ways which are impossible to disentangle.

[3]As [66] points out, the pervasiveness of "skewed" distributions in the real world, is one of the reasons why careful model development is crucial.

[4]I use the terms "minoritised" and "marginalised" to highlight that these positions are the result of a socio-historical and political process. For example, women and non-binary people make up slightly more than half of the population (globally and in many nations) but are nevertheless marginalised.

[5]The framing of some languages as "left-behind" and fundamentally in need of language resources is particularly problematic. As I discuss below, it may well be that some communities do not want or need these technologies and in any case would like to be actively involved in their creation [22, 23, 65].

context of algorithmic bias in SLTs.[6] However, many of the observations on English also apply to other languages, where SLTs perform much better on a dominant (standard) variety than other varieties and variants used by marginalised communities. This predictive bias can, for example, be seen in sentiment analysis and hate speech detection. [39] show that Google's sentiment analysis tool Perspective API classifies tweets by popular drag queens as "more toxic" than those by white nationalists. Perspective flags tweets containing reclaimed slurs like *gay* and *queer* and "obscene" language in neutral, positive or non-offensive contexts as "'toxic" (see also [42]), but does not account for the fact that "innocuous" words can be used in ways that are deeply hateful. In other words, it doesn't capture the social context of the "obscene" language.[7] Of course, which (or whose) language is considered "obscene" is itself an ideological choice imposed by the hearer [112]. Large language models are prone to reproduce structural oppression in a very direct way by "parroting" biases in the training data [16], for example islamophobic content [1, 43, 86].[8] Similar problems also exist in in machine translation [107] and word embeddings [27] where gender bias proves particularly persistent. Gender neutral nouns or pronouns are often translated reflecting stereotypes or are simply ungrammatical [38, 107]. Machine translation also introduces stylistic bias, where translated text "sound[s] older and more male" than the original [68]. Recent work on US English automatic speech recognition systems show substantial performance differences between Mainstream US English and varieties used by marginalised communities such as African American English (AAE) [80, 90, 126]. [80] find that commercial ASR systems by Apple, IBM, Google, Microsoft and Amazon produce a much higher rate of errors for Black speakers of AAE than comparable White speakers of Californian English. Notably, they also find that error rates were influenced by both gender and race, with particularly high rates for Black men, who tend to use a very high rate of "non-standard" linguistic features in the recordings used in the study (sourced from CORAAL [77]) [80]. This highlights the need for not just disaggregated approaches to SLT evaluation, but specifically intersectional ones, which recognise that interlocking systems of oppression (such as race and gender) cannot be considered separately (as conceptualised in Black feminist thought [36, 62] and applied to other domains of machine learning evaluation [28, 59, 75]). Other work has found predictive bias for regional varieties of English such as Indian English [91], Scottish English and Southern US English [119, 120] (as compared to Mainstream US English). I add to this literature by considering predictive bias in British English commercial ASR systems as it affects first and second language speakers of English.

[6]Again, neither of these facts are an accident: (socio)linguistics, SLT research and related fields have been and continue to be anglo-centric.

[7]A problem not limited to language, several art museums have had images of their exhibits, including the 25,000 year old Venus of Willendorf figurine, flagged as "pornographic": https://www.theguardian.com/artanddesign/2021/oct/16/vienna-museums-open-adult-only-onlyfans-account-to-display-nudes

[8]It is worth noting that broader (potential) harms of large language models, like those related to climate change, misinformation and radicalisation are also likely to disproportionaly affect marginalised groups [16, 127].

3 PATTERNS OF LANGUAGE VARIATION AS PROXIES FOR SOCIAL IDENTITIES

As noted above, language is both inherently social and inherently characterised by variation. This variation may appear random or free when we first encounter it (for example when we enter a new language community). However, as [129] put in a very influential formulation, language variation (and language change) is characterised by "orderly heterogeneity". That is, patterns of language variation are not random, but are highly structured both in individuals and in communities and they can further be used to construct social identities in interaction [47]. As a result, particular linguistic features (or particular combinations of them) can be proxies for social identities. Worse SLT performance for particular language varieties and linguistic features thus often translates to worse performance for particular (usually marginalised) people. In the following section I explain the relationship between language variation and identity before outlining some work on variation in British Englishes, in particular how it relates to power and discrimination.

3.1 Beliefs about language are beliefs about speakers

Language and language variation are always situated in a larger social context. All speaking, writing, signing and listening originates from somewhere[9]. Sociolinguists have long been interested in how particular ways of using language can become associated with specific social identities and positionalities until they become indexical of them (i.e. until they *point to* a particular identity) [47, 73]. In short, as people using language we construct beliefs about language to make sense of the (arbitrary) correlations between particular linguistic forms and the people who use them. Put more precisely, we "locate linguistic phenomena as part of, and evidence for, what [we] believe to be systematic behavioral, aesthetic, affective and moral contrasts among the social groups indexed" [72, p 37]. These ideologies are used to justify and re-entrench particular power structures and construct notions of normativity, markedness, difference and similarity between social groups [35]. Like other ideologies, they can become deeply embedded in our understanding of the world and shape how we produce and interpret language (variation).

Language ideologies can surface in "attitudes about language", often framed as "apolitical", aesthetic preferences for one form over another. But these attitudes about language are almost always reflective of attitudes about the speakers who use them. This is evident in the fact that the same linguistic feature is often interpreted differently depending on who produced it. For example, creaky voice, a phonation type commonly also known as "vocal fry", is among English speakers, much more stigmatised and pathologised in young women's speech than men's [5, 31]. The terms used to evaluate the feature are also evaluations of the women who use them: "annoying", "grating", "too much to bear" [31, 53]. Similarly, linguistic features common in some varieties of British English, such as "glottal replacement" of /t/ in words like *butter* or *Scotland* are stigmatised when used by working class speakers in formal contexts, but interpreted as signalling authenticity and solidarity

[9]Compare Donna Haraway's "god trick of seeing everything from nowhere" [57, p 581] (or everywhere), the illusion of complete, "transcendent" objectivity (in science) which in reality is framed through a particular embodied lens.

when used by upper-class speakers (e.g. politicians) in those very same contexts [79, 111]. Listeners' judgements of speakers (e.g. attractiveness, trustworthiness, friendliness) are also influenced by their perceived race, gender and social class background [9, 41]. These attitudes also have structural implications (see also [35]). [5] ask listeners to rate speakers with and without vocal fry according to their "hireability", and find that those without creaky voice are preferred. This is just one example among many culturally-specific language ideologies around "professional", "educated" and "articulate" speech [11, 87]. In anglophone settings, hiring committees disprefer second language speakers [67, 121] who have also been found to be perceived as "less credible" [84] than first language speakers. Just like algorithmic oppression, language ideologies are not just underpinned by or reflective of structural oppression, but also serve to secure it [104]. It is the language used by powerful social groups in a given societal context (e.g. White, upper and middle class, men) that becomes the "prestigious" or "right" way to speak. [5] conclude that women should avoid creaky voice to avoid discrimination and similar advice is often given to anyone who doesn't speak the "standard variety"[35]. I strongly reject this conclusion - it is listeners (and hiring committees) and language technologies should resist sexist (language) attitudes [31].[10]

3.2 Linguistic variation in the British Isles

The British Isles encompass a lot of linguistic diversity. In addition to English, there are many minoritised languages, including Scottish Gaelic, Scots, Irish, Welsh, Manx, Polish, Punjabi and Urdu, which have different levels of legal recognition within the United Kingdom and Ireland [45][11]. English in the British Isles is also characterised by significant variation, conditioned both by region and social class (see e.g. [51, 70, 130]). This variation is apparent both in dialectal variation (broadly: variation in syntax, morphology and lexicon) and accent variation (variation in pronunciation). Linguists tend to define regional accent or dialect regions along "linguistic borders" (so-called "isoglosses") where two (or more) different ways of expressing the same concept or structure meet. These different pronunciations, words or syntactic structures are often rooted in the distinct historical developments of English in different regions. Especially in the context of accents, these differences are not isolated to individual words, but tend to affect the entire "inventory" of sounds in a particular accent (the "phonology"). For example, accents in the South of Britain distinguish between the vowel in words like *can* and the vowel in words like *can't*, while those in the North generally do not [70]. In addition to these geographical differences in the presence and distribution of particular sounds, speakers also vary in their language use depending on style, context and social class.

