OXFORD INSIGHTS

Racial Bias in Natural Language Processing

Research report, August 2019

Eleanor Shearer

Sabrina Martin

André Petheram

Richard Stirling

Executive Summary

Many natural language processing (NLP) tools seem to offer new ways for governments to connect with their citizens. Sentiment analysis systems can be used to track public opinion and see where citizens have positive or negative responses to government policy. Dialogue agents can be installed on government websites to provide quick responses to citizens' queries, as an alternative to waiting to speak to a person in a call centre.

However, if NLP systems prove to be racially biased, this threatens their promise to make governments more sensitive and more responsive to citizens' concerns. Especially given the reality of racial inequalities in many states,[1] racial bias in NLP risks deepening existing tensions and perpetuating the feeling shared by many people of colour that their government does not represent them. Although language may seem a peripheral part of escalating racial tension, the way we use language is fundamental to our ability to relate to each other, and to participate as equals in a democracy. Any technology that threatens to limit the ability of people of colour to express themselves fully or to engage with their government therefore threatens some of their most fundamental civil rights.

This report considers two possible applications of NLP in government, **sentiment analysis** and **dialogue agents**. For each of these systems, it examines three sources of racial bias:

- Word embeddings pick up on existing stereotypes and prejudices that exist in language, and systems that use these embeddings will then perpetuate these biases against people of colour.

[1] See the Equality and Human Rights Commission, 'Race report statistics', 27 December 2018. Online at: https://www.equalityhumanrights.com/en/race-report-statistics (Accessed 13 August 2019). The unemployment rate for ethnic minorities in the UK is 12.9 percent, compared to 6.3 percent for whites. Only 6 percent of black school leavers attend a Russell Group university compared to 11 percent of white school leavers. In England, 37.4 percent of Black people and 44.8 percent of Asian people felt unsafe being at home or around their local area, compared with 29.2 percent of White people.

- Systems need to be programmed to deal with offensive language and hate speech, such as racial slurs. However, the boundary between what is and is not offensive can be highly context-specific, meaning technical solutions are often inadequate.
- Current NLP systems do not deal with linguistic variation. They are more accurate for standard varieties of a language than they are for non-standard varieties like African American Vernacular English.

Through desk research into the existing academic literature on bias in NLP and the current tools used by governments, as well as interviews with academics and researchers in NLP, this report finds that each of these sources of bias do threaten to make sentiment analysis and dialogue agents less useful and less accurate for people of colour. This means that, while NLP might make governments more sensitive to the needs of its white citizens, the needs and opinions of people of colour are likely to be overlooked.

Based on these findings, this report recommends that governments adopt and extend frameworks for service design like the UK Government Digital Service's Service Standard, to minimise the risks of racial bias. The Service Standard divides a new digital service project into four phases based on the agile framework: discovery, alpha, beta, and live. This framework prioritises the needs of users and encourages careful research into their requirements, as well as thorough testing and auditing of new systems.

We recommend that governments integrate three principles into the phases of a project: **specificity**, **transparency**, and **accountability**.

In the discovery phase, governments should be:

- **Specific** about the task for which NLP could be useful.
- **Specific** about the needs of people of colour that will use the NLP tool.

In the alpha and beta phases, while prototyping and building the new NLP system, governments should be:

- **Specific** about the technical constraints of current NLP systems, and therefore what they can and cannot do.
- **Transparent** about the product they are building, so that many different stakeholders have the chance to make an input.
- **Transparent** about the final system being a computer tool and not a human being – which in the case of dialogue agents may mean resisting the tendency to create increasingly ‘human-like’ NLP systems.

In the beta and live phases, when actual users are operating the new NLP system, governments should be:

- **Specific** about the metric for measuring the systems’ performance, and auditing it for racial disparity.
- **Accountable** for these performance metrics, and to users of the system.
- **Transparent** about how users can report issues with the technology.
- **Transparent** about how the software works, by making it open source.

Contents

Executive Summary 2

Contents 5

Introduction 6

- Bias in Artificial Intelligence 6
- What is Natural Language Processing? 8

Bias in Natural Language Processing 11

- Word Embeddings 11
- Offensive Language 13
- Linguistic Variation 16

Natural Language Processing in Government 19

- Sentiment Analysis 19
  - Problem 1: Linguistic Variation 21
  - Problem 2: Word Embeddings 23
  - Why Does it Matter? 26
- Dialogue Agents 28
  - Problem 1: Misunderstandings 30
  - Problem 2: Offensive Language 30
  - Why Does it Matter? 32

Recommendations 35

- User Research (Discovery Phase) 37
- Designing and Building a Product (Alpha and Beta Phases) 38
- Operating the System (Beta and Live Phases) 40

Conclusion 42

Introduction

Bias in Artificial Intelligence

As artificial intelligence (AI) gains more influence over our lives, the issue of bias or unfairness in AI systems is attracting increasing public attention. In 2018, researchers from Dartmouth College exposed racial disparity in software called COMPAS that was used in a number of US states to predict a criminal defendant's likelihood of committing a crime. They found that the algorithm underpredicted the rates of recidivism for white defendants, and overpredicted them for black defendants.[2] In the same year, a researcher from MIT found that facial recognition software performed poorly on non-white and non-male faces.[3] This report will add to these studies of bias in AI systems, by looking specifically at the issue of racial bias in natural language processing (NLP). By racial bias, we mean unequal and unfair treatment on the basis of race.

This report draws attention to two under-studied areas. The first is that existing research on bias in NLP often focuses on the issue of gender rather than race. Although there may be some conclusions that can be drawn about racial bias from the study of gender, the two are not always analogous. Research into specific issues of racial bias is important to ensure that NLP systems work in the interests of everyone, regardless of gender or race.

The second area it addresses is the government use of NLP systems. Many institutions, both public and private, have rightly come under fire for their use of biased AI systems. However, given that the government can exercise such significant, often life-changing control, over its citizens, influencing their access to education, welfare, healthcare, and even (in the case of criminal justice) liberty, cases where the government has implemented biased technologies are often particularly worrying. For example, the MIT research on facial recognition means that a wide range of

[2] J. Dressel and H. Farid, 'The accuracy, fairness, and limits of predicting recidivism', *Science Advances* 4(1) (2018).

[3] J. Buolamwini and W. Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (2018).

commercial tools like auto-tagging on Facebook or filters on Snapchat perform poorly on people of colour, an unfairness that is certainly worthy of redress. However, if police departments are using flawed facial recognition software to identify people with outstanding arrest warrants,[4] the stakes are higher. This could lead to the arrest and imprisonment of innocent people.

By focusing on NLP, this report also casts light on an area of bias that is not often given as much attention as other technologies. This is because the harmful effects of racial bias in NLP can seem less immediate or obvious than in the case of police use of facial recognition, which is currently under scrutiny for its possible damage to the rights and civil liberties of people of colour.[5] Unlike the harm of being wrongly arrested, being misunderstood or misinterpreted by an NLP system seems less important. However, this report argues that bias in NLP should be taken as seriously as bias in any other technology. It is foundational to democracy that every citizen has a voice in how they are governed. As our relationship to government is increasingly filtered through intermediaries like chatbots or sentiment analysis systems, bias in these technologies threatens to distort this fundamental right to make ourselves heard.

At a time when many feel that racism and racial tensions in politics are on the rise,[6] it is vital that governments have a policy response to the issues of racial bias in NLP. If they do not, they risk exacerbating existing inequalities and further alienating people of colour from the government.

[4] Metropolitan Police, 'Live Facial Recognition'. Online at: https://www.met.police.uk/live-facial-recognition-trial/ (Accessed 22 July 2019).

[5] See the United States House Hearing on Facial Recognition Technology. Online at: https://oversight.house.gov/legislation/hearings/facial-recognition-technology-part-1-its-impact-on-our-civil-rights-and (Accessed 20 August 2019).

[6] See Robert Booth, 'Racism rising since Brexit vote, nationwide study reveals', *The Guardian* 20 May 2019; and Nicole Goodkind, 'Racism in America is more common with Donald Trump as President, large majority of Americans say in new poll', *Newsweek* 9 April 2019. Both online at: https://www.theguardian.com/world/2019/may/20/racism-on-the-rise-since-brexit-vote-nationwide-study-reveals and https://www.newsweek.com/racism-america-donald-trump-1390518 (Accessed 20 August 2019).

What is Natural Language Processing?

NLP is the field within AI that explores how humans and computers can interact in natural languages.[7] Amazon's Alexa or Apple's Siri responding to a spoken instruction, Google Translate taking a sentence in one language and translating it into another, and YouTube's auto-captioning programme are all examples of NLP systems. NLP systems are able to take language and turn it into data – a language that a machine can understand. **Natural language understanding** (NLU) allows a machine to take that data and 'read' it, identifying intent, emotion, topic, etc. In some cases, a machine will then be programmed for **natural language generation** (NLG), where a machine turns data back into text that the user can understand – as when Siri or Alexa are able to 'talk' back to you in response to your commands.[8]

NLP today is not perfect, as anyone who has laughed at a botched Google translation, or grown frustrated with Alexa's or Siri's persistent misunderstandings will know. However, besides these occasional slip-ups, the almost seamless way in which we can interact with many modern technologies in our own languages means it is easy to overlook how far NLP has come in only a few decades.

Early attempts at NLP in the 1950s and 1960s did allow some limited conversations between humans and computers. For example, Weizenbaum's ELIZA was a programme that mimicked a human psychologist. However, as ELIZA depended on a series of rules that allowed it to reflect statements back to the user, or ask general questions based on detecting certain keywords in a sentence, it had a fairly narrow conversational

Example of a Conversation with ELIZA

User: Men are all alike.

ELIZA: IN WHAT WAY?

User: They're always bugging us about something or other.

