Working papers in Applied Linguistics and Linguistics at York 1, 4-15

Artificial intelligence language models and the false fantasy of participatory language policies

Mandy Lau[1]
York University, Toronto, Canada

Abstract: Artificial intelligence neural language models learn from a corpus of online language data, often drawn directly from user-generated content through crowdsourcing or the gift economy, bypassing traditional keepers of language policy and planning (such as governments and institutions). Here lies the dream that the languages of the digital world can bend towards individual needs and wants, and not the traditional way around. Through the participatory language work of users, linguistic diversity, accessibility, personalization, and inclusion can be increased. However, the promise of a more participatory, just, and emancipatory language policy as a result of neural language models is a false fantasy. I argue that neural language models represent a covert and oppressive form of language policy that benefits the privileged and harms the marginalized. Here, I examine the ideology underpinning neural language models and investigate the harms that result from these emerging subversive regulatory bodies.

Keywords: language policy; artificial intelligence; neural language models

1 Introduction

Artificial intelligence (AI) language models predict the probability of sequences of words and sentences, forming the basis of Natural Language Processing (NLP), a branch of computer science. These models are applied to many common NLP interpretation and generation tasks (i.e., word prediction for autofill, speech recognition in digital assistants, machine translation, or text summarization). Recently, language models increasingly depend on the use of neural networks in their machine learning. Also known as deep learning, neural language models enable computers to autonomously seek out patterns in a given language dataset. This contrasts with older classical methods, where humans provide computers the linguistic rules for statistical machine learning. Neural networks far outperform classical methods in predictive accuracy, processing vast amounts of information quickly and efficiently while seeking new patterns that humans could not possibly have predicted. (For more on 'what is NLP', see Crash course computer science, 2017; Brownlee, 2019.) It is possible that this form of black-box machine learning could mitigate against programmer bias and manipulation, to the point that programmers themselves could not explain how or why computers have learned certain patterns. The learning is wholly dependent on the statistical patterns within the training dataset.

[1] *Corresponding author: laumandy@yorku.ca*

 ISSN 2564-2855

Neural language models that learn from a corpus of user-generated online content bypass traditional keepers of language policy and planning (such as top-down governments and institutions). The languages offered in digital spaces can be driven by user demand and actual usage, increasing the capacity for language personalization. Here lies the dream that the language of the digital world can bend towards individual needs and wants. Through the participatory language work of users, linguistic diversity, accessibility, and inclusion can be increased. However, the promise of a more participatory, just, and emancipatory language policy resulting from neural language models is a false fantasy. In this paper, I argue that neural language models represent a covert and oppressive form of language policy that benefits the privileged and harms the marginalized. I begin by summarizing how languages are organized online using Kelly-Holmes' (2019) framework. I will then examine the ideology underpinning English neural language models and investigate the negative consequences of this emerging subversive regulatory body. I conclude by drawing connections to the neoliberal context in which neural language models reside.

2 Online language organization

The ways in which languages are organized online are categorized by Kelly-Holmes (2019) into four eras:

1. Monolingualism was when English dominated the world wide web at the beginning period of the internet.
2. Multilingualism, described as a "partial and parallel multilingualism" (Kelly-Holmes, 2019, p. 28), was when the "big" global languages were resourced and available for user selection. This was made possible by the stabilization of non-ASCII-supported alphabets such as Devangari and Chinese characters. The typical framing of one language/per user/by territory within a shared web experience resembles multiple monolingualisms.
3. Hyperlingualism, which emerged at the time of Web 2.0, is characterized by dynamic interactivity, collaboration, and crowdsourcing. Both technological and ideological changes contributed to an unlimited number of languages in expanding digital spaces.
4. Idiolingualism is an "intensified but isolated hyperlingualism" (Kelly-Holmes, 2019, p. 33), marked by increased personalization and linguistic customization.