In the British Isles (in particular in the UK), the classic example of a highly prestigious accent is Received Pronunciation (RP)[12]. RP, also colloquially referred to as "the Queen's English", is "supra-local": rather than being interpreted as an index of the speaker's geographical origin or identity, it is interpreted as indicative of their social (class) and educational background [3, 49]. It is spoken by a small group of people and was historically particularly widely used in British media and in elite spaces (private schools, politics, aristocracy) [3]. As [3] highlights, the association between RP and upper class status is very strong, and has been reinforced over centuries through prescriptive teaching (inside and outside classrooms) and popular media. Crucially, other "native" accents especially those associated with working class speakers both in urban and rural areas of the north and northeast of England, the Scottish central belt, Wales, and London continue to be stigmatised in many elite spaces and rated as "less prestigious" and "less pleasant" [110][13]. Recent research shows that attitudes towards (some) second language accents have improved, or, at the very least, that increased awareness of the negative effects and arbitrary nature of linguistic discrimination lead study respondents to suppress negative judgments [103]. Nevertheless, accent discrimination and open prejudice against speakers of particular accents (especially second-language speakers) or people who use particular linguistic features appears more socially acceptable than other forms of discrimination. Recent work shows that accent discrimination still plays a role high-prestige hiring contexts such as corporate law, although not all regional accents are equally stigmatised [29, 85]. Accent bias has also been documented in teacher training and schools in the UK, affecting both first and second language speakers of English [11, 37].

4 INVESTIGATING ALGORITHMIC BIAS IN BRITISH ENGLISH ASR SYSTEMS

To add to the understanding of predictive bias in commercial automatic speech recognition systems, I tested the off-the-shelf systems by Google (Google Speech-to-Text) and Amazon (Amazon Transcribe) using two corpora of read speech[14]: the Speech Accent Archive (SAA) [128] and Intonational Variation in English (IViE) [54]. The subset of SAA contains a wide range of first and second language speakers of English, and is useful to illustrate that second language speakers are disadvantaged in the context of ASR systems as compared to first language speakers. IViE allows for a detailed analysis of how speakers of different varieties of British English are impacted by algorithmic bias in ASR.

4.1 Experimental setup

The Speech Accent Archive is a database of short English language speech samples. Each entry consists of a recording of a read elicitation passage which contains most sounds of English, some demographic information about the speaker (binary gender, age, native language and other languages, birthplace, current place of residence, age and mode of acquisition of English), a detailed phonetic transcript and some linguistic analysis. For this experiment, I initially chose a subset of 495 recordings[15] provided by first and second

[10] I'd like to thank an anonymous reviewer for their generous comments regarding the role of the "listening subject" in the context of language technology.

[11] Polish, Punjabi and Urdu are the most common "non-indigenous" languages in the UK, though the list of languages spoken by residents of the UK is, of course, very long and ever-changing.

[12] For simplicity, I use the term RP, rather than Southern Standard British English (SSBE).

[13] See also the Accentism Project which collects personal stories about this experience: https://accentism.org/

[14] Data, code and analysis available at https://github.com/ninamarkl/FAccT22_ASRBias

[15] Accessed here: https://www.kaggle.com/rtatman/speech-accent-archive

language speakers of English (as self-defined by each speaker). Both groups are internally heterogeneous: the first language speakers are from a range of regions in the UK, and all 430 second language speakers speak one of ten randomly chosen languages as a first language (Arabic, French, German, Hindi, Italian, Mandarin, Portuguese, Spanish, Thai or Urdu) and vary in how, and for how long, they have been learning English[16]. To ensure that any differences weren't due to systematic differences in recording quality, signal to noise ratio was measured using Praat [25], and recordings with a measure higher than 50dB were excluded from the subsequent statistical analysis. 445 recordings were retained.

The IViE corpus was collected around 2000 to study intonational variation in the British Isles. It contains audio recordings from 102 adolescent speakers from 9 cities in the British Isles: London, Dublin, Cambridge, Liverpool, Leeds, Bradford, Newcastle, Cardiff and Belfast. The IViE corpus does not contain any information about the speakers, aside from their age (16), binary gender, city and the fact that the speakers from Cardiff and Bradford are bilingual (Welsh and Punjabi, respectively) and the speakers from London are of Caribbean descent. I chose recordings of the speakers reading the first two paragraphs of a longer retelling of the fairy tale Cinderella for analysis (about one minute per speaker). All reference transcripts were validated and, where speakers made speech or reading errors, adjusted by me.

All audio recordings were converted to 16kHz FLAC files, uploaded to Google Cloud and Amazon S3 Storage and processed using their Python APIs. Both corpora were processed with the default models for British English ('en-GB'). The generated transcripts were evaluated against the reference transcripts using `sclite` from the SCTK toolkit[17]. Further analysis of the evaluation outputs was conducted in R and Matlab to compare performance on speaker subgroups within each corpus.

4.2 Results

I employ a mixed methods approach to analysing the experimental results. To quantify the extent of predictive bias experienced by second language speakers of English and speakers of different regional varieties of English, I report word error rate (WER), a standard metric in ASR evaluation for both experiments. I then apply a qualitative error analysis to the results of the IViE experiment to explore the effect of phonetic variation on word error rates.[18]

4.3 Quantitative results

WER is an edit-distance metric presenting the number of errors (deletions, insertions, substitutions) in an automatic transcript relative to the number of correct words in a reference transcript[19]. It is usually computed over an entire test set, often a well-established benchmark. This aggregated approach risks obscuring systematic differences between subsets of the test set. To avoid this, I analyse the WER using multiple linear regression models which include factors such as variety, gender, speech rate and age.

4.3.1 SAA: L1 vs L2 speakers. While SAA also contains information about the speakers' first language and when they started learning English, I decided to focus specifically on sex[20], age, speech rate and L1/L2 status. Error rates varied greatly by L1 (and individual) for both systems, but they were lowest for L1 speakers of English[21]. Age of second language acquisition, as well as phonological characteristics of a first language can influence speakers' accents in many ways which may also impact ASR error rates.[22] However, for the purposes of this paper, I am more interested in highlighting that a wide range speakers whose (potentially only) common characteristic is that they're not "native speakers" of English, likely encounter problems when using these ASR systems. While this category is itself problematic, it is also how many speakers are perceived (and judged) by, in particular, "native speakers".

For both systems, linear regression models show that word error rates are significantly higher for L2 speakers than L1 speakers. Categorical predictors (variety: L1/L2 English, sex: male/female) are deviance coded and numeric predictors (age, speech rate in syllables per second) are scaled and centered. There is a significant main effect for variety at $p<0.05$ for both systems (see Table 1a and Table 1b). Sex is not a significant factor for either model, and adding an interaction term of sex and variety does not improve model fit. Age is a significant factor for Google, with higher error rates for older speakers. Speech rate is a significant factor for Amazon, with higher speech rates corresponding to lower WER[23]. This effect is not observed for Google. Overall, Google produces higher error rates for both speaker groups.

4.3.2 IViE: Variation with British L1 varieties. To investigate the impact of accent variation, I chose the variety with the lowest error rate for each system as the reference levels in linear regression models. Amazon performs best on recordings from Cambridge, while Google performs best on those from London. Categorical predictors (gender: male/female) deviance coded and numeric predictors (speech rate in syllables per second) scaled and centered.[24]

For speakers from Newcastle, Liverpool, Belfast, and Bradford, Amazon produces error rates which are significantly higher than those for speakers from Cambridge ($p<0.05$). There is also a significant main effect of gender, whereby recordings by female speakers show significantly lower error rates ($p<0.05$). Adding an interaction term for gender and variety did not improve model fit. Compared to speakers from London, Google only performs significantly worse for speakers from Belfast ($p<0.05$) (see Fig 1b). There is an interaction effect between variety and gender (which improves model fit):

[16]For each "native language subgroup", I selected up to 70 recordings: recordings are numbered consecutively, so the dataset contains, e.g., files "arabic1" to "arabic70" and "urdu1" to "urdu16" as there are a total of 16 urdu speakers in the sample. See https://github.com/ninamarkl/FAccT22_ASRBias for further details.

[17]https://github.com/usnistgov/SCTK

[18]While this type of analysis would also be appropriate and interesting for the SAA dataset, I focus on IViE here as the speakers in SAA show a lot of individual variation, the analysis of which is outwith the scope of this paper.

[19]Note that despite the name, word error *rate* can be larger than 1 and is conventionally represented in percentages (i.e. WER * 100)

[20]Recordings in the Speech Accent Archive are labelled by "sex" as either male or female. Acknowledging that sex and gender are separate if inter-related social constructs, I assume that "sex" in this context aligns with "gender" for most, if not all, speakers.

[21]See https://github.com/ninamarkl/FAccT22_ASRBias for further details.

[22]As do other sociolinguistic factors not captured in the dataset like residential history, social networks, social class, education, etc.

[23]This counter-intuitive result appears to be the result of exceptionally high WER for with very low speech rates (more than 1 standard deviation away from mean). Commercial ASR systems handle disfluent speech poorly [88].