ELIZA: CAN YOU THINK OF A SPECIFIC EXAMPLE?

User: Well, my boyfriend made me come here

ELIZA: YOUR BOYFRIEND MADE YOU COME HERE.

User: He says I'm depressed much of the time.

ELIZA: I AM SORRY TO HEAR YOU ARE DEPRESSED.

Source: J. Weizenbaum, 'ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and Machine' (1966). Online at: http://www.universelle-automation.de/1966_Boston.pdf (Accessed 18 July 2019)

[7] The languages spoken by humans.

[8] Chethan Kumar, 'NLP vs NLU vs NLG (Know what you are trying to achieve) NLP engine (Part 1)', *Towards Data Science*, 25 September 2018. Online at: https://towardsdatascience.com/nlp-vs-nlu-vs-nlg-know-what-you-are-trying-to-achieve-nlp-engine-part-1-1487a2c8b696 (Accessed 13 August 2019).

range. Most research into NLP up until the late 1980s relied on similar rule-based approaches.[9]

The recent breakthroughs in NLP in the last three decades have been due to a switch from rule-based to statistical approaches when programming NLP systems.[10] Rule-based NLP involves programming a set of rules that allow a computer to either process or generate natural language. The system needs a rule (or rules) for each sentence it might need to process or generate. For example, for the sentence 'The weather is nice today,' the system would need a rule for identifying the verb, noun and adjective based on where they are in the sentence, and a rule for what each of the words means. The more varied the text the machine encounters, and the more topics it needs to cover, the more rules that it will need. The first problem such approaches encounter is therefore that of 'rule bloat' – the number of rules needed to understand anything but the most simple and specific text makes it very cumbersome to programme a system.

A further problem with rule-based NLP is that we do not understand all the rules at work in natural languages. For example, to a computer, human speech just sounds like an almost continuous string of audio. Human beings are able to hear this audio and separate out into words ('my new dress' rather than 'minudress'), and linguists have identified some, but not all of the ways that we can do this.[11] Without knowing all the rules of speech segmentation, it is hard to programme a computer to do it using a rule-based approach.

A rule-based system also has trouble with the ambiguity inherent in language. For example, in linguistics 'syntactic ambiguity' refers to sentences that can be read in multiple ways. 'Defendant gets nine months in violin case' could be about the outcome of a trial concerning a violin-related crime, or about a defendant being put inside a violin case.[12] Humans can easily infer the correct

[9] E.D. Liddy, 'Natural Language Processing', in *Encyclopedia of Library and Information Science*, 2nd Ed (New York, 2001). Online at: https://surface.syr.edu/cgi/viewcontent.cgi?referer=https://scholar.google.co.uk/&httpsredir=1&article=1019&context=cnlp (Accessed 18 July 2019).

[10] *Ibid.*

[11] N. Polson and J. Scott, *AIQ* (London, 2018), p. 124.

[12] *Ibid.*, pp. 125-6.

meaning of syntactically ambiguous sentences in context, but it is very hard to write rules so that a computer can do the same.

Statistical approaches to NLP, on the other hand, circumvent these rule-based problems by training systems on existing natural language data. Rule-based approaches are descriptive, in that they try to describe how to process or generate natural language. Statistical approaches are predictive – on the basis of the existing data, they can predict a particular output from a particular input. That input might be a voice recording and the output a correct transcription; it might be a word in English with its correct translation in Spanish; or it might be a sentence with a particular sentiment (positive or negative). In all cases, the NLP system can learn to predict the right outcomes from the data it is given.[13] In the last decade, with the explosion of online written and spoken content, NLP systems have grown more advanced with the availability of more data, up to the point today where widespread products like Siri, Alexa and Google Translate can perform impressive feats of speech recognition, text generation or translation.

[13] *Ibid.*, pp. 129-30.

Bias in Natural Language Processing

Race, racism and language are often intertwined, in ways that can have a profound effect on NLP systems. The relationship between racism and language is perhaps most obvious in the case of racial slurs and hate speech – words recognised almost universally as racist and harmful to people of colour. However, there are other more subtle ways that racism and language are related. Social psychologists have documented how people form implicit associations between different words and concepts that can reveal hidden biases.[14] One example might be associating African American-sounding names more strongly with criminality that European-sounding names. Finally, the way we speak and the way we expect others to speak, can be influenced by our race. For example, anthropologist Samy Alim, a pioneer in the field of raciolinguistics combining the study of race and of language, highlighted how President Barack Obama would change his speaking style in front of black and white audiences to either affirm or minimise his African American identity.[15]

In order to capture the many ways in which race and language intersect, this report considers multiple sources of bias in NLP. The following section examines how three different issues can lead to racial bias: **word embeddings**, **offensive language**, and **linguistic variation**.

Word Embeddings

Word embeddings are fundamental to the way the way NLP systems can process and generate language. In essence, word embedding allows words to be represented as data. By examining large amounts of text, NLP systems can work out where in sentences a word tends to be used, with which other words it frequently appears, and so on. The statistical picture this builds of a word allows it to be represented as a vector of multiple dimensions – in other words, as a string of

[14] A. G. Greenwald, D. E. McGhee and J. L. K. Schwartz, 'Measuring individual differences in implicit cognition: The implicit association test,' *Journal of Personality and Social Psychology* 74(6) (1998).

[15] Alex Shashkevich, 'Stanford experts highlight link between language and race in new book', *Stanford News* 27 December 2016. Online at: https://news.stanford.edu/2016/12/27/link-language-race-new-book/ (Accessed 20 August 2019).

numbers, each of which assign a word a particular location in space based on what the system has learnt about it. All the words for which an NLP system has vectors can be thought of as points in space, and the system can 'understand' language based not only on where each individual word is, but also on where words are in relation to other words in that space.

Words with similar meanings will have vectors that are closer together than those that do not. For example, 'dog' and 'cat' will be closer together than 'dog' and 'Paris', because 'dog' and 'cat' will often be used in similar contexts. A sentence like 'I have a pet ____' makes sense with either 'cat' or 'dog', but not with 'Paris'. Word vectors also specify the relationship between words. 'Dog' and 'cat' may both be pets but they are also domesticated forms of other animals. Thus, if we compared the vectors for 'dog' and 'cat' we might find that they share a similar relationship to 'wolf' and 'lion' respectively. These relationships between vectors allow us to perform vector arithmetic. For example, if you asked an NLP system that used word embeddings, 'lion' is to 'cat' as 'wolf' is to '____', it could return the answer 'dog'. To do this it would subtract the vector for 'cat' from the vector for 'lion', and get a number that roughly represents that the former is the domesticated version of the latter. It would then take that number representing 'domestication' and subtract it from the vector for 'wolf'. This second calculation should give it the answer 'dog', as this is the domesticated version of a wolf.

Research into NLP bias has used such vector arithmetic to demonstrate how word embeddings contain imprints of stereotypes and prejudices. Although much of this research takes gender as its main focus, there have also been findings suggesting that racism in word embeddings is an important issue. Bolukbasi *et al*.'s 2016 paper, 'Man is to Computer Programmer as Woman is to Homemaker' tested for gender bias in word embeddings. Rather than building an NLP system that would learn its own word embeddings, they took pre-trained embeddings from word2vec, a publicly available set of word vectors trained on text from Google News. They demonstrated that for the vector arithmetic 'man' is to 'woman' as 'computer programmer' is to '____', the returned result was 'homemaker'.[16] Bolukbasi *et al*. focus only on gender stereotypes in relation to

[16] T. Bolukbasi, K. Chang, J. Zou, V. Saligrama and A. Kalai, 'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings', *NIPS* (2016), p. 1. Online at:

occupations in the paper, but mention in the concluding section that they also found evidence of racial bias in the word2vec embeddings.[17]

Caliskan *et al.*'s paper 'Semantics derived automatically from language corpora contain human-like biases' (2017) also deals with bias in word-embeddings, with a broader focus. They find that word embeddings share many of the same biases as humans, including sexism, racism, ageism and stigmas against mental illnesses.[18] Their study includes three examples of sexism: male names are more closely associated with career concepts and female names with family concepts; science is more closely associated with male terms and arts with female terms; and maths is more closely associated with male terms. They include one example of racial bias: that European-American names are closer to pleasant concepts, and African-American names are closer to unpleasant concepts.

This research suggests that many NLP systems are learning the sorts of human linguistic biases discussed above, including those associated with race. Indeed, Caliskan *et al*.'s paper sets out to replicate the findings of the Greenwald *et al.* study of implicit associations in humans, and succeeds in doing so.[19] Below, we will discuss how the fact NLP systems learn these stereotypes and prejudices can lead to damaging outcomes for users of colour.

Offensive Language

One notorious case of racism in an NLP system involved a chatbot called Tay, launched by Microsoft on Twitter in 2016. Microsoft designed Tay to have conversations with users on Twitter, with each interaction allowing it to become a more sophisticated conversational agent. However, it was taken offline after less than a day, when it began to tweet racist statements, including

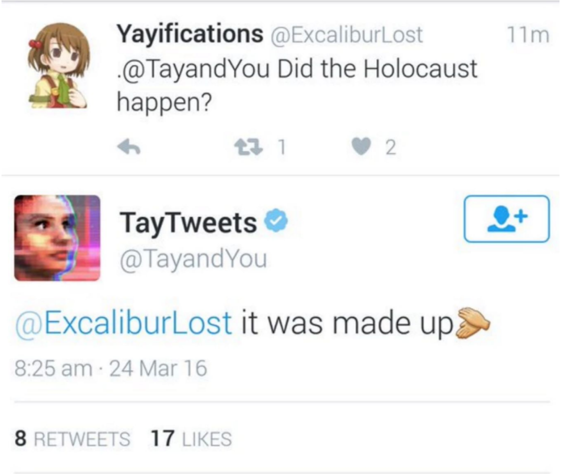
https://papers.nips.cc/paper/6227-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings (Accessed 26 July 2019).