The development of neural language models coincided with the simultaneous eras of hyperlingualism and idiolingualism. The technological and ideological setting of these eras enabled the work of language policy and planning to bypass the boundaries of geography, language standardization, official state policy, and language professional competence (Kelly-Holmes, 2019). In the hyperlingualism era, users often provided labour for free through models of crowdsourcing and the so called "gift economy" in which labour and data are given without any formal arrangements for reciprocation. As a result, many previously overlooked and undervalued languages are now online, particularly benefitting low-resource languages and oral language revitalization efforts. Facebook Translation is one example of how users volunteered to translate words and up-vote translations, with a final moderation by Facebook (Kelly-Holmes, 2019). This process contains the appearance of democratization as it did not include language standardization work by language professionals, nor did it require users to prove any legitimacy as language translators. It was completely community usage based.

Idiolingualism builds on the developments of hyperlingualism, adding algorithmic customization to create a personalized language filter bubble (Kelly-Holmes, 2019). Examples

include predictive text based on your past language behaviours, or tailored translation according to your past language use, time, and geography through apps and mobile devices such as Google Translate or the Translate One2One wearable (Kelly-Holmes, 2019). Common use of these apps and devices generates new data that can once again be fed back into neural language models for machine learning refinements without language interventions or regulations.

3 Machine learning datasets: The case of English online

Participatory language knowledge generated by actual users and leveraged to personalize online language experiences appears to be more linguistically inclusive and socially just. However, such language data sets are not neutral when used for machine learning. Hidden within neural language modeling are powerful sociopolitical ideologies. As Kelly-Holmes (2019) summarizes, "the web is a sociolinguistic machine—fueled by online language practices and choices and by widespread and common-sense ideologies and beliefs about language" (p. 25).

Large datasets used by the broader general public should be representative of diverse worldviews. However, the language practices, beliefs, and ideologies picked up through machine learning are unevenly distributed. For example, the dataset Common Crawl[2] derives its language data from the internet over the last eight years. While vast, the internet tends to overrepresent young users from the global north (Bender & Gebru et al., 2021; Pew, 2021). Datasets of American and British English tend to underrepresent the language practices of people of marginalized identities, such as speakers of African American Vernacular English (Martin, 2021), and overrepresent the views of white supremacy, misogyny, homophobia, ableism, ageism, etc. (Bender and Gebru et al., 2021). This is because the crawling method tends to derive language corpora from user-generated content sites with the most incoming and outgoing links, such as Reddit, Twitter, or Wikipedia, overlooking non-dominant views from less mainstream sites (Bender and Gebru et al., 2021). Reddit, Twitter, and Wikipedia are not as open and accessible as presented; these sites enable easy suppression of voices via false flagging for moderation, as well as systemic harassment, trolling, and violence upon marginalized communities, restricting and pushing their voices out (Bender and Gebru et al., 2021; Monteiro, 2019; Wachter-Boettcher, 2017). This restriction privileges and further amplifies the voices and worldviews of the dominant identities who do not experience online violence.

Even machine learning that does not derive its data from user-generated content reflects societal bias. The open-source Google algorithm Word2vec combs through Google News articles to learn relationships between words, creating word associations or word embeddings. The assumption is that Google News is neutral, without consideration of how the content may mirror historical, societal injustices in content and language, or how market forces may drive media reporting coverage. Word2vec will then pick up these biases from the dataset, and consequently return results that perpetuate and reinforce the same biases, such as sexist associations between words (Bolukbasi et al., 2016; Wachter-Boettcher, 2017).

Crowdsourcing language work could enable more human agency but is not without bias. For example, the labour market on Amazon's crowdsourcing platform Mechanical Turk is unevenly distributed across countries, inadvertently biasing the sample (Erlewine & Kotek, 2016). In a demographic survey of 1000 Turk workers, 46% reported to be from the United States, 34% from India, and the remaining 19% were from 64 other countries (Ipeirotis, 2010). These workers are also more likely to be female in the US, but male in India, tend to be born in the 1980s, have a

[2] http://commoncrawl.org

higher educational level but earn a lower income than the general population, and are single without kids (Ipeirotis, 2010). This results in a highly skewed sample producing large quantities of crowdsourced language data.