[24]Recall that speakers were the same age and attended the same school. Other potentially relevant information is not recorded.

Table 1: Speech Accent Archive data: Word Error Rate is significantly (p<0.05) higher for second language speakers of English than first language speakers of English for both ASR systems.

(a) Amazon: SAA – Reference L1 English

Variable	Estimate	Standard Error	t value
(Intercept)	19.58	0.84	23.37
English: L2	5.14	1.71	3.01
Gender: female	-1.53	1.12	-1.36
Speech rate	-4.40	0.60	-7.34
Age	-0.62	0.56	-1.08

(b) Google: SAA – Reference L1 English

Variable	Estimate	Standard Error	t value
(Intercept)	26.55	1.15	23.07
English: L2	7.91	2.34	3.37
Gender: female	-1.45	1.54	-0.94
Speech rate	0.09	0.82	0.11
Age	1.66	0.79	2.10

error rates are significantly higher for women from Belfast, Cardiff and Newcastle (see also Fig 1b).

4.4 Qualitative results: applying context-sensitive evaluation

Quantitative evaluation fails to capture the *context* of errors. WER (as computed above) does not distinguish between different error types (insertion, deletion or substitutions), linguistic contexts (word class, phrase position) or different "triggers" for errors (phonetic variation, speech errors, unusual phrases). WER therefore obscures both the origins and consequences of an error. While architectures vary [93, 131], most speech recognition systems make use of an acoustic model, which contains representations of speech sounds, a dictionary mapping sequences of sounds to words, and a language model, which is used to decode words into longer sequences. Because errors can be the result of a mismatch between the training and test data, the errors we can observe here can be the consequence of under-representation (of a particular pronunciation, turn of phrase or word) in any of those components. To understand origins and impacts of ASR errors, we can qualitatively analyse these errors [88]. The ability to pinpoint which linguistic features "trigger" errors with the help of sociolinguistic expertise could be very useful in developing more robust technologies [126].

4.4.1 Error types. WER considers three types of errors: "substitution errors" where a word is substituted with a wrong transcription, "insertion errors" where the ASR system inserts a word not present in the speech signal, and "deletion errors" where the ASR system fails to transcribe a word. The two systems differ in the distribution of those errors: substitutions are the most common type for both systems while insertion errors are rare, but Google has a much higher deletion rate than Amazon. These patterns are consistent across all speaker groups, which perhaps suggests different model settings. Systematic differences in error type could be problematic as they have distinct impacts on the transcripts. A very high deletion rate can render a transcript useless, in particular as they sometimes appear to cause knock-on effects (see also [89]). Substitution errors can vary in impact: substitutions tend to be phonetically similar (but semantically unrelated) or morphologically related (but not necessarily phonetically similar).

4.4.2 Errors related to phonetic variation. Analysing substitution errors more closely is useful to understand origins and impacts of the errors. We would expect the system to be most accurate for the variety the acoustic model was trained on (or the variety best represented in the training dataset). In addition to simply looking at WER by variety (recall: lowest WER for Cambridge & London), comparing what a speaker actually said to what the system transcribed can also provide clues to varieties the system was trained on. For example, for several of the Belfast speakers' the word *hair*, pronounced by most of them as /hɜːr/ is mis-transcribed as *her*.[25] In RP, the sequence /hɜː(r)/ is indeed most likely *her*, while the actual target *hair* is produced as something like /hɛə/. Transcribing /hɜː(r)/ as *her* is therefore entirely expected if the system was trained on RP. This is just one small example of a systematic difference in the phonology of different varieties, which can lead to predictive bias. This kind of approach could be applied in evaluation on a larger scale by systematically tracing correlations between error rates and sociolinguistic variation [126].

4.4.3 Morphological and syntactic errors. For both systems, substitutions are often morphologically related forms, differing from the target only in number or tense. Sometimes these substitutions are phonetically similar. For example, in more than half of the Google transcripts *lived* in the phrase *Cinders lived with her mother* is substituted with *live*. These errors may reflect differences between connected speech (and in particular, faster speech) and more careful speech a system was trained on. However, some substitutions are quite phonetically distinct. In 47 (Amazon)/41 (Google) of the 102 IViE recordings, the word *would* in *The ball would be held* is replaced with *will*. This error might be introduced by the language model used in the ASR system (for example, because it was trained on text containing mostly present tense verb forms).

5 DISCUSSION

5.1 British English commercial ASR & linguistic hierarchies

The quantitative analysis shows that the performance of Amazon Transcribe and Google Speech-to-Text differs broadly by speaker group, with higher error rates for second language speakers of English, male speakers (Amazon), and speakers of some varieties spoken in the North and Northeast of England (Newcastle, Liverpool, Bradford) and Northern Ireland (Belfast) as compared to L1 speakers, women, and those from the South of England. These differences are particularly notable considering the nature of the

[25] Belfast English (like some other varieties of British English) collapses the distinction between the vowels in *hair* and *her* [130].

Table 2: Word error rates differ by variety. For each system, the variety with the lowest error rate was chosen as the reference level (Amazon: Cambridge, Google: London). Amazon: sig. (p<0.05) worse: Bradford, Liverpool, Newcastle, Belfast; better for women. Google: sig. worse: Belfast.

(a) Amazon: IViE – Reference Cambridge

Variable	Estimate	Std Error	t value
(Intercept)	11.21	1.56	7.19
Cardiff Welsh	4.45	2.50	1.78
Bradford Punjabi	8.84	2.18	4.05
Leeds	1.43	2.23	0.64
Liverpool	6.00	2.20	2.72
London West Indian	4.80	2.39	2.01
Newcastle	5.25	2.19	2.40
Belfast	9.03	2.19	4.13
Dublin	1.18	2.23	0.53
Gender: female	-2.46	1.11	-2.42
Speech rate	0.22	0.63	0.35

(b) Google: IViE – Reference London

Variable	Estimate	Std Error	t value
(Intercept)	35.01	2.42	14.40
Cambridge	-0.07	3.38	-0.02
Cardiff Welsh	-0.12	3.45	-0.03
Bradford Punjabi	0.48	3.41	0.14
Leeds	-5.06	3.41	-1.49
Liverpool	2.82	3.22	0.88
Newcastle	-0.34	3.24	-0.11
Belfast	9.11	3.25	2.80
Dublin	-1.68	3.37	-0.50
Gender: female	-8.72	4.38	-1.99
Speech rate	6.41	0.96	6.67
Cambridge*female	8.31	6.08	1.37
Cardiff*female	15.55	6.83	2.28
Bradford*female	-2.21	6.24	-0.35
Leeds*female	9.68	6.31	1.53
Liverpool*female	10.48	6.12	1.71
Newcastle*female	12.73	6.09	2.09
Belfast*female	28.34	6.09	4.65
Dublin*female	1.35	6.28	0.22

speech data tested: both corpora only contain read speech. This kind of careful speech is generally much easier to process for ASR systems than "real" conversational speech, as it is less affected by phonetic reduction and does not tend to contain hesitations or repetitions [117]. Because all speakers read the same passage, we can isolate differences in pronunciation, speech rate and prosody as the only sources of variation. The small but significant gender gap in Amazon's system with better performance for female speakers echoes findings by [80] in the US. A quantitative analysis of the phonetic features (e.g. vowel quality) of the speakers is outwith the scope of this paper, but prior research in sociolinguistics has found time and again that women tend to avoid dialectal and stigmatised features more than men, speaking "closer to" the standard [81].[26]

Overall, these findings add to the evidence that algorithmic bias not only extends to speech and language technologies but specifically reinforces and reproduces existing linguistic hierarchies and language ideologies. British English ASR systems appear to work best for prestigious varieties such as RP. Conversely, they work worst for second language speakers and speakers of (more or less) stigmatised regional varieties, groups who already experience (linguistic) discrimination [110]. The fact that Google performs best for the London speakers of Caribbean heritage is an interesting complication here, as some varieties within London, especially Multicultural London English (MLE) spoken in particular by younger people of various (minority) ethnic backgrounds, is also subject to some stigma [78, 110].