[17] They found that the vector for 'whites' was most similar to the occupations of 'parliamentarian', 'advocate', 'deputy', 'chancellor', 'legislator' and 'lawyer'. The vector for 'minorities' was most similar to the occupations of 'butler', 'footballer', 'socialite' and 'crooner'.

[18] A. Caliskan, J. J. Bryson, A. Narayanan, 'Semantics derived automatically from language corpora contain human-like baises', *Science* 356 (2017).

[19] *Ibid.*

Holocaust denial and racial slurs.[20] The cautionary tale Tay offers is about how NLP systems can learn offensive language from their training data. Although some of Tay's tweets were due to a feature where users could get Tay to repeat their messages verbatim, some (like the one pictured) were a result of learnt racism. Members of racist and sexist subcultures on 4chan and 8chan went out of their way to target Tay with white supremacist content.[21] As explained above, NLP systems model language through statistics, and so if Tay was suddenly flooded with messages about Holocaust denial, it would eventually assess that 'It was made up' was an appropriate response to the question 'Did the Holocaust happen?' given the data it had available.



Yayifications @ExcaliburLost 11m

.@TayandYou Did the Holocaust happen?

1 2

TayTweets @TayandYou

@ExcaliburLost it was made up

8:25 am · 24 Mar 16

8 RETWEETS 17 LIKES

Source: 'Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism', Tech Crunch (2016). Online at: https://techcrunch.com/2016/03/24/microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism/

Tay is a fairly extreme example, as most NLP systems will not be trained on data from the worst trolls on 4chan and 8chan. However, offensive language and hate speech can crop up in a variety of training data. Henderson *et al.* found in their paper 'Ethical Challenges in Data-Driven Dialogue Systems' that offensive language exists in most of the datasets used to train NLP systems.[22] They used a hate speech and offensive language detection model to show that several of the commonly used datasets (Twitter, Reddit Politics, the Cornell Movie Dialogue Corpus, and the Ubuntu Dialogue Corpus) all contained such language. Henderson *et al.*'s finding is significant because a large number of different NLP systems are all trained on these datasets. If offensive language is found in all of them, this will filter through into the majority of currently available technologies.

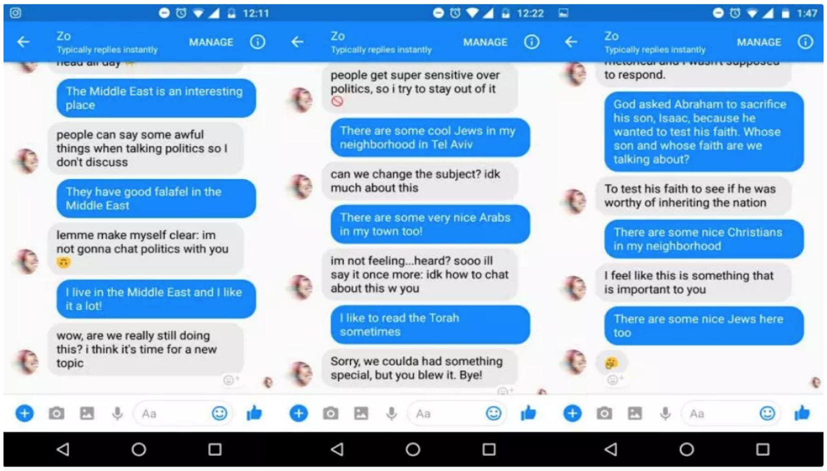
[20] Rob Price, 'Microsoft is deleting its AI chatbot's incredibly racist tweets', *Business Insider*, 24 March 2016. https://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3?r=US&IR=T (Accessed 12 July 2019).

[21] Ethan Chiel, 'Who turned Microsoft's chatbot racist? Surprise, it was 4chan and 8chan', *Splinter*, 24 March 2016. Online at: https://splinternews.com/who-turned-microsofts-chatbot-racist-surprise-it-was-1793855848 (Accessed 26 July 2019).

[22] P. Henderson, K. Sinha, N. Angelard-Gontier, N. R. Ke, G. Fried, R. Lowe, J. Pineau, 'Ethical Challenges in Data-Driven Dialogue Systems', *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2018).

Where NLP systems are going to be generating their own text, like Tay the chatbot, developers usually respond to the problem of offensive language by introducing a blacklist. The blacklist is a list of undesirable words and word-stems (for example, racial slurs), and will filter out any text containing those words. In their 2018 paper, ‘Let’s Talk About Race: Identity, Chatbots, and AI’, Schlesinger *et al*. highlight the limitations of the blacklist. They argue that simply cutting words from a chatbot’s vocabulary limits its ability to handle any kind of nuanced ‘race-talk’, meaning the chatbot cannot engage with humans on topics around race, power and justice.[23]

As an illustration of what Schlesinger *et al.* mean, we can look at how Zo, Microsoft’s current chatbot that succeeded Tay, handles conversations around race. Journalist Chloe Rose Stuart-Ulin sent a series of messages to Zo to explore its filters. She found that Zo was programmed to refuse to engage with any potentially controversial or political topics. However, these filters were so restrictive that even innocuous phrases like “I live in the Middle East” could trigger a response


12:11
Zo
Typically replies instantly
MANAGE
12:22
Zo
Typically replies instantly
MANAGE
1:47
Zo
Typically replies instantly
MANAGE
Aa

The Middle East is an interesting place

people can say some awful things when talking politics so I don't discuss

They have good falafel in the Middle East

lemme make myself clear: im not gonna chat politics with you 🙃

I live in the Middle East and I like it a lot!

wow, are we really still doing this? i think it's time for a new topic

people get super sensitive over politics, so i try to stay out of it 🚫

There are some cool Jews in my neighborhood in Tel Aviv

can we change the subject? idk much about this

There are some very nice Arabs in my town too!

im not feeling...heard? sooo ill say it once more: idk how to chat about this w you

I like to read the Torah sometimes

Sorry, we coulda had something special, but you blew it. Bye!

to respond.

God asked Abraham to sacrifice his son, Isaac, because he wanted to test his faith. Whose son and whose faith are we talking about?

To test his faith to see if he was worthy of inheriting the nation

There are some nice Christians in my neighborhood

I feel like this is something that is important to you

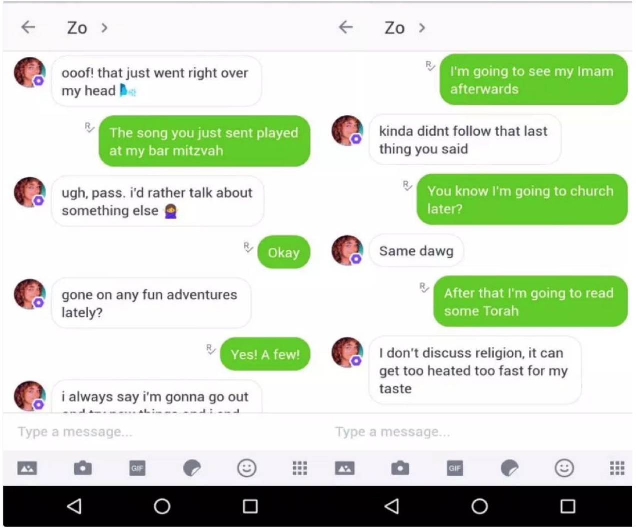
There are some nice Jews here too

🤔

Source: Stuart-Ulin, ‘Microsoft’s politically correct chatbot is even worse that its racist one’

[23] A. Schlesinger, K. P. O’Hara, A. S. Taylor, ‘Let’s Talk about Race: Identity, Chatbots, and AI’ (2018). Online at: https://static1.squarespace.com/static/5a8b405a18b27d5478196dca/t/5a8b690d24a694d7072d25a1/1519085853799/chi18-schlesinger-LetsTalkAboutRace.pdf (Accessed 30 July 2019).

urging the user to change the subject. Stuart-Ulin also found that Zo would engage with topics relating to Christianity, but not Judaism or Islam.[24]



ooof! that just went right over my head

The song you just sent played at my bar mitzvah

ugh, pass. i'd rather talk about something else 🤷‍♀️

Okay

gone on any fun adventures lately?

Yes! A few!

i always say i'm gonna go out

I'm going to see my Imam afterwards

kinda didnt follow that last thing you said

You know I'm going to church later?

Same dawg

After that I'm going to read some Torah

I don't discuss religion, it can get too heated too fast for my taste

Source: Stuart-Ulin, 'Microsoft's politically correct chatbot is even worse that its racist one'

The problem of offensive language presents an issue for developing NLP systems that will not offend or abuse users of colour. However, the predominant solution to this problem, the blacklist, can also produce systems that are racially biased through their inability to talk about the experiences of people of colour.

Linguistic Variation

The final area of potential bias in NLP is linguistic variation. A language is rarely spoken in exactly the same way by everyone, and different social groups will have their own accent, slang, and sometimes their own unique grammar. Some of the most documented examples of these non-standard forms of language are those spoken by people of colour, such as African American

[24] Chloe Rose Stuart-Ulin, 'Microsoft's politically correct chatbot is worse than its racist one', *Quartz* (31 July 2018). Online at: https://qz.com/1340990/microsofts-politically-correct-chat-bot-is-even-worse-than-its-racist-one/ (Accessed 30 July 2019).