Generally, the version of a language that holds the highest prestige tends to follow the language users with the highest prestige (Milroy, 2001). Knowing which version of online language carries the most prestige would require knowing which users carry the most weight: the language used by young, straight, American, white males—English. This becomes the default language when it comes to language participation or content on the internet. As of March 2021, over 60% of online content is in English, followed by Russian at 8.3%, and Turkish, Spanish, and Persian at just over 3% (W3Techs, 2021). This is the case despite English-language users being estimated in 2020 as comprising 25% of all internet users, followed by Chinese (19%), Spanish (7%), and Arabic (5%) (Internet World Stats, 2021).[3] The default cultural perspective then also becomes that of the young, straight, American, white male.

Efforts to make language more inclusive in online environments focus on filtering out hateful and offensive language. However, filtering methods generally remove all words on a list classified as offensive without considering the meaning or context of the word in use. For example, the Colossal Clean Crawled Corpus discards pages that contain 400 "dirty, naughty, obscene or otherwise bad words," with most of the words related to sex and some related to white supremacy (Bender and Gebru et al., 2021). This effectively wipes out pornography but also inevitably filters out the discourse of communities who have reclaimed some of the words on their list. These communities include LGBTQ communities or the #metoo movement. (Bender and Gebru et al., 2021). Filtering algorithms that weed out fake news and conspiracy theories work in similar ways, having the effect of wiping out the discourse of human rights activists or political justice by equating shared politically charged words with shared values (Nakov & Da San Martino, 2020). Similar to the aforementioned Facebook translation example, the work of verbal hygiene (Cameron, 1995) in regulating the appropriacy of online language has also bypassed traditional gatekeepers and is now relegated to AI technologists at large media corporations, who are also overwhelmingly young, straight, American, white, Judeo-Christian, and male (Monteiro, 2016).

4 Neural language models: A de-facto language policy

Large language models, particularly open-sourced ones, are used reflexively in common applications found on our mobile devices and computers. Any online activity mediated by language is an interaction with neural language models and their schema. In other words, language models act as a mechanism mediating between ideology and practice, leading to de-facto language policies (Shohamy, 2008). Language models enforce their de-facto policies with stealth and effectiveness, governing our decisions and behaviours in subversive ways. Public or private institutions that use applications powered by these language models are in effect surrendering control, at least partially, of their language policy to AI technologists. Their own carefully constructed language policy need not apply to the AI tools they choose to use; instead, by choosing to use the tools, they are inadvertently choosing to subscribe to whatever policy is embedded in their tools' language models.

[3] In this estimate, only one dominant language is assigned per internet user, when, in reality, many people are multilingual and speak English as a Global English (Internet World Stats, 2021). This type of framing of language is characteristic of what Kelly-Holmes (2019) refers to as the "multilingualism era" of language organization.

The human impact is serious; the following sections explain how language models can regulate human language production to include an AI audience and manipulate humans through filtering and synthetic language production. The consequences range from changing human linguistic behaviour to systemic experiences of identity-based microaggressions, discrimination, and violence.

4.1 Regulating human language behaviour

Interactions between artificial intelligence and humans create new forms of data that inform future human decisions and behaviours. AI sifts this data through their language models, using language forms to deduce the probability of future events and behaviours. Without actual language understanding, AI uses language proxies to arrive at interpretations, which guide human decisions. For decisions that have important consequences, some humans have learned to adapt and produce language tailored to language models to achieve the best possible results. An example of this is found in corporate disclosures. CEOs have learned to adjust their speech tone and word-combination choices to improve algorithmic scores and minimize triggering red flags, which affect the share-trading decisions of human analysts and traders (Cao et al., 2020; Wigglesworth, 2020).

While this may seem innocuous, producing language for a proxy-based machine is an opportunity cost. Human energy is redirected to comply with algorithms instead of communicating in meaningful ways. Further, knowledge of how opaque algorithms work in specific sectors is privileged and requires additional resources to gain an algorithmic edge, such as professional coaching or access to simulation algorithms. This serves to further widen socioeconomic inequities and exacerbate historical, social injustices.