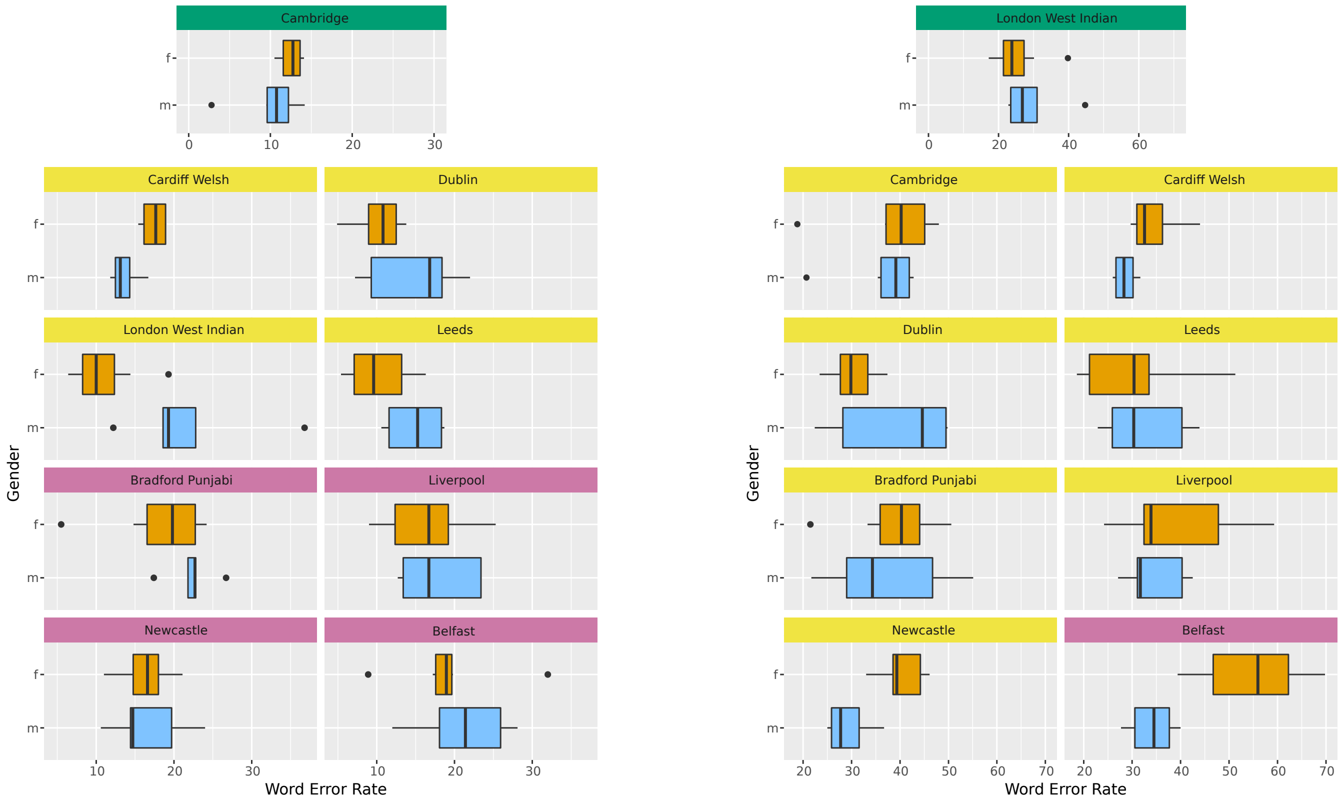
5.2 Specific origins of "bias" in SLTs

My finding that Amazon Transcribe and Google Text-to-Speech perform best for Southern British varieties of English (and L1 speakers), suggests that some regional varieties of British English, especially Northern Ireland and the North of England are under-represented in the training datasets for these systems. Because the lexical content of the recordings was tightly constrained, these biases most likely originate in the training data for the acoustic models. Similarly, in the US context, [80] conclude that the higher error rates for African American English speakers are due to under-representation of AAE in the acoustic models [90] further suggest that some AAE constructions are also under-represented in the language models used by commercial and open-source ASR. They find that Google Speech-to-Text and Mozilla's DeepSpeech produce significantly higher error rates in the vicinity of "habitual be"[27], a feature absent in Mainstream US English, than other uses of "be" [90]. The way training datasets are sourced thus warrants particular attention. Like many commercial machine learning systems, commercial ASR systems are trained on proprietary datasets. Documentation[28] of the voice user interfaces of both Amazon (Alexa) and Google (Google Assistant) suggest that voice data collected from users is part of this training data. Even setting aside any privacy concerns, this reliance on customer data is problematic because customers of large technology corporations who already use an ASR tool are unlikely to be representative of any given language community. According to the British Office for National Statistics, 35% of adults in Great Britain used voice user interfaces in 2020 [50]. The survey only considered variation in age and sex, with younger age groups

[26] Google shows the opposite pattern for some varieties (Belfast, Cardiff, Newcastle) - perhaps as a result of different error types: one recording by a Belfast woman has a WER of 9% for Amazon but 60% for Google with 80% deletion errors.

[27] E.g. AAE "I be at my office at 7.30" is equivalent to Mainstream US English "I am usually at my office at 7.30", see also [55].

[28] https://www.amazon.co.uk/gp/help/customer, https://safety.google/assistant/



(a) Amazon: IViE – Reference Cambridge

(b) Google: IViE – Reference London

Figure 1: Word error rates differ by variety. For each system, the variety with the lowest error rate was chosen as the reference level (Amazon: Cambridge, Google: London) - as indicated by the green panel. Varieties with significantly higher (p<0.05) error rates than the reference level are indicated in pink. (Yellow panels indicate no significant difference).

being much more likely to use these technologies than older ones (and no difference by sex), but similar studies on smartphone and home broadband use from the United States suggest that income is also an important factor [30]. While 96% of households in the UK had broadband access in 2020, speeds vary by region [10], which, among many other impacts, means that users in some geographical regions (e.g. Wales and the Scottish highlands) are probably less likely to make use of devices reliant on cloud-computing (such as ASR systems). This points to a more fundamental problem with commercial SLTs and predictive bias: their market-oriented design. Smaller or marginalised language communities are less likely to be considered valuable markets for technology companies [see also 82]. [19] presents a striking quote by a former engineer working on Apple's ASR systems who was told by a manager that African American English was not being developed because "Apple products are for the premium market". This especially egregious example of a (racist) language ideology linking all speakers of a particular language variety to a particular social and economic position, implying that AAE speakers are not part of the "premium market" emphasises that corporations design technologies with particular users in mind. Communities who aren't perceived as "desireable" markets are less likely to be catered to. In the British context, this means that regional varieties, especially relatively stigmatised ones, are not developed for. This is probably also in part due to the "standard language ideology" [92], the belief that there's an "ideal" or "standardised" variety of the language which is often the most prestigious, or "canonical" form of the language. Importantly, as this ideology becomes part of a "common sense" understanding of the world, so does the belief that all speakers of the language should (strive to) be able to speak the "standard variety", as failing to do so is simply speaking "incorrectly". In the context of SLTs, a consequence of this standard language ideology is that catering for different accents or dialects of the same language is considered less important as speakers are expected to be able to switch to the "standard". While open-source crowdsourced datasets like Mozilla's CommonVoice [7] could in theory be more representative, in practice, they are also unbalanced. This reflects broader issues with crowd- or community-sourced language data (e.g., text) used to train machine learning systems. A well-researched example to illustrate this point is Wikipedia, the text of which is frequently used to train SLTs. Wikipedia is a microcosm of the larger problem [16] identify where marginalised groups are both under- and potentially misrepresented in the text used to train large language models and other SLTs. For example, text by and about women and non-binary people is both under-represented on English language Wikipedia [122], and qualitatively different from entries about men [115, 125]. This leads on the one hand, to a very particular skewed perception of the world by readers, and, much less obviously to people editing or reading Wikipedia, to, for example, "co-occurrence bias" (e.g. the word *family* co-occurs more often with *woman* than *man*) [61], which is reproduced in machine translation systems [107], word embeddings [108], and other tools relying on large language models [97]. This "gender gap" on Wikipedia is an interesting example because it highlights how large structural factors and platform-specific policies and contexts work together to, to some extent inadvertently, create deeply skewed collections of knowledge and text which are then used to build SLTs. The majority of Wikipedia editors are men [6, 58] due to (perceived) skills gaps [58], lack of leisure time and perceptions of the existing (misogynistic or antagonistic) culture on Wikipedia. A seemingly common sense but in practice pernicious Wikipedia policy further favours biographies linking to existing Wikipedia articles [2, 122], making it difficult to add articles about notable people from marginalised communities who have been excluded from mainstream histories. [122] shows that, in addition to only making up 19% of all 1.5 million biographies about inventors, scientists and writers on English language Wikipedia, biographies about notable women are much more likely to be flagged for deletion or disputed on Wikipedia than those about men. The way women are written about is also different with more focus on personal relationships and a high rate of gendered terms (e.g. *feminine*, *woman* etc) [115, 125]. In short, women and non-binary people are under- and misrepresented on Wikipedia, and by extension in the "training data" for many SLTs.