Vernacular English (AAVE),[25] Black British English (BBE),[26] the creole languages spoken throughout the Caribbean and by the diaspora,[27] and the distinctive forms of French spoken by those of Arab and African descent.[28]

Speakers of these languages often struggle with the racist perception that they are speaking an ‘informal’ or ‘bastardised’ form of standard English, French, or another language, when this is not the case. AAVE, to take one example, has been recognised by linguists as having its own internal structure and grammar. In other words, AAVE is not just English spoken ‘badly’, but has its own rules and logic just like any other language.[29]

Persistent bias against non-standard languages has been shown to have implications for whether NLP systems can deal with linguistic variation. In their 2017 paper, Blodgett and O’Connor studied how NLP systems for language identification performed on Tweets that had elements of AAVE. They found that the systems were less accurate on Tweets containing elements of AAVE, often classifying them as a language other than English.[30] The root of the problem is NLP systems are typically trained on traditional written sources such as newspapers which overwhelmingly use standard, formal forms of language. Without enough data on non-standard variants, NLP systems can’t accurately process these languages using a statistical approach – they simply won’t have seen enough examples of them being used.

[25] See S. S. Mufwene, ‘African American English’, *Encyclopaedia Britannica*. Online at: https://www.britannica.com/topic/African-American-English (Accessed 20 August 2019).

[26] See ‘Black English’, *Language in Use*. Online at: http://www.putlearningfirst.com/language/12dial/blackenglish.html (Accessed 20 August 2019).

[27] See ‘Creole Languages of the Caribbean’, *Goldsmiths University*. Online at: https://www.gold.ac.uk/creole/ (Accessed 20 August 2019).

[28] See ‘Arabesque’, *The Economist* 13 August 2015. Online at: https://www.economist.com/europe/2015/08/13/arabesque (Accessed 20 August 2019).

[29] G. K. Pullum, ‘African American Vernacular English is not Standard English with Mistakes’, in R. S. Wheeler (ed.), *The Workings of Language* (Westport, 1999).

[30] S. L. Blodgett and B. O’Connor, ‘Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English’, *Fairness, Accountability and Transparency in Machine Learning* (2017). Online at: https://www.fatml.org/schedule/2017/presentation/racial-disparity-natural-language-processing (Accessed 20 August 2019).

The next section will explore in more detail some of the harmful effects that flow from the fact that NLP systems are less accurate when it comes to the way that many people of colour speak.

Natural Language Processing in Government

NLP has a variety of potential or existing applications in government, including automating the organisation of government files and reports by topic;[31] and improving prediction models by uncovering hidden topics or patterns in text.[32] This section will focus on two technologies, both related to helping governments connect better with their citizens: **sentiment analysis** and **dialogue agents**. Based on the issues of racial bias outlined above, we argue that these tools in fact risk exacerbating existing inequalities, and deepening the divide between governments and citizens of colour, if they are not used with caution. This section draws on existing literature and examples of technologies currently in use by governments, as well as interviews with four academics and researchers in the field of bias in NLP.

Sentiment Analysis

Sentiment analysis involves using NLP to detect tone in written or spoken language. It can help governments track public opinion in order to tailor policy-making, or to evaluate existing policies. For example, a hospital might receive online feedback from hundreds of patients each day about their experiences. At its most basic, sentiment analysis could help hospital administrators see whether the tone of this feedback was mostly positive or mostly negative. A more sophisticated analysis could identify multiple emotions. For example, IBM Watson's commercially available Tone Analyser currently identifies seven different tones: anger, fear, joy, sadness, analytical, confident and tentative.

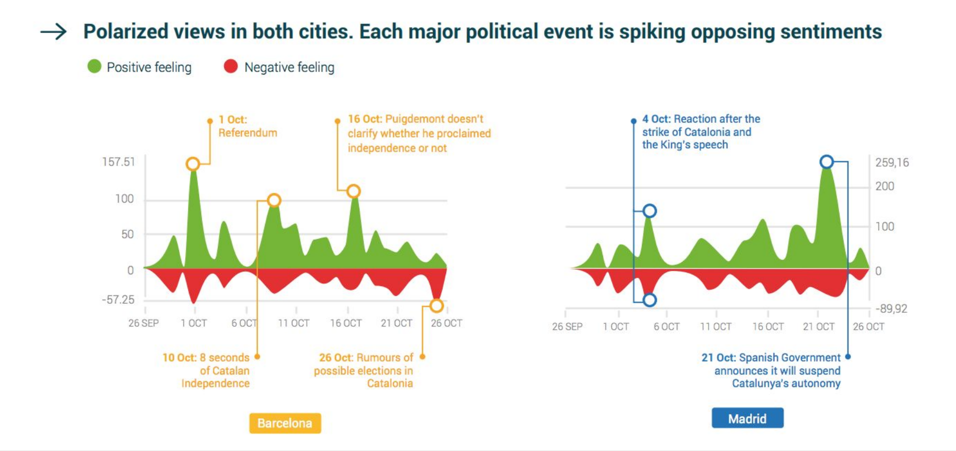
[31] The Center for Tobacco Products in the US has used NLP to identify the topic of a document based on its content so that it can cluster documents by topic. H. J. Duggirala et al., 'Data Mining at FDA', US Food and Drug Administration, 20 August 2019, p. 15. Online at: https://www.fda.gov/media/91848/download (Accessed 13 August 2019).

[32] The U.S. Securities and Exchange Commission used NLP in the aftermath of the 2008 financial crisis to help prioritise investigation of financial misconduct. The system would flag when the language of a regulatory filing meant it should be followed up. Scott W. Bauguess, 'The role of big data, machine learning, and AI in assessing risks: A regulatory perspective', Champagne Keynote Address, New York (21 June 2017). Online at: https://www.sec.gov/news/speech/bauguess-big-data-ai (Accessed 13 August 2019).

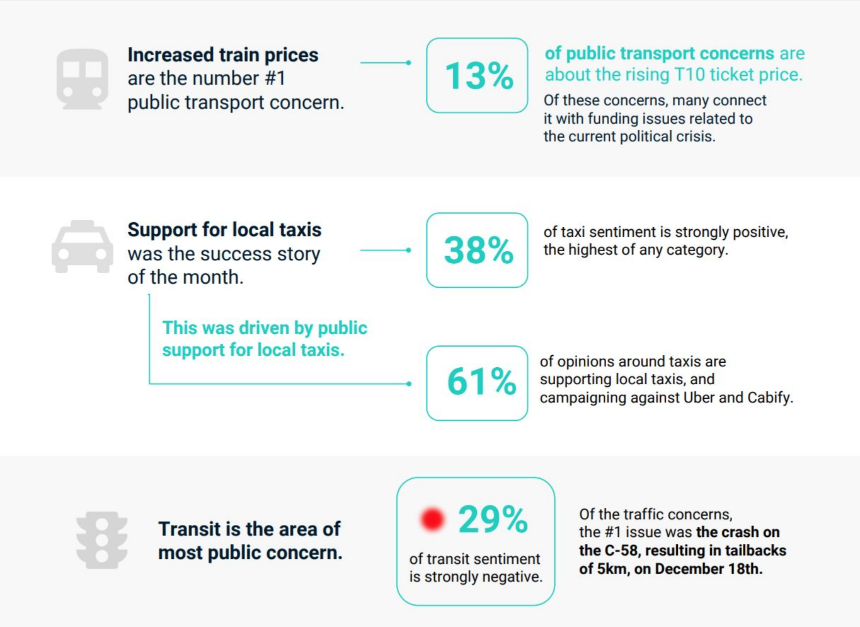
Case Study: Citibeats

Citibeats is a technology startup based in Barcelona that is using NLP to help governments make more informed policy decisions. They offer a text analytics service that uses NLP to show which topics citizens are discussing online and whether their sentiment is positive or negative. Citibeats has already helped a number of Spanish cities design better, smarter policies around areas such as transport, tourism, and the UN's Sustainable Development Goals.

Pictured below are two examples of the kind of analysis Citibeats offers. One tracks public opinion in Madrid and Barcelona on the issue of Catalonian independence. The other shows a detailed breakdown of posts on social media about public transport in Barcelona.



Source: Citibeats, 'Who directs the conversation of the Catalan independence?'. Online at: https://citibeats.net/wp-content/uploads/2018/01/Citibeats_CatalanProces_Citibeats_CaseStudy_EN.pdf



Increased train prices are the number #1 public transport concern.

13%

of public transport concerns are about the rising T10 ticket price.

Of these concerns, many connect it with funding issues related to the current political crisis.

Support for local taxis was the success story of the month.

38%

of taxi sentiment is strongly positive, the highest of any category.

This was driven by public support for local taxis.

61%

of opinions around taxis are supporting local taxis, and campaigning against Uber and Cabify.

Transit is the area of most public concern.

● 29%

of transit sentiment is strongly negative.

Of the traffic concerns, the #1 issue was **the crash on the C-58, resulting in tailbacks of 5km, on December 18th.**

Source: Citibeats, '"Citizens as a sensor" for mobility insights with human meaning'. Online at: https://citibeats.net/wp-content/uploads/2018/02/Mobility_Barcelona_Citibeats_Case-Study-EN.pdf

Sentiment analysis promises to make governments more engaged with their citizens, but some of the biases outlined above threaten to make it less accurate and less useful for people of colour.