A case in point is the AI system Hirevue. Over a million job candidates have been screened by Hirevue, which looks for proxies such as word choice, voice tone, and gestures to generate an employability score (Harwell, 2019). These scores determine who gets to the interview stage. Since the "look for's" are opaque, candidates often seek out recruitment coaches and training programs to learn to optimize their scores (Harwell, 2019; O'Neil, 2016). Those who are unable to seek this type of support due to financial or time constraints are disadvantaged. The more severe consequence is that employment and human rights laws do not apply to algorithm design (O'Neil, 2016). Further, some candidates cannot manipulate the scoring algorithms simply because of who they are. Hirevue's application uses up to 500,000 data points in a 30-minute, 6 question interview recording to generate a score (Harwell, 2019). Like many other automated recruitment software, some of these data points are linked to personality and mental illness assessments (O'Neil, 2016). Other performance proxies are less accurately assessed for certain populations who are not the default, such as those who speak with a non-American or non-British Standard English accent (Harwell, 2019), or Black women with dark skin tones (Buolamwini & Gebru, 2018). These proxies enable illegal employment discrimination by language, disability, race, ethnicity, gender, age, and more.

Corporate disclosures and employee recruitment are among many examples of how language models in AI systems enforce their de-facto language policies upon individuals and institutions. Other examples of AI language models used for high-stakes decisions include algorithmic recidivism risk prediction systems used in Canadian and American criminal court (COMPAS; O'Neil, 2016; Robertson, Khoo & Song, 2020), policing software, and higher education admission systems (O'Neil, 2016). These are only the most explicit and visible ways in which language is directly used as a proxy for AI systems to justify decisions and regulations. More implicit ways that neural language models control our language activities and, more

generally, our lives, include filtering algorithms and synthetic language generation, described in the next section.

4.2 We see what we want to see

The appeal of AI personalization is that it could help us navigate through massive amounts of information and increase our human agency in choosing the content we want to see in the languages we want to see it in. However, this is only an "illusion of increased choice…we are being steered through the global, multilingual web in a monolingual bubble" (Kelly-Holmes, 2019, p. 34). Machine learning determines your choices based on location data and past language behaviours, reducing your exposure to other languages (Kelly-Holmes, 2019). A narrowing of languages also means a narrowing of worldviews, with filters that recommend the content you may be interested in based on other users whose language practices are similar to yours. As a result, we become more socially, linguistically, and ideologically isolated within our own echo chamber.

The exception, however, are the users who are farthest away from the default identity of the young, straight, American, white male. Instead, these "edge case" users tend to be misrepresented through the white male gaze (Noble, 2018). Between 2009-2015, Noble (2018) documented how Google Search, using its language model BERT, misrepresented and stereotyped social identities, such as presenting hyper-sexualized Black women and girls as the first results generated in a broad search. Popular belief may be that search results provide the most relevant or useful information, and that racist and sexist results may be a mere reflection of society. However, Noble (2018) debunks this myth, noting that racist and sexist search results are the outcome of Google prioritizing clicks that generate advertising profits, underrepresenting results from competitors, less profitable smaller advertisers, and personal blogs. In this way, personalization is not a service to users but a service for advertisers, helping them to find the best match in terms of consumers. Therefore, personalization efforts have not actually resulted in as much variation as the public may believe (Noble, 2018, p. 55). Search results that misrepresent, stereotype, and dehumanize people lead to multiple harms, such as microaggressions, stereotypes, discrimination, to physical violence.

4.3 Garbage in, garbage out: Machine generated language

Since synthetic language, or language generated by language models, reflects the biased language and worldviews of the default young, straight, American, white male, it can be inaccurate at best and violent at worst. Inaccuracies in machine translation could have terrifying consequences if the translation service is leveraged as part of an apparatus for state surveillance as with the case of the Israeli occupation of Palestine. In the aforementioned crowdsourced Facebook Translation application, inaccurate machine translation led the Israeli police to the wrongful arrest of a Palestinian construction worker in the West Bank (Hern, 2017). Facebook inaccurately translated the man's caption of a photo of himself next to a bulldozer, "يصبحهم" (yusbihuhum) meaning "good morning", as "hurt them" in English or "harm them" in Hebrew (Hern, 2017). The trust placed on the system was so ingrained that the Israeli police did not even verify the translation with Arabic-speaking officers before making the arrest.