Data bias can also be introduced to training data sets through (usually crowdsourced) annotation. Predictive bias in hate speech detection [39, 42, 106] has been shown to be affected by both data bias and annotation bias. [106] report a higher error rate (erroneous "toxic" label) for African American English phrases than Mainstream US English. They also show that White annotators are more likely than Black annotators to label non-toxic AAE tweets a "toxic" [106]. In the US context, White speakers of "Standard American English" often associate African American English with racist stereotypes [87], while "raciolinguistic" ideologies position African American English as inferior to (white) Mainstream US English [105]. [39] and [42] conclude that the disproportionate associations between "obscene" words and reclaimed slurs and "toxic" labels in training data give rise to the predictive bias they observe.

5.3 Concrete harms of biased SLT

Worse performance on particular language varieties is often equivalent to worse performance for particular communities. Because SLTs tend to be trained on "prestigious" varieties, such as Standard English and RP, these communities are likely to be already marginalised. In this way SLTs further cement the power of "high-status" varieties (and speakers) – and contribute to the devaluation of all other varieties (and speakers). Availability of SLTs in powerful varieties could even accelerate language shift in some domains, such as computing and digital media [63].

Allocative harms [116], or "adverse decisions" [13] of biased SLTs are still understudied. But SLTs are now commonly embedded in high stakes contexts where bias could prevent a language community from being allocated resources. Automatic speech recognition systems are part of complex algorithmic systems used to automatically rank job application videos [101]. Predictive bias of the kind shown in this paper, with worse performance for second language speakers or speakers of stigmatised varieties, could have very concrete negative consequences for marginalised applicant groups. As voice user interfaces become increasingly important tools to access services (e.g. banking and customer service) and devices, predictive bias in ASR can both "degrade service[s]" [13] and actively deny services. Predictive bias in hate speech detection can also represent an allocative harm, if users are unfairly excluded from, for example, social media platforms because a post was erroneously flagged as "toxic". While the prescriptive linguistic rules of social media platforms which are enforced by algorithmic tools, have been shown to inspire (fascinating) new language practices to avoid penalties [114, 123], creative resistance to algorithmic systems is not enough to avoid harm.

Stereotypical, discriminatory and hateful predictions in the context of language about communities in natural language generation tasks are representational harms [13, 86, 116]. For example, the islamophobic predictions by GPT-3 documented in [1] are both harmful to any Muslim users who see the offensive output and feed into (and, of course, result from) much larger islamphobic discourses harming all Muslims. Similarly, machine translation systems which reproduce gender and racial bias, can harm the immediate users and further perpetuate sexist and racist discourses [38, 107]. Hate speech detection tools also create representational harm to marginalised groups when they fail to detect actually hateful content, as highlighted in [39] by not "protecting" marginalised users from this content, while simultaneously reinforcing existing ideologies about, for example, what is and is not "obscene" language [112].

5.4 Mitigating bias in SLTs

Mitigation of data bias is an active field of research, but there are fundamental limitations to the extent to which data can be "debiased". Data is often likened to a "natural resource" [113] and specifically often (positively) compared to oil, due to its central role in many societies and economies. As [118] points out, the metaphor is also apt because machine learning is fundamentally extractive, and, I would add, just like the fossil fuel industry it does a lot of harm, in particular to marginalised populations. Unlike (unrefined) oil, however, data is not a "naturally occurring resource" that can simply be "collected" [18] – it is socially constructed [40, 99]. As such, it can never be fully free of "bias" (see also [60]). It is thus crucial that we account for all (structural) factors that have shaped a dataset, and highlight the labour that went into it [18, 40]. This includes where, how, when, why, about whom or what, by whom and for whom the dataset was compiled and, where relevant, who annotated it [52]. Understanding datasets as infrastructure and prioritising good documentation is a first step to mitigating harms of biased systems as it allows us to anticipate them [15, 71, 99].

Despite inherent limitations, "debiasing" approaches are increasingly popular in the field and have seen some success. In the context of hate speech detection [42] show that relatively simple interventions such as adding non-toxic examples of identity terms which are associated with disproportionate levels of toxicity such as"gay". While far from perfect, the Mozilla's CommonVoice corpora do cover a much broader set of varieties, including some minoritised languages which do not have broad SLT support. Recent work in ASR furthermore shows that different model setups [74, 124] can mitigate accent bias. [108] show that recent efforts in increasing the representation of women on Wikipedia has had some limited positive effect on bias in word embeddings. [106] highlight that annotator bias can be limited when annotators are informed about sociolinguistic variation before labelling data. Other research on crowd-sourcing also highlights the importance of balancing annotators from diverse backgrounds and accounting for variation in values and opinions [12, 69]. [43] highlight that filtering large text corpora to mitigate bias is very challenging, especially because using a simple list of "bad words" risks removing all kinds of "non-toxic" uses of those terms by, in particular, marginalised groups (e.g. reclaimed slurs, other "obscene language"). As suggested in 4.2, better evaluation strategies of SLTs are also central to mitigating harms. An intersectional approach to quantitative evaluation can identify predictive bias which may be missed by more "aggregated" techniques. Furthermore adding qualitative methods allows us to pinpoint the exact implications, and sometimes, causes of undesirable, "biased" system behaviours.

Finally, it should be noted, as raised by [66], "not everything is a data problem". While imbalances in training and test datasets are one important source of predictive bias in machine learning, particular model structures can amplify or even introduce biases [66]. But focusing on the *data* we use is crucial, especially as concerns about "data bias" are often dismissed (including by senior figures in the field [116]) as "non-technical" issues which fall (by extension perhaps) outwith the remit of machine learning engineers. Within this perspective, predictions of machine learning models which reflect, reproduce or actively worsen structural oppression are not considered an injustice (or even erroneous), but the "neutral" consequences of "accurately" reflecting the (racist, sexist, queerphobic etc.) world as it exists. But of course, we can (and, I would argue, should) instead choose to build technologies that work towards more liberated futures [34, 40]. Creating more representative datasets and carefully selecting models can limit harms of machine learning technologies. Another important consideration is whether to deploy (machine learning) technology at all [14].

5.5 Limits of bias and fairness

While the recent interest algorithmic oppression (or "bias") in machine learning, including SLT applications, is a step in the right direction, there are fundamental limits to these discourses of bias and fairness [20, 56, 64]. As noted by [56], the framing of "bias" elides the structural nature of the issue and decouples it from power and oppression. Shifting our attention from "bias" to both *oppression* and concrete *harms* of algorithmic systems, allows us to account for power, forces us to reflect on our normative position towards harms, and focus on people's lived experiences [24]. It also allows us to expand the discussion to include harms of systems which aren't "biased" in a narrow sense. For example, an algorithmic system ranking job applicants based on "voice data" may not produce a higher rate of transcription errors for one group of speakers, but still applies language ideologies about "professional speech" when ranking applicants. This is particularly pernicious as, contrary to the standard language ideology, not everyone has equal access to these "right" ways of speaking (e.g. through education). Other seemingly "trivial" or "harmless" SLT applications like grammar checkers also encode deeply harmful language ideologies about "good" or "correct" language use without showing any predictive bias in a narrow sense. As [4] notes, those who can make themselves "legible" to the algorithm (e.g. by using the right words), according to the model of the world (or language) it has constructed will succeed (and continue to do so). Arguably, even hate speech detection tools flagging positive uses of (reclaimed) slurs as "toxic" are not "wrong" – they just lack access to the social context that licences only some people to use particular words depending on their positionality[29]. That does not mean that excluding users or removing content based on those decisions is not harmful.

Aside from reinforcing (linguistic) ideologies, entirely "unbiased" systems can also be applied in deeply harmful ways. Right now, Amazon's ASR, machine translation and other natural language processing tools are being used to facilitate surveillance[30] of incarcerated people (and their contacts) in the United States [8]. Arguably, the harms resulting from this use of SLTs are particularly large if these systems exhibit predictive bias as they could result in, for example, criminal investigation based on incorrect transcripts[31]. On the other hand, "fixing" or mitigating predictive bias in this context, risks rendering marginalised populations even more legible [4] (to the state and to corporations) against their wishes (or at least, without their consent). As [100] put it: "[b]ias is real, but it's also a captivating diversion". Even when focusing on the harms of predictive bias of an algorithmic system, we risk overlooking the harms of the algorithmic system, full stop. Sometimes, conversations about the technical challenges or even the social contexts of "bias" (such as this paper), distract us from perhaps more urgent political conversations about the kinds of technologies we want to build and the kinds of futures we want these technologies to exist in. One way of integrating these conversations in the development of not just "fairer" but fundamentally (more) just SLTs [94] is to actively involve language communities. Participatory, community-led approaches to the development of speech and language infrastructure (such as Masakhane [95]) could be particularly beneficial for smaller or marginalised language communities, which are often overlooked or purposefully excluded by large technology corporations and academic institutions. Participatory and community-oriented approaches to data creation, curation and compilation are a way to ensure that ownership of the data stays with the community [18, 33]. Crucially, it would also give them a say in how their language(s) are represented and space for refusal [32].