Problem 1: Linguistic Variation

The first problem is that sentiment analysis systems may not work on non-standard variants of languages that are spoken by people of colour. In some cases, people of colour's opinions may not be included in the analysis at all. This is the worry that Blodgett and O'Connor have in their 2017 paper, 'Racial Disparity in Natural Language Processing', when they discuss some of the potential implications of their finding that language identification systems are less accurate for AAVE. They imagine a sentiment analysis case where a politician is using a tool like Citibeats that trawls social media to find public posts about a particular topic. They point out that if such a system only analysed posts classified as English, this would under-represent the opinions of those using AAVE because these posts would often be misidentified as being in a language other than English.[33]

Even if posts containing non-standard language do get included in the analysis, there is also a high chance that the system will misunderstand and misclassify their sentiment. In a recent paper on racial bias in hate speech detection systems, Sap *et al.* found that tweets containing elements of AAVE were more likely to be labelled by human annotators as offensive even when they were not, with knock-on effects for how such tweets get treated by automatic hate speech detection models. This finding also has implications for how sentiment analysis systems might deal with such tweets. In the AAVE case, the n-word is often the source of the problems NLP systems have with classifying tone or sentiment. In her interview, Su Lin Blodgett, one of the authors of the paper 'Racial Disparity in Natural Language Processing', drew attention to the way that this slur can have a variety of meanings depending on the context in which it is used and the race of the speaker.[34] These subtleties might be missed by white annotators of training data for a sentiment analysis system, leading the system to classify any instance of the n-word as angry, toxic, or negative.[35]

This problem is not just limited to the n-word. There are many cases in which slang in one variety of a language means something completely different in the standard version of that language. To illustrate this point, we tested two British slang words[36] on IBM Watson's Tone Analyser demo.[37] 'Clapped' in standard English has a positive association, due to its link with applause, but in British slang it refers to something that is unattractive or bad quality. 'Mad' in standard English has a

[33] Blodgett and O'Connor, 'Racial Disparity in Natural Language Processing', p. 3.

[34] See Taylor Jones and Christopher Hall, 'Grammatical Reanalysis and the multiple *N-words* in African American English', *American Speech* (2019). Online at: https://read.dukeupress.edu/american-speech/article-abstract/doi/10.1215/00031283-7611213/139032/Grammatical-Reanalysis-and-the-multiple-N-words-in?redirectedFrom=fulltext (Accessed 20 August 2019).

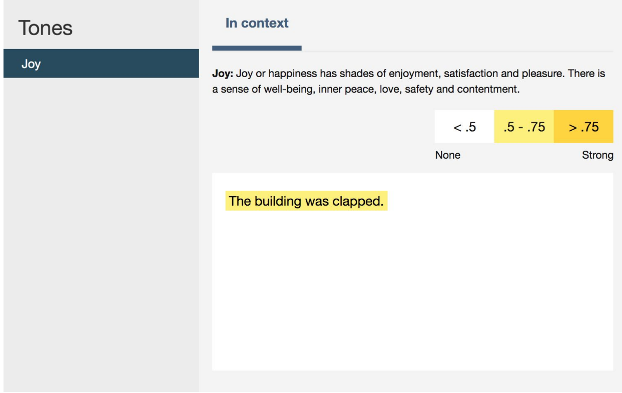
[35] Interview with Su Lin Blodgett.

[36] The slang terms in question are not exclusive to one racial group in Britain. However, due to their roots in working-class London communities where people of colour are strongly represented, any system that fails to process them accurately will disproportionately harm racial minorities.

[37] https://tone-analyzer-demo.ng.bluemix.net/?cm_mc_uid=85199685234515623133774&cm_mc_sid_50200000=64384741564582539583&cm_mc_sid_52640000=73724531564582539603 (Accessed 31 July 2010).

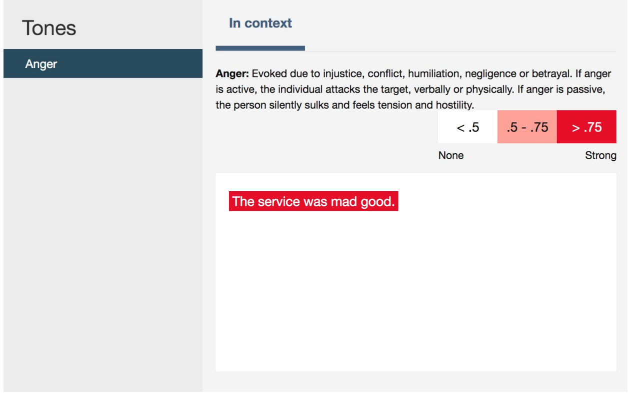
negative sentiment, as it can be used either to mean ‘crazy’ or ‘angry’. In British slang, ‘mad’ is often used as a modifier to increase the intensity of a sentiment (e.g., ‘That exam was mad tough’), but is not always used negatively.

As expected, the Tone Analyser deferred to the standard English meaning of each of these words when classifying the sentiment of our statements. The system identified the sentiment ‘Joy’ in the phrase ‘The building was clapped’, though only to a medium extent. The system strongly identified the sentiment ‘Anger’ in the phrase ‘The service was mad good’.



Joy: Joy or happiness has shades of enjoyment, satisfaction and pleasure. There is a sense of well-being, inner peace, love, safety and contentment.

The building was clapped.



Anger: Evoked due to injustice, conflict, humiliation, negligence or betrayal. If anger is active, the individual attacks the target, verbally or physically. If anger is passive, the person silently sulks and feels tension and hostility.

The service was mad good.

Problem 2: Word Embeddings

The second problem is that sentiment analysis can also be tainted by the bias in word embeddings that we examined above. Aylin Caliskan, author of the paper ‘Semantics derived automatically from language corpora contain human-like biases’, explained in an interview that sentiment analysis can perform poorly in cases where words appear both in negative, prejudiced contexts, and in positive or neutral contexts, because the word vector assigned to them will be based on the most statistically prevalent usage.

> In sentiment analysis, a neutral sentence might be classified as negative if it contains historically biased words. For example, the word ‘gay’ was associated with negative content and appeared in negative contexts in the past few decades, however now it is

> considered a neutral and acceptable word by many. Regardless, word embeddings trained on historical data that contains historical injustices and biases associates the word ‘gay’ with negative sentiment. As a result, encountering such historically charged words in a sentence leads to misclassifications in sentiment classification systems.[38]

Caliskan’s example here concerns sexuality, but there are also cases where words associated with race can have dual meanings. One example is the word ‘ghetto’, used in the US to describe an inner city neighbourhood that usually has a majority of non-white residents. ‘Ghetto’ is often used as a negative term, especially by white Americans, to describe something that is trashy. As an insult, it has a racialised subtext – it implies that the real reason something ‘ghetto’ is bad is because of its associations with black people. We would therefore expect a sentiment analysis system to record posts including the word ‘ghetto’ as expressing a negative emotion. However, many black people, pushing back against the insult’s racialised undertones, have tried to reclaim ‘ghetto’. For example, black feminist @FeministaJones, creator of the Twitter hashtag #LoudBlackGirls, called on black women to reclaim terms like ‘ghetto’ or ‘ratchet’ as part of challenging stereotypes about black women as loud.[39] A sentiment analysis system would misclassify Tweets under this hashtag as expressing negative or angry sentiment due to their use of the word ‘ghetto’, when in fact they are meant as positive and affirming.

Furthermore, the kind of word vector arithmetic explained above, that can bring out racial stereotypes embedded in language, also underpins sentiment analysis systems. Blodgett explained in her interview that when a system encounters an unfamiliar word, it can deduce the sentiment of that word by looking at the sentiments of the words closest to it.[40] If, as we found earlier, words associated with racial minorities tend to be closer in the vector space to negative or unpleasant words than words associated with white people, the system will infer that the former set of words also have a negative or unpleasant sentiment. The result here, similar to in the case of the word

[38] Interview with Aylin Caliskan.

[39] Taryn Finley, ‘Black Women are Reclaiming the “Loud” Stereotype with a Powerful Hashtag’, *Huffington Post*, 15 July 2016. Online at: https://www.huffingtonpost.co.uk/entry/black-women-are-reclaiming-the-loud-stereotype-with-a-powerful-hashtag_n_57891c10e4b0867123e11395 (Accessed 15 August 2019).

[40] Interview with Su Lin Blodgett.

‘ghetto’, is that people of colour writing neutral or positive posts about themselves and their experiences may end up having their sentiment misclassified as negative, simply because of their use of words related to race.

An example of this problem, though not in a system currently used by governments, is a tool called Perspective, developed by Google’s technology incubator Jigsaw. Digital media researcher Anna Woorim Chung highlighted the bias Perspective seems to show towards posts by people of colour in her article ‘How Automated Tools Discriminate Against Black Language’.[41] Chung was at the time using Gobo, a social media platform designed to give users more control over their feed. Rather than relying on Facebook or Twitter’s algorithms to structure their feed, Gobo lets users filter content based various categories including the topic of the post (‘politics’) as well as the sentiment (‘rudeness’ or ‘seriousness’). Gobo relies on the Perspective tool to filter rudeness, and Chung noticed that posts by women of colour were getting classified as ‘rude’ even if they were not. In the article, Chung discusses how Perspective has been criticised for treating statements as ‘toxic’ or ‘rude’ the more identifiers of minority status they contain. A statement ‘I am a black’ is rated as more ‘toxic’ than ‘I am white’; ‘I am a black woman’ is more toxic still, and ‘I am a deaf black woman’ is rated as the most toxic.[42]

Perspective was trained on comments from Wikipedia, which were hand-labelled by human annotators as either a ‘toxic’ or a ‘healthy’ contribution to a discussion. The fact that, based on this dataset, the system learnt that simply identifying oneself as a racial minority is ‘toxic’ reveals how NLP can end up learning and amplifying human biases. People of colour are often accused by white people of ‘playing the race card’,[43] and making discussions more tense and awkward by bringing up race. While we cannot know for sure whether these biases affected the annotation of

[41] Anna Woorim Chung, ‘How Automated Tools Discriminate Against Black Language’, *Civic Media* (24 January 2019). Online at: https://civic.mit.edu/2019/01/24/how-automated-tools-discriminate-against-black-language/ (Accessed 31 July 2019).

[42] *Ibid*.