Trust in synthetic language makes humans particularly vulnerable to ideological manipulation. Language models can be deployed to generate vast amounts of coherent synthetic text quickly, making it an effective and efficient tool to create oppressive misrepresentation, propaganda, conspiracy theories, and fake news. This was brought to public attention in the 2016 United States election when fake news flooded social media via social bots to sway public opinion in Donald Trump's favour (Bovet & Makse, 2019; Noble, 2018; Wachter-Boettcher, 2017). Other

examples include Google's computer vision and language models linking Black people, most prominently Michelle Obama, to apes or animals (Noble, 2018), or the LA Times' use of automated text generation, leading to tweets that tend to be more racist and misrepresentative than human-generated tweets, as in the case of their tweet misrepresenting police-shooting victim Keith Lamont Scott as a criminal (Ascher, 2017; Noble, 2018).

McGuffie and Newhouse (2020) demonstrated how one of the largest and most powerful language models, OpenAI's GPT-3, can be weaponized to create synthetic interactional and informational texts easily and efficiently for far-right extremist radicalization and recruitment efforts (Bender & Gebru et al., 2021). While less advanced language models would require hours of labour and sophisticated technological resources to create ideologically biased texts, an advanced model like GPT-3 has the capacity to produce realistic and consistent fake text when it is fed a few simple inputs, such as a few tweets, paragraphs, forum threads, or emails, without technical know-how (McGuffie & Newhouse, 2020). Accordingly, GPT-3 can be prompted to perform language tasks such as "producing polemics reminiscent of Christchurch shooter Brenton Tarrant, reproducing fake forum threads casually discussing genocide and promoting Nazism in the style of the defunct Iron March community, answering questions as if it was a heavily radicalized QAnon believer, and producing multilingual extremist texts, such as Russian-language anti-Semitic content, even when given English prompts" (McGuffie & Newhouse, 2020). Currently, GPT-3 is not open-sourced, but it is still vulnerable to attack and copying.

Drawing from this, it is easy to imagine how language models can be weaponized to generate and distribute synthetic hate-speech, racist memes, conspiracy theories, or pseudoscience on the internet for many more malicious causes. We are now seeing just how dangerous online propaganda is during the current global COVID-19 pandemic (Romer & Jamieson, 2020; van der Linden, Roozenbeek & Compton, 2020). What the World Health Organization (2021) terms the "infodemic" poses a threat to public trust and public health, from the largest toxic alcohol outbreak in Iran following fake news suggesting its preventative use against COVID-19 (Delirrad & Mohammadi, 2020), to spikes in hate crimes and xenophobic violence. In Vancouver, anti-Asian hate crimes increased 717% from 2019 to 2020 (Manojlovic, 2021). In a September 2020 report, British Columbia was found to have the most reported anti-Asian racist incidents per capita in North America, followed by California, New York, and Ontario (Project 1907, 2020). Women are impacted the most, accounting for 60-70% of all reported incidents (Project 1907, 2020). In the UK, Awan and Khan-Williams (2020) documented how COVID-19 triggered an increase in Islamophobic fake news and theories, such as blaming Muslims for spreading the virus by attending mosques and not following social distancing rules.

The extent to which language models, synthetic text generation, and the proliferation of fake news online are connected to acts of violence is difficult to prove. However, some examples show a correlation between persistent engagement with online conspiracy theories and hate crimes (Noble, 2018). In the 2015 mass shooting at Emanuel African Methodist Episcopal Church, White supremacist Dylann Roof shot at twelve Black church members during worship, killing nine (Kaadzi Ghansah, 2017). His manifesto indicated that his racist attitudes were stoked by engagement with white supremacist online material, including fake news, conspiracy theories, and dehumanizing stereotypes, following a Google search of "Black on White crime" (Kaadzi Ghansah, 2017; Noble, 2018).

Together, these examples demonstrate how language policies embedded within language models can and do manipulate human behaviour, subjecting already marginalized communities to

further injustices and systemic violence. Those who choose to use tools powered by neural language models and those who are most negatively impacted by these tools need to consider their ethical and legal rights and responsibilities carefully.