[29]Note that this is distinct from the high toxicity scores for entirely "neutral" terms like *gay*, *trans* or *lesbian*, which does constitute predictive bias.
[30]Leo Technologies frames their services as safeguarding inmates, but [8] report that they are also being used to monitor "conversations involving mention of the Spanish word for lawyer or accusations that detention facilities were covering up COVID-19 outbreaks"
[31]Incorrect court transcriptions of speakers of non-standard varieties are also a dangerous problem of human transcribers [102]

6 CONCLUSION

Like other machine learning technologies, SLTs (re)produce the structures of oppression which shape the contexts in which they are designed and deployed, as a form of algorithmic oppression. SLTs tend to be designed for and by (language) communities which hold more power (within and across societies). As a result they are not only less useful for marginalised communities, but because of the complex interaction of (social) meaning, social context and identity, SLTs can inflict allocative and representational harms on marginalised communities. For example, as this paper shows, British English commercial ASR performs significantly worse for second language speakers of English and speakers of regional non-standard accents. Beyond this predictive bias, SLTs can also entrench existing ideologies about communities and the linguistic varieties they speak. Shifting our focus to the experiences of, in particular marginalised, people who use SLTs (or to whom they are applied) forces us to think carefully about what to do with "biased" systems and invites us to actively involve them in or let them lead technology design.

ACKNOWLEDGMENTS

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences. I'd like to thank Catherine Lai, Lauren Hall-Lew and three anonymous reviewers for their generous feedback. Special thanks to Lukas Eigentler for help with Matlab.

REFERENCES

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

[2] Julia Adams, Hannah Brückner, and Cambria Naslund. 2019. Who Counts as a Notable Sociologist on Wikipedia? Gender, Race, and the "Professor Test". *Socius* 5 (2019), 2378023118823946. https://doi.org/10.1177/2378023118823946

[3] Asif Agha. 2003. The social life of cultural value. *Language & Communication* 23, 3 (2003), 231–273. https://doi.org/10.1016/S0271-5309(03)00012-0 Words and Beyond: Linguistic and Semiotic Studies of Sociocultural Order.

[4] Ali Alkhatib. 2021. To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 95, 9 pages. https://doi.org/10.1145/3411764.3445740

[5] Rindy C Anderson, Casey A Klofstad, William J Mayew, and Mohan Venkatachalam. 2014. Vocal fry may undermine the success of young women in the labor market. *PloS one* 9, 5 (2014), e97506–e97506.

[6] Judd Antin, Raymond Yee, Coye Cheshire, and Oded Nov. 2011. Gender Differences in Wikipedia Editing. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration* (New York, NY, USA) *(WikiSym '11)*. Association for Computing Machinery, 11–14. https://doi.org/10.1145/2038558.2038561

[7] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. 4211–4215.

[8] Avi Asher-Schapiro and David Sherfinski. 2021. U.S. prisons are installing AI-powered surveillance to fight crime, documents seen by the Thomson Reuters Foundation show, but critics say privacy rights are being trampled. *Thomson Reuters Foundation News* (Nov 2021). https://news.trust.org/item/20211115095808-kq7gx/

[9] Melissa M. Baese-Berk, Drew J. McLaughlin, and Kevin B. McGowan. 2020. Perception of Non-Native Speech. *Language and Linguistics Compass* 14, 7 (2020), e12375. https://doi.org/10.1111/lnc3.12375

[10] Carl Baker. 2021. Constituency data: broadband coverage and speeds. *House of Commons Library: Data Dashboard* (2021). https://commonslibrary.parliament.uk/constituency-data-broadband-coverage-and-speeds/

[11] Alex Baratta. 2017. Accent and linguistic prejudice within British teacher training. *Journal of Language, Identity & Education* 16, 6 (2017), 416–423. https://doi.org/10.1080/15348458.2017.1359608

[12] Natã M. Barbosa and Monchu Chen. 2019. Rehumanized Crowdsourcing: A Labeling Framework Addressing Bias and Ethics in Machine Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow Scotland UK). ACM, 1–12. https://doi.org/10.1145/3290605.3300773

[13] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[14] Eric P.S. Baumer and M. Six Silberman. 2011. When the Implication Is Not to Design (Technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA) *(CHI '11)*. Association for Computing Machinery, 2271–2274. https://doi.org/10.1145/1978942.1979275

[15] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (12 2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[16] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[17] Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

[18] Garfield Benjamin. 2021. What We Do with Data: A Performative Critique of Data 'Collection'. *Internet Policy Review* 10, 4 (2021). https://doi.org/10.14763/2021.4.1588

[19] Ruha Benjamin. 2019. *Race after technology: Abolitionist tools for the New Jim Code.* Polity Press, Newark.

[20] Cynthia L. Bennett and Os Keyes. 2020. What is the Point of Fairness? Disability, AI and the Complexity of Justice. *SIGACCESS Access. Comput.* 125, Article 5 (March 2020), 1 pages. https://doi.org/10.1145/3386296.3386301

[21] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic Speech Recognition for Under-Resourced Languages: A Survey. *Speech Communication* 56 (2014), 85–100. https://doi.org/10.1016/j.specom.2013.07.008

[22] Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 3504–3519. https://doi.org/10.18653/v1/2020.coling-main.313

[23] Abeba Birhane. 2020. Algorithmic Colonization of Africa. *SCRIPTed* 17, 2 (Aug. 2020), 389–409. https://doi.org/10.2966/scrip.170220.389

[24] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, Online, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[25] Paul Boersma. 2002. Praat, a system for doing phonetics by computer. *Glot international* 5 (2002).

[26] Meredith Broussard. 2019. *Artificial Unintelligence: How Computers Misunderstand the World.* The MIT Press.

[27] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the Origins of Bias in Word Embeddings. In *International Conference on Machine Learning*. PMLR, 803–811. http://proceedings.mlr.press/v97/brunet19a.html

[28] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st conference on fairness, accountability and transparency (Proceedings of machine learning research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[29] Amanda Cardoso, Erez Levon, Devyani Sharma, Dominic Watt, and Yang Ye. 2019. Inter-speaker variation and the evaluation of British English accents in employment contexts. In *Proceedings of the International Congress of Phonetic Sciences*. 1615–1619.

[30] Pew Research Centre. 2021. *Internet/Broadband Fact Sheet*. Technical Report. Pew Research Centre. https://www.pewresearch.org/internet/fact-sheet/internet-broadband/

[31] Monika Chao and Julia R. S. Bursten. 2021. Girl talk: Understanding negative reactions to female vocal fry. *Hypatia - A Journal of Feminist Philosophy* 36, 1 (2021), 42–59. https://doi.org/10.1017/hyp.2020.55 Publisher: Cambridge University Press.

[32] M. Cifor, P. Garcia, T.L. Cowan, J. Rault, T. Sutherland, A. Chan, J. Rode, A.L. Hoffmann, N. Salehi, and L. Nakamura. 2019. Feminist Data Manifest-No. (2019). https://www.manifestno.com