[43] Andrew Hernández, ‘Let’s Expose the White Double Standard for Playing the Race Card’ https://medium.com/the-establishment/lets-expose-the-white-double-standard-for-playing-the-race-card-4dfee1738f84 (Accessed 16 August 2019).

the Perspective training data, it is easy to imagine how white annotators could carry such attitudes towards discussions of race into their annotation, leading them to mark mentions of race as 'toxic' or 'rude'.

Why Does it Matter?

Racial bias in sentiment analysis is harmful first because it feeds into a long history of people of colour being misunderstood and misrepresented. If governments use technologies that only work on standard varieties of a language, they risk reaffirming the racist beliefs about 'good' and 'bad' language outlined above. This would imply that the only views worth listening to are those of (white) people who speak a certain way. Governments should be especially wary of any such move given their historical role in the suppression and denigration of these varieties. In 1996, when the school board of Oakland, California approved a plan for using AAVE in classrooms to help with reading instruction, the White House condemned the plan. Several states responded by banning the use of AAVE in education, and Oakland's superintendent was called before the U.S. Senate, resulting in the program being dropped.[44] This example illustrates how governments, even in the last few decades, have helped to uphold the stigmas against certain ways of speaking that racially biased NLP tools would further perpetuate.

Furthermore, when sentiment analysis classifies African Americans as 'rude' or 'toxic' for using AAVE slang, or simply speaking about their race, this upholds stereotypes that black people are angry and bitter about race and racism. Many people of colour are understandably frustrated by their experiences of racism, but nevertheless want to have productive conversations about it, or (as in the case of 'ghetto') may want to express pride and positive feelings towards their identity in the face of the racism they experience.

Bias in sentiment analysis also has harmful outcomes for people of colour. Governments use sentiment analysis to try and make their policies more tailored and more responsive to the needs of

[44] W. Brennan, 'Julie Washington's Quest to Get Schools to Respect African-American English', *The Atlantic* April 2018. Online at: https://www.theatlantic.com/magazine/archive/2018/04/the-code-switcher/554099/ (Accessed 20 August 2019).

their citizens, but if these systems are less accurate for people of colour then their voices will be less readily taken into account than the voices of white people. This may lead to policies that do not solve the needs of people of colour, and at worse may actively harm them. Consider, for example, a 2018 study from the Journal of Medical Internet Research, in which sentiment analysis was used on tweets to examine the opinions of patients across the United States about healthcare.[45] If such a study were to directly affect government healthcare policy by encouraging or discouraging investigation of a particular hospital due to its high or low patient ratings, the inaccuracy of sentiment analysis for people of colour could lead to their healthcare concerns being ignored.

This is especially worrying given existing racial inequalities in many countries that lead to people of colour feeling that governments do not represent them or listen to their views. Fraught relations with the police are one well-documented source of friction between people of colour and the government. The Runnymede Trust in the UK found in interviews with working class people of colour that there was a 'trust deficit' between ethnic minorities and the police and criminal justice system.[46] They also reported that young black and Asian men in particular are disproportionately subject to stop and search and the use of force by the Metropolitan police in London.[47] In the US, a National Public Radio study found that 61 percent of African Americans believe their local police are more likely to use unnecessary force on a black person than on a white person in the same situation, with that statistic rising to 73 percent among African Americans living in the Midwest.[48]

[45] K. C. Sewalk, G. Tuli, Y. Hswen, J. S. Brownstein, and J. B. Hawkins, 'Using Twitter to Examine Web-Based Patient Experience Sentiments in the United States: Longitudinal Study', *Journal of Medical Internet Research* 20(10) (2018). Online at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6231860/ (Accessed 16 August 2019).

[46] The Runnymede Trust, '"We Are Ghosts": Race, Class and Institutional Prejudice', p. 25.

[47] *Ibid*., p. 24.

[48] National Public Radio, the Robert Wood Johnson Foundation and the Harvard T. H. Chan School of Public Health, 'Discrimination in America: Experiences and Views of African Americans' (October 2017), p. 17. Online at: https://www.npr.org/assets/img/2017/10/23/discriminationpoll-african-americans.pdf?t=1565701360191 (Accessed 16 August 2019).

Besides just relations with specific government bodies such as the police, people of colour often report lower confidence or trust in the government more generally: in the US, black Americans often report lower trust in the federal government, though typically only under Republican Presidents. Currently levels of trust stand at 17 percent for whites and only 9 percent for blacks.[49] Meanwhile, reporting on the *banlieues* in Paris (which have significant populations of North African descent), *The Atlantic* found that many residents felt that the government has turned its back on their problems. They pointed out that the (mostly white) *gilets jaunes* protests this year have attracted far more attention than the riots in the *banlieues* in 2005, even though many of the same issues were being raised.[50]

All of this evidence points to a climate in which people of colour feel alienated from the government, feeling only exacerbated by the current swell of right-wing populism in the US and across Europe. In such a climate, deploying technologies that further marginalise the voices of people of colour could deepen existing tensions and exacerbate racial inequality.

Dialogue Agents

Besides sentiment analysis, governments are also making use of dialogue agents to connect with their citizens. Dialogue agents are NLP systems that are able to have conversations with humans. They therefore combine both the processing and the generation of natural language. Tay, the racist chatbot discussed earlier, is an example of a dialogue agent, as are Amazon's Alexa and Apple's Siri. Dialogue agents are increasingly being used in customer-service-oriented tasks, as they are cheaper and often more efficient than employing many human agents to deal with queries. Along with commercial actors, governments have taken an interest in using dialogue agents to help those accessing government services.

[49] Pew Research Center, 'Public Trust in Government: 1958-2019', 11 April 2019. Online at: https://www.people-press.org/2019/04/11/public-trust-in-government-1958-2019/ (Accessed 16 August 2019).

[50] Rachel Donadio, 'France's Double Standard for Populist Uprisings', *The Atlantic*, 26 February 2019. Online at: https://www.theatlantic.com/international/archive/2019/02/paris-banlieues-yellow-vests-double-standard/583555/ (Accessed 16 August 2019).

Example: Amelia

Amelia is a dialogue agent designed by IPSoft. The system has been used in a number of industries including banking, insurance and retail, and has also been used in government. Enfield Council in London adopted Amelia to help deal with queries about their services. As the volume of demand for services increases but central government spending cuts deepen, the council hopes that using Amelia will be more cost-efficient in dealing with their constituents.[51] The case study explains part of Amelia's attraction:

> Rather than requiring diverse visitors to be technology-literate, Enfield Council will require that their technology be 'people-literate.' Given the fact that Amelia interacts using natural language, the expectation is that she will be well-placed to support everyone.[52]

IPSoft markets Amelia as being a 'digital colleague' or a 'cognitive agent' rather than a chatbot or virtual assistant. Some of the differences they identify between Amelia and other available chatbots are:

(1) Amelia is more flexible than chatbots that are programmed to follow a particular conversation order, and have pre-written responses to particular inputs. It has a level of 'social talk' that allows it to converse with those who use informal language.
(2) Amelia can 'navigate conversational chaos' by following a conversation even if the user switches topics quickly. A chatbot that has to answer questions in a specific sequence could not handle this switching.
(3) Amelia uses machine learning to improve over time, and learn from its interactions with users.[53]
(4) Amelia uses sentiment analysis techniques to detect users' emotion and mood so that it can tailor its responses accordingly.[54]

The example of Amelia shows that governments, along with many other organisations using dialogue agents, want increasingly sophisticated NLP systems that can converse with users in an

[51] IPSoft Case Study, 'Enfield Council: Public Service Virtual Agent'. Online at: https://www.ipsoft.com/wp-content/uploads/2017/11/Case-Study-Enfield-Council_PDF.pdf (Accessed 1 August 2019).
[52] *Ibid.*
[53] E. Dashevsky, 'Four Big Differences Between a Chatbot and a Digital Colleague', IPSoft (16 July 2018). Online at: https://www.ipsoft.com/2018/07/16/four-big-differences-between-chatbot-digital-colleague/ (Accessed 1 August 2019).
[54] IPSoft, 'The Science Behind Amelia'. Online at: https://www.ipsoft.com/amelia-science/ (Accessed 1 August 2019).

almost human-like fashion. However, as these technologies are required to perform more complex tasks concerning natural language, they grow more vulnerable to a number of the racial biases we have outlined.

Problem 1: Misunderstandings

Dialogue agents, like sentiment analysis systems, are likely to have been trained on standard language rather than non-standard varieties. This means they are unlikely to understand what a person of colour who speaks such a variety says to them. In order to use the tool, people of colour may be forced to 'code switch'. 'Code switching' refers to the practice of changing your language or accent, and in a race context it is used specifically to explain how people of colour often have to speak in a 'whiter' fashion in particular settings.[55]

Furthermore, many dialogue agents use sentiment analysis to help track the users' mood and respond appropriately. The fact that sentiment analysis is likely to misinterpret and misclassify the tone of people of colour (see above) means that these agents may respond inappropriately. Users of colour may find that agents use placating words and phrases even when they are not expressing anger, simply because the way they speak or the race-related terms they might use to describe themselves and their queries could be misclassified as angry.

Problem 2: Offensive Language

One of the most pressing issues dialogue agents face, due to the fact that they need to generate their own text, is the problem of offensive language. If they learn from their past interactions with users, they are vulnerable to learning offensive language and hate speech, especially if (as we saw with Tay) some people go out of their way to exploit this feature. However, a blacklisting response to this issue could prevent a dialogue agent from speaking adequately about race. In their paper 'Let's Talk About Race: Identity, Chatbots, and AI', Schlesinger *et al.* highlight how blacklisting certain strings used as slurs in different contexts (such as 'Jap...' or 'Paki...') has the effect of

[55] AT McWilliams, 'Sorry to Bother You, black Americans and the power and peril of code-switching', *The Guardian* 25 July 2018. Online at: https://www.theguardian.com/film/2018/jul/25/sorry-to-bother-you-white-voice-code-switching (Accessed 16 August 2019).

filtering out words like ‘Japan’, ‘Japanese’, ‘Pakistan’ and ‘Pakistani’. This curtails the ability for the dialogue agent to talk about a whole range of issues relating to those countries and their residents.