5 The reality of machine learning

If language is "neutral," then perhaps a participatory approach to language models and policy might work in ways that respect all humans. However, as previous examples demonstrate, language is ideologically bound. In thinking through verbal hygiene theory (Cameron, 1995), a naturalist might argue that even if language data is ideologically bound, we should "leave it alone". It suggests that verbal hygiene practices should not be enacted because it is a natural reflection of actual language use in a complex society. Part of the appeal of machine learning is its ability to relearn according to the most current language practices that are fed back into the system and improve when given larger datasets. However, this is not as easy as it seems. Bender et al. (2021) outlined the harmful environmental footprint of large datasets, as well as the argument that increasing the size of the dataset only increases the many inherent problems previously mentioned. They also note that retraining large language models is expensive and could not possibly reflect the most current social movements or movements that are poorly documented or insufficiently reported by the media (Bender et al., 2021). As a result, large language models tend to be more static and fixed.

Further, it is unclear who is accountable for the consequences of neural language models. Stakeholders who might be included could be the company that releases the language model, the language model engineers, the software company that uses the language model to create its application, the application user (institution or individual), or state legislators. Currently, the stakeholders with the most power in online language policy and the most interest in the opaqueness of these systems are the companies that create the language models and leverage them for applications. These technology companies thrive on deregulation, a freedom of speech logic, and profit-models built on engagement metrics. Thus, their language policy and planning are based on shareholder interest, even if that means protecting this highly profitable driver of engagement: the "free" speech of white supremacy and misogyny.

At the same time, technology companies must pacify public resistance and control public pressure for regulation by installing ethics teams, such as the Ethical AI team at Google. However, these teams tend to hold only symbolic power. Those within who dare to challenge company practices are seen as threats. For example, the co-founder of Google's Ethical AI team, Timnit Gebru, one of few Black women in leadership, was fired due to a research paper set for release (Schiffler, 2021). Titled "On the dangers of stochastic parrots: Can language models be too big?" (Bender & Gebru et al., 2021) and referenced in this paper, the authors consider the harm caused by various large language models, implicating Google in a third of the large language models under study: BERT, ALBERT, GShard, and Switch-C. One month following Gebru's exit, the other Google Ethical AI co-founder and co-author of the paper, Margaret Mitchell, was placed on administrative leave and subsequently fired after criticizing the company's behaviour and actions towards Gebru and their opposition towards the paper (Dickey, 2021). This reveals Google's contradictory position: in this case, we see how Google actively engaged in verbal hygiene practices by suppressing the voices of marginalized employees (who were fulfilling their job duties in critiquing the ethics of AI) to protect its profit-making model while claiming to be invested in "AI for social good…[and] to bring the benefits of AI to everyone" (Google, 2021).

6 Conclusion

Neural language models act as a de-facto language policy (Shohamy, 2008), enabling what Noble (2018) terms *technological redlining*: how algorithms and big data (such as neural language models) reinforce social injustices. How languages are organized online in the hyperlingualism and idiolingualism eras (Kelly-Holmes, 2019) leads to language models that overrepresent the views of the default user: English-speaking, young, white males (Bender & Gebru et al., 2021; Wachter-Boettcher, 2017) and increasing isolation in the languages and worldviews we access online. It can be true that crowdsourcing and the gift economy are able to increase linguistic vitality and support the development of under-resourced languages. However, bypassing traditional sources of language expertise and state structures can lead to mistakes and misuse. It can also have the unintended effect of maintaining the status quo in a state's official language rights (Kelly-Holmes, 2019). Personalization in the forms of devices, such as the Translate One2One wearable, only serves as a linguistic accommodation that keeps you in a monolingual bubble, rather than enabling exposure to a more multilingual society. Rather than bending to our individual needs and wants, personalization exposes us to commercial interests and regulates our language behaviours (Bender & Gebru et al., 2021; Noble, 2018). Democratization, diversity, and inclusion via participatory language work is only a false fantasy; in reality, participatory language work is oppressive and exploitative, organized to benefit classes of powerful elites financially and politically (Noble, 2018). Just the six big tech companies alone (Apple, Microsoft, Amazon, Alphabet/Google, Facebook, and Tesla) are worth over $9 trillion, making up a quarter of the S&P index funds and are bigger than the entire European stock market in 2020 (Klebnikov, 2020; La Monica, 2021). Their sheer power makes it more important than ever to include state governance in the work of online language policy and planning. A naturalistic, laissez-faire approach "is nothing but a policy for making powerful interests and strong forces even stronger and more powerful" (Kristiansen, 2003, p. 69). Therefore, stronger, more transparent language work and policies involving more diverse stakeholders in the form of democratic regulation on neural language models are necessary next steps.