[33] Donavyn Coffey. 2021. Māori are trying to save their language from Big Tech. *Wired* (April 2021). https://www.wired.co.uk/article/maori-language-tech
[34] Sasha Costanza-Chock. 2020. *Design Justice.* MIT Press. https://design-justice.pubpub.org/
[35] Justin T. Craft, Kelly E. Wright, Rachel Elizabeth Weissler, and Robin M. Queen. 2020. Language and Discrimination: Generating Meaning, Perceiving Identities, and Discriminating Outcomes. *Annual Review of Linguistics* 6, 1 (2020), 389–407. https://doi.org/10.1146/annurev-linguistics-011718-011659
[36] Kimberle Crenshaw. 1991. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review* 43, 6 (1991), 1241–1299. https://doi.org/10.2307/1229039
[37] Ian Cushing and Julia Snell. 2022. The (White) Ears of Ofsted: A Raciolinguistic Perspective on the Listening Practices of the Schools Inspectorate. *Language in Society* (2022), 1–24. https://doi.org/10.1017/S0047404522000094
[38] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M Phillips, and Kai-Wei Chang. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. arXiv:2108.12084 [cs.CL]
[39] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture* 25, 2 (2021), 700–732. https://doi.org/10.1007/s12119-020-09790-w
[40] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism.* The MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001
[41] John A. Dixon, Berenice Mahoney, and Roger Cocks. 2002. Accents of Guilt?: Effects of Regional Accent, Race, and Crime Type on Attributions of Guilt. *Journal of Language and Social Psychology* 21, 2 (2002), 162–168. https://doi.org/10.1177/02627X02021002004
[42] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 67–73. https://doi.org/10.1145/3278721.3278729
[43] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Online and Punta Cana, Dominican Republic). Association for Computational Linguistics, 1286–1305. https://aclanthology.org/2021.emnlp-main.98
[44] Matthew S. Dryer and Martin Haspelmath (Eds.). 2013. *WALS Online.* Max Planck Institute for Evolutionary Anthropology, Leipzig. https://wals.info/
[45] David M. Eberhard, Gary F. Simons, and Charles D. Fennig (Eds.). 2021. *Ethnologue: Languages of the World* (24th edition ed.). SIL International, Dallas. http://www.ethnologue.com
[46] Penelope Eckert. 2008. Variation and the Indexical Field. *Journal of Sociolinguistics* 124 (2008), 453–476.
[47] Penelope Eckert. 2012. Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annual Review of Anthropology* (2012), 87–100. Issue June. https://doi.org/10.1146/annurev-anthro-092611-145828
[48] Virginia Eubanks. 2018. *Automating inequality : how high-tech tools profile, police, and punish the poor.* St. Martin's Press, New York, NY.
[49] Anne H. Fabricius. 2018. *Social Change, Linguistic Change and Sociolinguistic Change in Received Pronunciation.* Palgrave Macmillan UK, London, 35–66. https://doi.org/10.1057/978-1-137-56288-3_3
[50] Office for National Statistics. 2020. *Internet access – households and individuals, Great Britain: 2020.* Technical Report. Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2020
[51] Paul Foulkes and Gerald J. Docherty. 1999. *Urban Voices: Accent Studies in the British Isles.* Arnold–Oxford UP, London, England–New York, NY.
[52] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (nov 2021), 86–92. https://doi.org/10.1145/3458723
[53] Ira Glass. 2015. If you don't have anything nice to say, SAY IT IN ALL CAPS: Freedom fries. *This American Life* 545 (Jan 2015). https://www.thisamericanlife.org/545/if-you-dont-have-anything-nice-to-say-say-it-in-all-caps
[54] Esther Grabe and Francis Nolan. 2002. The IViE Corpus: English Intonation in the British Isles. http://www.phon.ox.ac.uk/files/apps/old_IViE/
[55] Lisa J. Green. 2002. *African American English: A linguistic introduction.* Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511800306
[56] Lelia Marie Hampton. 2021-03-03. Black Feminist Musings on Algorithmic Oppression. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event Canada). ACM, 1–11. https://doi.org/10.1145/3442188.3445929
[57] Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14, 3 (1988), 575–599. http://www.jstor.org/stable/3178066
[58] Eszter Hargittai and Aaron Shaw. 2015. Mind the Skills Gap: The Role of Internet Know-How and Gender in Differentiated Contributions to Wikipedia. *Information, Communication & Society* 18, 4 (2015), 424–442. https://doi.org/10.1080/1369118X.2014.957711
[59] Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. 2021. Unpacking the Interdependent Systems of Discrimination: Ableist Bias in NLP Systems through an Intersectional Lens. In *Findings of the Association for Computational Linguistics: EMNLP 2021.* Association for Computational Linguistics, Punta Cana, Dominican Republic, 3116–3123. https://aclanthology.org/2021.findings-emnlp.267
[60] Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2020. Situated data, situated systems: A methodology to engage with power relations in natural language processing research. In *Proceedings of the second workshop on gender bias in natural language processing.* Association for Computational Linguistics, Barcelona, Spain (Online), 107–124. https://www.aclweb.org/anthology/2020.gebnlp-1.10
[61] Thomas Hellström, Virginia Dignum, and Suna Bensch. 2020. Bias in Machine Learning What is it Good (and Bad) for? *CoRR* abs/2004.00686 (2020). arXiv:2004.00686 https://arxiv.org/abs/2004.00686
[62] Patricia Hill Collins. 2000 [1990]. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment* (second ed.). Routledge.
[63] Amanda M. Hilmarsson-Dunn and Ari P. Kristinsson. 2009. Iceland's language technology: policy versus practice. *Current Issues in Language Planning* 10, 4 (2009), 361–376. https://doi.org/10.1080/14664200903554966 arXiv:https://doi.org/10.1080/14664200903554966
[64] Anna Lauren Hoffmann. 2019. Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse. *Information, Communication & Society* 22, 7 (2019), 900–915. https://doi.org/10.1080/1369118X.2019.1573912
[65] Anna Lauren Hoffmann. 2021. Terms of Inclusion: Data, Discourse, Violence. *New Media & Society* 23, 12 (2021), 3539–3556. https://doi.org/10.1177/1461444820958725
[66] Sara Hooker. 2021. Moving beyond "algorithmic bias is a data problem". *Patterns* 2, 4 (April 2021), 100241. https://doi.org/10.1016/j.patter.2021.100241 Publisher: Elsevier BV.
[67] Megumi Hosoda and Eugene Stone-Romero. 2010. The effects of foreign accents on employment-related decisions. *Journal of Managerial Psychology* 25, 2 (Feb. 2010), 113–132. https://doi.org/10.1108/02683941011019339 Publisher: Emerald.
[68] Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Online, 1686–1690. https://doi.org/10.18653/v1/2020.acl-main.154
[69] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow Scotland UK). ACM, 1–12. https://doi.org/10.1145/3290605.3300637
[70] Arthur Hughes, Peter Trudgill, and Dominic Watt. 2013. *English Accents and Dialects.* Routledge. https://doi.org/10.4324/9780203784440
[71] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 560–575. https://doi.org/10.1145/3442188.3445918
[72] J. T. Irvine and S. Gal. 2000. Language ideology and linguistic differentiation. In *Regimes of language: Ideologies, polities, and identities*, P. V. Kroskrity (Ed.). School of American Research Press, Santa Fe, 35–84.
[73] Alexandra Jaffe. 2016. Indexicality, Stance and Fields in Sociolinguistics. In *Sociolinguistics: Theoretical Debates*, Nikolas Coupland (Ed.). Cambridge University Press, 86–112. https://doi.org/10.1017/CBO9781107449787.005
[74] Abhinav Jain, Vishwanath P. Singh, and Shakti P. Rath. 2019. A Multi-Accent Acoustic Model Using Mixture of Experts for Speech Recognition. In *Proc. Interspeech 2019.* 779–783. https://doi.org/10.21437/Interspeech.2019-1667
[75] May Jiang and Christiane Fellbaum. 2020. Interdependencies of gender and race in contextualized word embeddings. In *Proceedings of the second workshop on gender bias in natural language processing*. Association for Computational Linguistics, Barcelona, Spain (Online), 17–25. https://www.aclweb.org/anthology/2020.gebnlp-1.2
[76] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560
[77] Tyler Kendall and Charlie Farrington. 2021. The Corpus of Regional African American Language. https://doi.org/10.35111/EXQ3-X930

[78] Ruth Kircher and Sue Fox. 2021. Multicultural London English and its speakers: a corpus-informed discourse study of standard language ideology and social stereotypes. *Journal of Multilingual and Multicultural Development* 42, 9 (2021), 792–810. https://doi.org/10.1080/01434632.2019.1666856

[79] Sam Kirkham and Emma Moore. 2016. Constructing social meaning in political discourse: Phonetic variation and verb processes in Ed Miliband's speeches. *Language in Society* 45, 1 (2016), 87–111. https://doi.org/10.1017/S0047404515000755 Publisher: Cambridge University Press.

[80] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (April 2020), 7684–7689. https://doi.org/10.1073/pnas.1915768117

[81] William Labov. 1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2, 2 (1990), 205–254. https://doi.org/10.1017/S0954394500000338

[82] Halcyon M. Lawrence. 2021. Siri Disciplines. In *Your Computer Is on Fire*, Thomas S. Mullaney, Benjamin Peters, Mar Hicks, and Kavita Philip (Eds.). The MIT Press, 179–198. https://doi.org/10.7551/mitpress/10993.003.0013

[83] Dave Lee. 2021. The next Big Tech Battle: Amazon's Bet on Healthcare Begins to Take Shape. *Financial Times* (2021). https://www.ft.com/content/fa7ff4c3-4694-4409-9ca6-bfadf3a53a62

[84] Shiri Lev-Ari and Boaz Keysar. 2010. Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology* 46, 6 (2010), 1093–1096. https://doi.org/10.1016/j.jesp.2010.05.025

[85] Erez Levon, Devyani Sharma, Dominic J. L. Watt, Amanda Cardoso, and Yang Ye. 2021. Accent Bias and Perceptions of Professional Competence in England. *Journal of English Linguistics* 49, 4 (2021), 355–388. https://doi.org/10.1177/00754242211046316

[86] Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing Stereotyping Biases via Underspecified Questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (Online). Association for Computational Linguistics, 3475–3489. https://doi.org/10.18653/v1/2020.findings-emnlp.311

[87] Rosina Lippi-Green. 2012. *English with an accent language, ideology, and discrimination in the United States.* Routledge, London ; New York.