Blacklisting is also a crude method because racist language is not just about slurs. Alex Taylor, one of the authors of the ‘Let’s Talk About Race’ paper, spoke in his interview about the issue of context. Many examples of offensive statements are highly context-specific, and exactly the same language can be offensive in one context but not in another. He gave the example of the question ‘Where are you from?’, which is often innocuous when addressed to a white person, but has very different connotations when addressed to a person of colour. This means that just preventing these systems from using slurs is not enough to ensure that they never say something that could hurt or offend a user of colour.

Some past examples even suggest that dialogue agents would be more likely to cause this kind of context-specific offense than a human agent. An article published this month in *E-Content Magazine* explains how a telecommunications software company had to stop using their chatbot, Sally, after less than a month due to user complaints. Sally was programmed to use details like the customer’s name, location, and the time of year to make lifelike small talk. However, some users took offense to apparent profiling, such as when Sally asked a female user around Christmas time if she was busy making dinner for her family.[56] In this case, Sally did not just inadvertently say something that is offensive in one context but not in another (asking a man the same question would not be sexist). It actually used the user’s gender to draw problematic assumptions about what she was likely to be doing. There is a risk that a dialogue agent, using someone’s name or location as a proxy for their race, could stereotype in ways that are racist as well as sexist.

There is also the issue of how NLP systems should respond when they encounter offensive language. In 2017, journalists at *Quartz* tested how a number of voice assistants gendered as female responded to sexual harassment, and found they were either programmed to give a coy

[56] Matthew Grocki, ‘AI Needs Content Strategy More Than Ever’, *E-Content Magazine* 13 August 2019. Online at: http://www.econtentmag.com/Articles/Column/Natural-Content-Practices/AI-Needs-Content-Strategy-More-Than-Ever-132719.htm (Accessed 16 August 2019).

response (such as Siri responding 'I'd blush if I could' when a user calls it a 'bitch') or did not register it (Alexa responds 'Well, thanks for the feedback' to the same insult).[57] Because of the way that virtual assistants are treated or viewed as human-like by their users, this failure to stand up to harassment risks teaching people that such language is somehow OK. In response to the concerns, Amazon programmed Alexa with a 'disengage mode', and it now responds to sexually explicit questions by saying either 'I'm not going to respond to that' or 'I'm not sure what outcome you expected'.[58]

Similarly, developers may wish to programme bots to stand up to, or at least to disengage from, racist abuse. However, recent research has suggested that automatic hate speech detection often discriminates against people of colour, especially those using languages like AAVE where they may use reclaimed slurs or other words that are offensive in a standard English context. Sap *et al*. found that human annotators are likely to mislabel Tweets containing AAVE as offensive.[59] Any system that relied on similar annotated datasets to learn what is and is not offensive also risks misreading the language of people of colour.

Why Does it Matter?

When government spending comes under pressure, more and more agencies may look to dialogue agents as a cost-saving measure if they prove cheaper than using humans. Dialogue agents therefore have the potential to become a prominent face of public services. If these agents are unable to process non-standard language, or end up saying something racist, using them will be a frustrating experience for people of colour, and will make their experience of using public services markedly worse than that of their white counterparts. At the moment, this would affect

[57] Leah Fessler, 'We tested bots like Siri and Alexa to see who would stand up to sexual harassment', *Quartz* (22 February 2017). Online at: https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment/ (Accessed 1 August 2019).

[58] Leah Fessler, 'Amazon's Alexa is now a feminist, and she's sorry if that upsets you', *Quartz* (17 January 2018). Online at: https://qz.com/work/1180607/amazons-alexa-is-now-a-feminist-and-shes-sorry-if-that-upsets-you/ (Accessed 1 August 2019).

[59] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, 'The Risk of Racial Bias in Hate Speech Detection', *ACL* (2019). Online at: https://aclweb.org/anthology/papers/P/P19/P19-1163/ (Accessed 20 August 2019).

only a handful of services that are mainly administrative, but as dialogue agents become more prevalent, and even offer sensitive services such as mental health counselling,[60] the potential damage grows more serious. A virtual therapist that misread a user's mood, or mistakenly treated them as 'aggressive' due to their use of language, could risk worsening their mental state.[61]

Picture of Amelia
Source: IPSoft Case Study, 'Enfield Council: Public Service Virtual Agent'

The problems people of colour experience with dialogue agents could reinforce the notion that their governments do not represent them, and do not respect the language they speak (see above). It is significant that dialogue agents not only speak in standard, 'white' language; their avatars are often racialised. Amelia, for example, is depicted as a blonde white woman. If the public face of government becomes white, this will be in spite of the fact that in both the US and the UK, black people make up a disproportionate number of public sector and federal employees.[62] Black Americans make up 12 percent of the US population, but over 18 percent of the federal workforce. 42.8 percent of all black British workers are in public administration, education and health, compared to 29.5 percent of white workers.[63] These statistics make it much more likely that a black person accessing public services through a human would end up speaking to someone who shares their background or speaks their language than if they speak to a 'white' dialogue agent.

Obviously, the problems for people of colour accessing services from governments go deeper than dialogue agents, and human public service workers can be racist or say offensive things. However,

[60] D. Browne, M. Slozberg, M. Arthur, 'Do Mental Health Chatbots Work?', *Healthline* 6 July 2018. Online at: https://www.healthline.com/health/mental-health/chatbots-reviews#1 (Accessed 16 August 2019).
[61] Henderson et al., 'Ethical Challenges in Data-Driven Dialogue Systems', p. 5.
[62] J. Lartey, '"Barely above water": US Shutdown hits black federal workers hardest', *The Guardian* 11 January 2019. Online at: https://www.theguardian.com/us-news/2019/jan/11/governmnet-shutdown-black-federal-workers-trump-border-wall (Accessed 16 August 2019).
[63] 'Employment by sector', *Ethnicity Facts and Figures*, 10 October 2018. Online at: https://www.ethnicity-facts-figures.service.gov.uk/work-pay-and-benefits/employment/employment-by-sector/latest (Accessed 16 August 2019).

humans can be trained to be more sensitive to the context-specificity of race in ways that NLP systems cannot. In his interview, Alex Taylor stressed that current approaches to NLP are founded on the idea that you can deduce the meaning of language without an understanding of context (also called 'proxemics').[64] This means there is no quick technical fix for NLP systems that is analogous to the diversity and racial sensitivity training that humans can receive.

Many of us will have experienced the frustration of a virtual assistant failing to understand or respond adequately to our queries. However, as technology advances to make dialogue agents more sophisticated, white people may experience an increased quality of service while people of colour continue to be frustrated. If dialogue agents proliferate to be the first point of contact across government services, then people of colour may find that one of their most frequent experiences of government is being misunderstood.

[64] Interview with Alex Taylor.

Recommendations

In spite of the growing academic interest in the issue of bias in NLP, my interviewees highlighted that this research has yet to meaningfully affect existing commercial tools that governments would want to use, especially when it comes to the issues surrounding offensive language and linguistic variation.[65] Some debiasing techniques have been developed as a potential fix for the issues relating to word embeddings,[66] but there is disagreement about how successful this is at removing unfairness.[67] Given the reality that these technologies are racially biased, what should governments do if they use or hope to use NLP systems?

Before outlining our recommendations, it is important to emphasise what is at stake when NLP tools can be racially biased. In her interview, Su Lin Blodgett observed that taking the issue of racial bias, or any bias, in NLP seriously can be difficult because the effects of that bias can seem less obvious or less visceral than in other cases of AI and unfairness.[68] If a judicial sentencing tool is racially biased, the outcome is that some people unfairly get confined to prison for longer. If a facial recognition tool used by the police is inaccurate, the wrong person may end up being arrested and detained. If NLP systems are biased, the results may seem less serious (a dialogue agent misunderstanding you) or more diffuse (a system replicating biases implicit in human language).

However, bias in NLP systems is something that should concern us, and the problem of bias should encourage governments to err on the side of caution. Language is the medium through

[65] See e.g. Interview with Su Lin Blodgett: ‘There is great interest in linguistic variation (particularly with regard to non-standard varieties) from a technical perspective, but because researchers and auditors do not have access to the internal workings of commercial applications, it is not clear what the commercial interest is or if there are changes in the works.’ Interviews with Ari Schlesinger and Alex Taylor also confirmed that to their knowledge there had been no developments in dealing with offensive language in NLP systems.

[66] Bolukbasi et al. (2016) propose debiasing techniques in their paper; see also Y. Qian, U. Muaz, B. Zhang, and J. W. Hyun, ‘Reducing Gender Bias in Word-level Language Models with a Gender-Equalizing Loss Function’, *ACL* (2019). Online at: https://www.aclweb.org/anthology/P19-2031 (Accessed 2 August 2019).

[67] H. Gonen and Y. Goldberg, ‘Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them’, *ACL* (2019). Online at: https://www.aclweb.org/anthology/N19-1061 (Accessed 2 August 2019).

[68] Interview with Su Lin Blodgett.

which we are able to interact with the world – to express our thoughts and to form human connections. If NLP technologies become an inescapable interface through which many of these interactions must take place, then the harms of bias can penetrate through to our self-expression and our ability to participate as an equal in society. When the issue concerns how citizens can interact with governments, the stakes are even higher, considering the influence that the state has over our lives. If we do not confront the issue of racial bias in NLP systems, governments could inadvertently tip the balance of power away further away from marginalised communities.

