QNAS

QnAs with Sharad Goel and Allison Koenecke

Sandeep Ravindran, *Science Writer*

Sharad Goel works at the interface of computer science, statistics, and the social sciences. An assistant professor of management science and engineering at Stanford University, Goel has applied computational and statistical techniques to study a variety of socially relevant, policy-related topics, including voter fraud and political polarization. Goel and a graduate student in his laboratory, Allison Koenecke, have also been interested in studying fairness in algorithmic systems, a focus of his PNAS article on racial disparities in automated speech recognition published earlier this year (1). Goel and Koenecke discuss their findings with PNAS.

PNAS: How did you become interested in studying racial disparities in automated speech recognition (ASR)?

Goel: Most of my research is aimed at measuring and reducing social stratification. We had been working to understand some of the more theoretical aspects of algorithmic fairness, and the different ways in which one can evaluate the extent to which a machine-learning system is "fair." There are various mathematical definitions of fairness, but we wanted to see how real systems operate and understand the ways in which design choices can lead to systematic disparities.

Koenecke: We were initially interested in this problem because of the many applications of voice-recognition systems in society. Much of this project was motivated by the work of Joy Buolamwini, a computer scientist at the [Massachusetts Institute of Technology] Media Lab, and Timnit Gebru, a research scientist at Google, who found disparities by race and gender in the context of computer vision, specifically in facial recognition by a few large corporations (2). We wanted to find out whether this would also apply to speech recognition. There are a lot of places, from voice assistants to medical and court transcription, where these ASR applications may be in use and could lead to further racial disparities.

PNAS: You looked at racial disparities in commercial speech-to-text tools developed by five large technology firms. How did you go about studying this?

Sharad Goel. Image credit: Sharad Goel.

Koenecke: The first challenge was finding new datasets that would not have been used as training data by these companies. By collaborating with the Stanford linguistics department, we found new linguistic datasets that were representative and had similar interview formats for both White speakers and Black speakers. We compared the human-generated transcriptions from each of the datasets to the ASR-generated transcriptions from the five commercial speech-to-text tools. Then we assessed the word error rate, the standard metric that's used by linguists to express the accuracy of a transcription. [The study assessed the five commercial ASR services by accessing their public speech-to-text interfaces either through an application programming interface or a software development kit. The ASR algorithms tested may not be the same as those used in proprietary technologies, although they may share similarities. The researchers accessed the ASR services in 2019, and changes since then may cause the services to function differently at present.]

PNAS: What did you find?

Koenecke: We found that all five of these systems exhibited racial disparities as measured by average

First published August 10, 2020.

www.pnas.org/cgi/doi/10.1073/pnas.2015356117

word error rate: 0.35 for Black speakers versus 0.19 for White speakers, which is roughly a doubled error rate for Black speakers.

PNAS: What were the underlying reasons for these racial disparities?

Koenecke: We can't know what exactly is under the hood in the five ASR systems that we studied, but most modern ASR systems tend to use two components as part of their models. One is the language model, which works on what you're saying, so things like grammar and lexicon. The second component is the acoustic model, which works on how you say something, so things like intonation and patterns of stress. We did a series of different analyses and found that the racial disparities in ASR performance are linked to the acoustic model. This is likely related to the acoustic differences between African American vernacular English and standard English, such as in pronunciation and prosody [the patterns of stress or intonation].

PNAS: What are some of the implications of your findings?

Koenecke: The implications are far-reaching, and could affect anyone from people with disabilities who are using these speech-to-text systems to interact with web browsers, to doctors who are using these medical transcription systems to record their patient notes. African American vernacular English speakers could be harmed if companies are still creating products that are disproportionately bad at transcribing their speech.

PNAS: What can companies do to prevent these sorts of racial disparities?

Koenecke: We believe that the speech-recognition community should invest more broadly in ensuring that these ASR systems are inclusive. This ranges from making sure that their training data are inclusive, to making sure that the engineers working on these systems are diverse and care about ensuring that different varieties of English are represented all of the way down the pipeline. In particular, we hope that companies can collect more diverse data both of African American vernacular speech and of other varieties of English. Lastly, we hope that developers regularly assess and publicly report their progress over time by evaluating their word error rates on different sets of test data and showing that they are making improvements across different varieties of English.

Allison Koenecke. Image credit: Allison Koenecke.

Goel: I think we also need to be more intentional when developing these types of systems. A big underlying problem is that issues of equity typically aren't foregrounded in the development process. That's really the fundamental barrier that we're up against. Once equity is top of mind, it often becomes clear what you need to do to ensure that systems are inclusive. In this case, going out and collecting more diverse datasets would have gone a long way. But until we get to the point where equity is part of the grander design objective, we're going to continue seeing systems that shut out large groups of people and exacerbate existing inequities.

PNAS: Where do you see this project going in the future?

Goel: We hope these results prompt developers, researchers, and government agencies to work to ensure that this technology is broadly inclusive in the future.

1 A. Koenecke *et al.*, Racial disparities in automated speech recognition. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 7684–7689 (2020).

2 J. Buolamwini, T. Gebru, (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the Conference on Fairness, Accountability and Transparency. Friedler SA, Wilson C, Eds. (Association for Computing Machinery, New York, NY, 2018). 77–91.