[88] Nina Markl and Catherine Lai. 2021. Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, Online, 34–40. https://aclanthology.org/2021.hcinlp-1.6

[89] Joshua L Martin. 2021. Spoken corpora data, automatic speech recognition, and bias against african american language: The case of habitual 'be'. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (FAccT '21).* Association for Computing Machinery, New York, NY, USA, 284. https://doi.org/10.1145/3442188.3445893 Number of pages: 1 Place: Virtual Event, Canada.

[90] Joshua L. Martin and Kevin Tang. 2020. Understanding racial disparities in automatic speech recognition: The case of habitual "be". In *Proc. Interspeech 2020.* 626–630. https://doi.org/10.21437/Interspeech.2020-2893

[91] Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell. 2020. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of the 12th language resources and evaluation conference.* European Language Resources Association, Marseille, France, 6462–6468. https://www.aclweb.org/anthology/2020.lrec-1.796

[92] James Milroy. 2001. Language Ideologies and the Consequences of Standardization. *Journal of Sociolinguistics* 5, 4 (2001), 530–555. https://doi.org/10.1111/1467-9481.00163

[93] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. 2019. Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access* 7 (2019), 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880

[94] Julia Nee, Genevieve Macfarlane Smith, Alicia Sheares, and Ishita Rustagi. 2021. Advancing Social Justice through Linguistic Justice: Strategies for Building Equity Fluent NLP Technology. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (New York, NY, USA) *(EAAMO '21).* Association for Computing Machinery, 1–9. https://doi.org/10.1145/3465416.3483301

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

[96] Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On Learning and Representing Social Meaning in NLP: A Sociolinguistic Perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Online). Association for Computational Linguistics, 603–612. https://doi.org/10.18653/v1/2021.naacl-main.50

[97] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York University Press.

[98] Cathy O'Neil. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Penguin Books.

[99] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and Its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research. *Patterns* 2, 11 (2021), 100336. https://doi.org/10.1016/j.patter.2021.100336

[100] Julia Powles and Helen Nissenbaum. 2018. The Seductive Diversion of 'Solving' Bias in Artificial Intelligence. https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53

[101] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20).* Association for Computing Machinery, New York, NY, USA, 469–481. https://doi.org/10.1145/3351095.3372828

[102] John R. Rickford and Sharese King. 2016. Language and Linguistics on Trial: Hearing Rachel Jeantel (and Other Vernacular Speakers) in the Courtroom and Beyond. *Language* 92, 4 (2016), 948–988. https://doi.org/10.1353/lan.2016.0078

[103] Janin Roessel, Christiane Schoel, and Dagmar Stahlberg. 2020. Modern Notions of Accent-ism: Findings, Conceptualizations, and Implications for Interventions and Research on Nonnative Accents. *Journal of Language and Social Psychology* 39, 1 (2020), 87–111. https://doi.org/10.1177/0261927X19884619

[104] Jonathan Rosa and Christa Burdick. 2016. Language Ideologies. In *Oxford Handbook of Language and Society*, Ofelia García, Nelson Flores, and Massimiliano Spotti (Eds.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190212896.013.15

[105] Jonathan Rosa and Nelson Flores. 2017. Unsettling Race and Language: Toward a Raciolinguistic Perspective. *Language in Society* 46, 5 (2017), 621–647. https://doi.org/10.1017/S0047404517000562

[106] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics.* Association for Computational Linguistics, Florence, Italy, 1668–1678. https://doi.org/10.18653/v1/P19-1163

[107] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. arXiv:2104.06001 [cs.CL]

[108] Katja Geertruida Schmahl, Tom Julian Viering, Stavros Makrodimitris, Arman Naseri Jahfari, David Tax, and Marco Loog. 2020. Is Wikipedia succeeding in reducing gender bias? Assessing changes in gender bias in Wikipedia using word embeddings. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science.* Association for Computational Linguistics, Online, 94–103. https://doi.org/10.18653/v1/2020.nlpcss-1.11

[109] Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th annual meeting of the association for computational linguistics.* Association for Computational Linguistics, Online, 5248–5264. https://doi.org/10.18653/v1/2020.acl-main.468

[110] Devyani Sharma, Erez Levon, and Yang Ye. 2022. 50 years of British accent bias: Stability and lifespan change in attitudes to accents. *English World-Wide* (2022). https://doi.org/10.1075/eww.20010.sha

[111] Jennifer Smith and Sophie Holmes-Elliott. 2018. The unstoppable glottal: tracking rapid change in an iconic British variable. *English Language and Linguistics* 22, 3 (2018), 323–355. https://doi.org/10.1017/S1360674316000459

[112] Arthur K Spears. 1998. African-American Language Use: Ideology and so-Called Obscenity. In *African-American English*, Guy Bailey, John Baugh, Salikoko S. Mufwene, and John R. Rickford (Eds.). Routledge, 240–264.

[113] Luke Stark and Anna Lauren Hoffmann. 2019. Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture. *Journal of Cultural Analytics* 1, 1 (2019), 11052. https://doi.org/10.22148/16.036

[114] Ian Stewart, Stevie Chancellor, Munmun De Choudhury, and Jacob Eisenstein. 2018. #Anorexia, #anarexia, #anarexyia: Characterizing Online Community Practices with Orthographic Variation. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017* (2018), 4353–4361. https://doi.org/10.1109/BigData.2017.8258465

[115] Jiao Sun and Nanyun Peng. 2021. Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia. In *Proceedings of the 59th Annual Meeting*

of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (Online). Association for Computational Linguistics, 350–360. https://doi.org/10.18653/v1/2021.acl-short.45

[116] Harini Suresh and John V. Guttag. 2021. A Framework for Understanding Unintended Consequences of Machine Learning. *CoRR* abs/1901.10002v4 (2021). arXiv:1901.10002v4 http://arxiv.org/abs/1901.10002v4

[117] Piotr Szymański, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. 2020. WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3290–3295. https://doi.org/10.18653/v1/2020.findings-emnlp.295

[118] Sy Taffel. 2021. Data and oil: Metaphor, materiality and metabolic rifts. *New Media & Society* 0, 0 (2021), 0. https://doi.org/10.1177/14614448211017887 arXiv:https://doi.org/10.1177/14614448211017887

[119] Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, 53–59. https://doi.org/10.18653/v1/W17-1606

[120] Rachael Tatman and Conner Kasten. 2017. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2017-Augus (2017), 934–938. https://doi.org/10.21437/Interspeech.2017-1746

[121] Andrew R Timming. 2016. The effect of foreign accent on employability: a study of the aural dimensions of aesthetic labour in customer-facing and non-customer-facing jobs. *Work, Employment and Society* 31, 3 (April 2016), 409–428. https://doi.org/10.1177/0950017016630260 Publisher: SAGE Publications.

[122] Francesca Tripodi. 2021. Ms. Categorized: Gender, Notability, and Inequality on Wikipedia. *New Media & Society* (2021), 14614448211023772. https://doi.org/10.1177/14614448211023772

[123] Emily van der Nagel. 2018. 'Networks that work too well': intervening in algorithmic connections. *Media International Australia* 168, 1 (2018), 81–92. https://doi.org/10.1177/1329878X18783002 arXiv:https://doi.org/10.1177/1329878X18783002

[124] Thibault Viglino, Petr Motlicek, and Milos Cernak. 2019. End-to-End Accented Speech Recognition. In *Proc. Interspeech 2019*. 2140–2144. https://doi.org/10.21437/Interspeech.2019-2122

[125] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *Ninth International AAAI Conference on Web and Social Media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10585

[126] Alicia Beckford Wassink. 2021. Uneven Success: Racial Bias in Automatic Speech Recognition. *Martin Luther King, Jr. Colloquium* (Jan 2021). https://www.youtube.com/watch?v=CFKTxUmLByo

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

[128] Steven Weinberger. 2015. The speech accent archive. online. http://accent.gmu.edu

[129] Uriel Weinreich, Marvin Herzog, William Labov, and Winfred Lehmann. 1968. Empirical foundations for a theory of language change. In *Directions for historical linguistics*, Yakov Malkiel (Ed.). University of Texas, 95–188.

[130] John C. Wells. 1982. *Accents of English*. Vol. 2. Cambridge University Press. https://doi.org/10.1017/CBO9780511611759

[131] Dong Yu and Li Deng. 2015. *Automatic Speech Recognition: A Deep Learning Approach*. Springer London, London. https://doi.org/10.1007/978-1-4471-5779-3_1