References

Ascher, D. (2017). *The new yellow journalism: Examining the algorithmic turn in news organizations' social media information practice through the lens of cultural time orientation.* (Proquest ID: Ascher_ucla_0031D_16033) [Doctoral dissertation, University of California, Los Angeles]. eScholarship.

Awan, I., & Khan-Williams, R. (2020). *Research briefing report 2020: Coronavirus, fear and how Islamophobia spreads on social media 2020*. Anti-Muslim hatred working group. https://antimuslimhatredworkinggrouphome.files.wordpress.com/2020/04/research-briefing-report-7-1.pdf

Bender, E., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *FAcct '21: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 610-623. https://doi.org/10.1145/3442188.3445922

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *30th Conference on Neural Information Processing Systems*, 1–9. https://papers.nips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications, 10*(7), 1–14. https://doi.org/10.1038/s41467-018-07761-2

Brownlee, J. (2019). *What Is Natural Language Processing?* Machine learning mastery. https://machinelearningmastery.com/natural-language-processing/

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 1–15. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Cameron, D. (1995). *Verbal Hygiene*. Routledge.

Cao, S., Jiang, W., Yang, B., Zhang, A. L., & Robinson, J. M. (2020). *How to talk when a machine is listening: Corporate disclosure in the age of AI* (NBER Working Paper No. 27950). National Bureau of Economic Research. https://www.nber.org/papers/w27950

Crash course computer science. (2017, November 22). *Natural language processing: Crash course computer science #36* [Video]. YouTube. https://www.youtube.com/watch?v=fOvTtapxa9c

Delirrad, M., & Mohammadi, A. B. (2020). New methanol poisoning outbreaks in Iran following COVID-19 pandemic. *Alcohol and Alcoholism*, *55*(4), 347–348. https://doi.org/10.1093/alcalc/agaa036

Dickey, M. R. (2021, February 19). Google fires top AI ethics researcher Margaret Mitchell. *Tech Crunch.* https://techcrunch.com/2021/02/19/google-fires-top-ai-ethics-researcher-margaret-mitchell/

Erlewine, M. Y., & Kotek, H. (2016). A streamlined approach to online linguistic surveys. *Natural Language & Linguistic Theory*, *34*(2), 481–495. https://doi.org/10.1007/S11049-015-9305-9

Google. (n.d.). *AI for Social Good.* Google AI. https://ai.google/social-good/

Harwell, D. (2019, November 6). HireVue's AI face-scanning algorithm increasingly decides whether you deserve the job. *The Washington Post.* https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/

Hern, A. (2017, October 24). Facebook translates "good morning" into "attack them", leading to arrest. *The Guardian*. https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest

Internet World Stats. (2021). *Internet world users by language: Top 10 languages.* Internet World Stats: Usage and population statistics. https://www.internetworldstats.com/stats7.htm

Ipeirotis, P. (2010). *The new demographics of Mechanical Turk*. A Computer Scientist in a Business School. https://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html

Kaadzi Ghansah, R. (2017, August 21). A most American terrorist: The making of Dylann Roof. *GQ*. https://www.gq.com/story/dylann-roof-making-of-an-american-terrorist

Kelly-Holmes, H. (2019). Multilingualism and technology: A review of developments in digital communication from monolingualism to idiolingualism. *Annual Review of Applied Linguistics*, *39*, 24–39. https://doi.org/10.1017/S0267190519000102