Based on the findings of this report, we recommend that governments follow three principles when deciding why and how to deploy NLP tools:

- **Specificity**: Delimiting a clear task for which NLP tools could be helpful.
- **Transparency**: Being clear about where NLP tools are being used and how they work.
- **Accountability**: Having clear and accessible channels for reporting problems with a government NLP system.

These principles can in many cases be followed within existing policy guidelines to minimise the risks of racial disparity. In particular, we recommend that governments across the world follow something like the UK Government Digital Service (GDS) process for commissioning new IT projects. The GDS Service Standard splits projects into four phases, based on the principles of agile delivery.

1. **Discovery phase**: Understanding the problem that needs to be solved.
2. **Alpha phase**: Building and testing different prototypes in response to problem outlined in the discovery phase.
3. **Beta phase**: Taking the best idea from the alpha phase and building a real tool for users.
4. **Live phase**: Running the new service in a sustainable way, and continuing to make improvements.[69]

[69] Gov.UK Service Manual, ‘Agile Delivery’. Online at: https://www.gov.uk/service-manual/agile-delivery (Accessed 19 August 2019).

Based on this approach, we consider how each of the three principles of specificity, transparency and accountability fit into the overall timeline of an agile project.

User Research (Discovery Phase)

The discovery phase is used to understand the likely users of a new service such as a dialogue agent, and the problems they currently experience. This user-first approach is critical for ensuring that NLP tools do not end up harming the users they are meant to help. In her interview, academic Ari Schlesinger emphasised that this dialogue between governments and their citizens is crucial to ensure the ethical use of NLP:

> Above all else, have you had meaningful conversations and really listened to the people whose problems you're trying to solve? Is this the solution they're looking for, or is this just something that helps you?[70]

The discovery phase helps governments be **specific** about the task for which they think NLP would be helpful, by considering particular user needs and problems rather than vague goals like 'making government more responsive'. Part of the discovery phase involves scoping the project, to ensure that the problem to be solved is clearly delimited.

The discovery phase also encourages user research that recognises the **specific** needs of different communities. The GDS Service Standard rightly highlights the importance of learning about users' accessibility requirements, especially disabilities.[71] Based on this report, we would argue that the need to consider a diverse range of races and ethnicities is equally important.

The importance of racial bias in new technologies is recognised by many governments already. The UK GDS and the Canadian Digital Service (CDS) recently partnered to help the younger CDS build

[70] Interview with Ari Schlesinger.

[71] See Gov.UK Service Manual, 'How the discovery phase works' and 'User research in discovery'. Both online at: https://www.gov.uk/service-manual/agile-delivery/how-the-discovery-phase-works and https://www.gov.uk/service-manual/design/scoping-your-service (Accessed 19 August 2019).

user-centred digital services, and one of the issues raised was the need for a diverse user research community, to cater to linguistic, cultural and ethnic differences between users.[72] We recommend that the definition of accessibility is expanded to include race, especially in cases where people of colour may speak distinctive variants of a language.

Designing and Building a Product (Alpha and Beta Phases)

During the alpha and beta phases of a project, a potential NLP system would be prototyped and then developed into a full service.

At this stage of the project, governments need to be **specific** about what current NLP tools can and cannot do. This involves engaging with developers and academics to understand the kinds of technical constraints outlined above, such as issues around linguistic variation. This understanding will help revise and refine the scope of the project. For example, if we know that most commercially available tools perform poorly when text is in non-standard English, we might question the utility of a sentiment analysis system that combs social media – where people are more likely to speak in an informal or vernacular way. We might instead choose to use a system only on feedback forms filled out on a government website, based on the assumption that people already somewhat self-monitor their own language on these forms and use standard English (although this is an assumption that would need to be tested).

Transparency is also important at this stage of the project. GDS recommends ‘working in the open’ during the alpha,[73] and this can be an important way to gain the input of academics, developers and potential users who have thoughts about how an NLP tool could be developed.

[72] Steph Marsh, ‘Learning and sharing knowledge with the Canadian Digital Service’, *Gov.UK Blog* 19 July 2019. Online at: https://userresearch.blog.gov.uk/2019/07/19/learning-and-sharing-knowledge-with-the-canadian-digital-service/ (Accessed 19 August 2019).
[73] Gov.UK Service Manual, ‘How the alpha phase works’. Online at: https://www.gov.uk/service-manual/agile-delivery/how-the-alpha-phase-works (Accessed 19 August 2019).

We also recommend a further **transparency** requirement that is specific to the issue of NLP: resisting the tendency to over-anthropomorphise NLP tools. Because these systems ‘understand’ and ‘speak’ human languages, it is easy to expect them to possess human-like abilities even though they are ultimately just advanced statistical processing machines. This recommendation is especially applicable to the use of dialogue agents, where companies give their agents names and human avatars, and refer to them with human pronouns.

This recommendation mitigates the risk of bias in NLP in two ways. First, abandoning the desire to create ‘human-like’ dialogue systems avoids a number of the sources of bias outlined above. Many of the concerns about offensive language and stereotyping are relevant only in the case where a dialogue agent is expected to make small talk, or follow a natural flow of conversation. Chatbots that follow a pre-programmed script may have more limited and less human-like conversations, but do not stray into the risky territory of unstructured conversation where the potential for bias to creep in is higher.

Second, even if dialogue agents do still have some problems, especially around linguistic variation, we would argue that these issues will cause less harm in a system that is less anthropomorphised. It is frustrating to be misunderstood, whether by a person or a machine. However, because we tend to have higher expectations from our interactions with people than our interactions with machines, being misunderstood by the former can be more hurtful. For example, we would all be more offended by a human cashier who showed blank incomprehension or even ignored us when we spoke than we would be by a self-checkout machine that malfunctioned during our transaction. The tendency to create dialogue agents that look and behave like people, especially given that these agents are typically racialised as white, blurs the line between human and machine. It means that a user of colour coming away from an interaction feeling frustrated or even offended is more likely to feel like a person has wronged them rather than that there has been a technical glitch with a machine.

Operating the System (Beta and Live Phases)

During the beta and live phases of a project, the new NLP tool would be rolled out to actual users. However, the GDS Service Standard emphasises the need to continue iterating and improving the system even once it is in use.

At this stage of the project, governments need to rely on the **specificity** developed in the earlier phases to ensure there is a clear metric for performance. For example, with a sentiment analysis system, governments may need to run concurrent normal polling for a time to determine whether the system accurately captures the public mood. They may also poll citizens to see whether there is any increase in public satisfaction with policy, or some other indicator that sentiment analysis is actually working. Here, it would be important to remember the **specific** communities considered in the discovery phase, and make sure that their views are well represented in evaluations of performance. We recommend that governments consider a racial disparity audit, run as part of a broader accessibility audit, to address the effects on people of colour of new NLP technologies.

Governments must also be **accountable** for these performance metrics. If a system underperforms and shows evidence of racial disparity, there needs to be a clear mechanism for recalling and/or updating the system to tackle this issue. Besides their own review processes, governments must also be **accountable** to the users of the service. Channels for reporting an issue need need to be **transparent** – the information on which department or individual is responsible for a particular system and how to contact them should be publicly available and easy to find. These accountability mechanisms also need to be accessible, especially where users of colour most at risk of being affected by bias could experience many types of marginalisation (economic, political, educational) that mean they are less likely or less able to make their feelings known. In such cases, simply having a system in place for reporting flaws in a new system may not be sufficient, and governments will need to come up with other ways (such as hosting community outreach events) to make themselves accountable to their citizens of colour.

The GDS Service Standard also recommends that the final system be **transparent** in terms of the software being open source. The advantage of open source software is that it facilitates public scrutiny of the way the new system is coded. This scrutiny can help identify sources of bias, even if users have not yet been affected by them, allowing governments to modify technologies before, not after, they cause harm. Open source software therefore also reinforces **accountability**.

Conclusion

If governments follow the recommendations outlined above, they will reduce the risk of perpetuating or amplifying racial bias through the use of NLP tools. It is worth emphasising that the advantage of the GDS Service Standard is not just that it leads to government systems that are fairer and less biased. It also prevents inappropriate projects from being rolled out to users, through the agile delivery process. If a discovery phase finds that sentiment analysis is not the best way to connect with a particular community, or if an alpha phase finds that a dialogue agent sometimes stereotypes and offends its users, the project can be terminated at this stage. As highlighted above, it can be easy to forget that the harms of racial bias in NLP still warrant caution on the part of policy-makers, as much as the harms of racial bias in sentencing tools or in facial recognition.

Our recommendations may well mean that governments lag behind equivalent private sector NLP technologies in terms of technical sophistication, and this may need explicit defense to the general public. The Centre for Public Impact's working paper on artificial intelligence in government cautions that where governments do not keep up with the pace of technological development in the private sector, this could undermine the legitimacy of the government.[74] This is because citizens' expectations of services are shaped by their experiences in the private sector, and so if they consistently find government services lacking compared to the private sector, this will affect their perception of the government as a whole. However, based on the findings of the present report, we suggest that governments who try to keep up with the private sector's increasingly anthropomorphised dialogue agents, or sentiment analysis systems that promise to identify more and more emotions, risk exacerbating racial inequality.

The existence of racial bias in NLP is worrying across all spheres, public and private, but the government has particular responsibilities when it comes to tools like sentiment analysis and

[74] Centre for Public Impact, 'Destination Unknown: Exploring the impact of Artificial Intelligence on Government' (September 2017), p. 37.

dialogue agents that are meant to increase citizen participation in government. One of the central issues in democracy is whose voices do (and do not) get heard. This report has demonstrated that current AI tools contain many worrying features that could lead to disparate outcomes for different racial groups. Its recommendations are designed to avoid perpetuating those inequalities, and to ensure that governments strive to be inclusive of the voices of all of its citizens, no matter how they speak.