Klebnikov, S. (2020, August 28). U.S. tech stocks are now worth more than $9 trillion, eclipsing the entire European stock market. *Forbes.*

https://www.forbes.com/sites/sergeiklebnikov/2020/08/28/us-tech-stocks-are-now-worth-more-than-9-trillion-eclipsing-the-entire-european-stock-market/

Kristiansen, T. (2003). Language attitudes and language politics in Denmark. *International Journal of the Sociology of Language*, *159*(2003), 57–71. https://doi.org/10.1515/ijsl.2003.009

La Monica, P. (2021, January 6). Proof Big Tech is way too big: It's a quarter of your portfolio. *CNN*. https://www.cnn.com/2021/01/06/investing/stocks-sp-500-tech/index.html

Manojlovic, D. (2021). *Report to the Vancouver Police Board: Year-end 2020 Year-to-date key performance indicators report.* Vancouver Police Department. https://vancouverpoliceboard.ca/police/policeboard/agenda/2021/0218/5-1-2102P01-Year-end-2020-KPI-Report.pdf

Martin, J. L. (2021). Spoken corpora data, automatic speech recognition, and bias against African American language: The case of habitual 'be.' *FAcct '21: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 284. https://doi.org/10.1145/3442188.3445893

McGuffie, K., & Newhouse, A. (2020). The radicalization risks of GPT-3 and advanced neural language models. *ArXiv*. http://arxiv.org/abs/2009.06807

Milroy, J. (2001). Language ideologies and the consequences of standardization. *Journal of Sociolinguistics*, *5*(4), 530–555.

Monteiro, M. (2019). *Ruined by design: How designers destroyed the world, and what we can do to fix it.* Mule Books.

Nakov, P., & Da San Martino, G. (2020, November 19). *Fact-checking, fake news, propaganda, and media bias: Truth seeking in the post-truth era* [Conference presentation]. EMNLP 2020 Conference. https://virtual.2020.emnlp.org/tutorial_T2.html

Noble, S. (2018). *Algorithms of oppression*. NYU Press.

O'Neil, C. (2016). *Weapons of math destruction*. Crown Books.

Pew Research Center. (2021). *Internet/broadband fact sheet.* Pew research center. https://www.pewresearch.org/internet/fact-sheet/internet-broadband/

Project 1907. (2020). *Racism incident reporting centre: A community-based reporting tool to track incidents of racism*. Project 1907. https://www.project1907.org/reportingcentre

Robertson, K., Khoo, C., & Song, Y. (2020). *To surveil and predict: A human rights analysis of algorithmic policing in Canada*. Citizen Lab and the International Human Rights Program. https://citizenlab.ca/wp-content/uploads/2020/09/To-Surveil-and-Predict.pdf

Romer, D., & Jamieson, K. H. (2020). Conspiracy theories as barriers to controlling the spread of COVID-19 in the U.S. *Social Science and Medicine*, *263*. https://doi.org/10.1016/j.socscimed.2020.113356

Schiffler, Z. (2021, March 5). Timnit Gebru was fired from Google - then the harassers arrived. *The Verge*. https://www.theverge.com/22309962/timnit-gebru-google-harassment-campaign-jeff-dean

Shohamy, E. (2008). Language policy and language assessment: The relationship. *Current Issues in Language Planning*, *9*(3), 363–373. https://doi.org/10.1080/14664200802139604

van der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating against fake news about COVID-19. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.566790

W3Techs. (2021). *Historical trends in the usage statistics of content languages for websites*. Web technology surveys. https://w3techs.com/technologies/history_overview/content_language

Wachter-Boettcher, S. (2017). *Technically wrong: Sexist apps, biased algorithms, and other threats of toxic tech*. W.W. Norton & Company.

Wigglesworth, R. (2020, December 5). Robo-surveillance shifts tone of CEO earnings calls. *Financial Times.* https://www.ft.com/content/ca086139-8a0f-4d36-a39d-409339227832

World Health Organization. (2021). *Infodemic*. World Health Organization. https://www.who.int/health-topics/infodemic#tab=tab_1