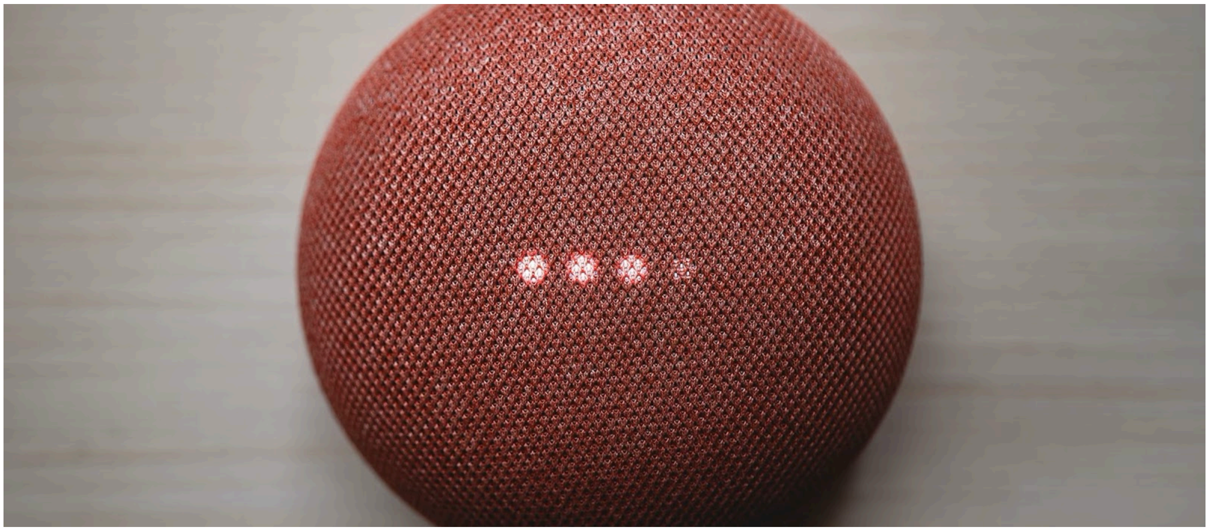
This caselet was developed by student Chloe Evered as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and supervised by Dr. Vandana P. Janeja, Dr. Christine Mallinson and Dr. Karen Chen at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.



Spotting Unintended Harms of Algorithmic Bias

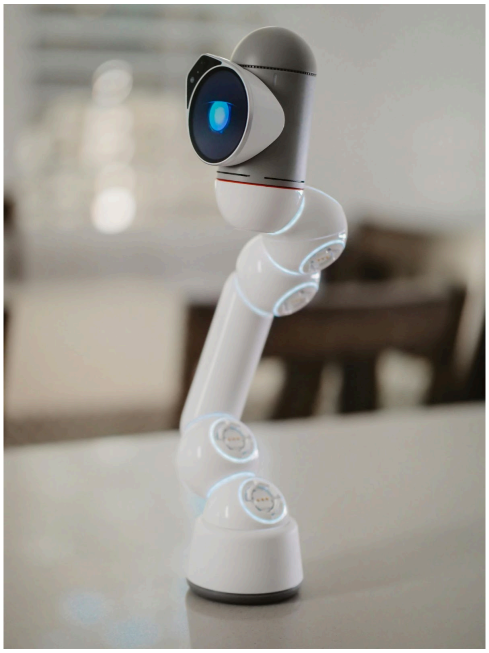
In this caselet, you'll explore a benevolent use case for speech recognition and synthesis technology. Even when these technologies aren't employed for malicious purposes, like deepfakes, they can still cause unintended harms. Read the background information carefully, and use it to answer the multiple choice questions that follow.

By the end of this caselet, you will have a better understanding of linguistic discrimination and a framework for understanding the harms that biased algorithms can cause.

Problem Context

As a data scientist, you've worked hard to combat harms caused by audio deepfakes—malicious applications of voice cloning and speech synthesis technology—first, for a news network fighting disinformation, then for a bank protecting customers from fraud. There is now an exciting new opportunity for you to help develop a speech synthesis system with a positive application: a voice-enabled chatbot for a game show. This televised game show is science-fiction themed, so your client wants an automated chatbot to interact with contestants in the place of a human host. This chatbot will involve an automatic speech recognition (ASR) system to convert spoken language into text as well as a text-to-speech (TTS) system, so the contestants can interact with it with only their voices.

This caselet was developed by student Chloe Evered as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and supervised by Dr. Vandana P. Janeja, Dr. Christine Mallinson and Dr. Karen Chen at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.



You successfully build this chatbot, and the production company uses it in an untelevised rehearsal one month before the show airs.

It's great at both presenting questions to contestants and responding to their answers!

However, on the first day of filming with real contestants, something goes awry. When one of the contestants, who is from Australia, answers questions correctly, the chatbot mistakenly outputs that their answer was wrong, and gives the next contestant the opportunity to steal. This does not seem to be a problem for the other contestants, who are all from the United States. In effect, the Australian contestant is being penalized purely because of the way they speak: in other words, their dialect or language variety.

Q1: Why is the chatbot unable to recognize this contestant's answers?

a. The Australian contestant has a heavy accent, and the speakers from the data the chatbot was trained on did not have accents.
b. The Australian contestant's variety of English was not well-represented in the audio data the chatbot was trained on, which primarily came from speakers of American English varieties.

Understanding potential harms of algorithmic bias

Importantly, there is no such thing as un-accented speech. We all have accents! Often, when we perceive someone to have no accent, that speaker probably has a similar variety to our own or one that we consider to be a "standard" way of speaking. Linguistic diversity is a normal and natural part of human language and culture, but it is not always attended to in the development of speech technologies: for example, ASR systems are often bad at recognizing certain varieties

This caselet was developed by student Chloe Evered as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and supervised by Dr. Vandana P. Janeja, Dr. Christine Mallinson and Dr. Karen Chen at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

of spoken English, particularly African American English (Martin & Wright, 2023). Furthermore, an overwhelming majority of resources and datasets for audio deepfake detection are built around only English speech data, though there are thousands of other languages spoken around the world. The lack of resources for deepfake detection in languages other than English, and particularly for non-spoken (signed) languages, is a significant issue. In this caselet, though, we focus on harms that can arise when a system is biased with respect to different dialects, or varieties, of English.

Failure of an ASR system to recognize speech of a particular language variety is an example of algorithmic bias, or computational discrimination, in which systems replicate, reinforce, or magnify existing social biases and prejudices against a specific group of people. As you can imagine, a multitude of different harms can arise from algorithmic bias and discrimination. These harms fall under two broad categories: allocational and representational. Allocational harms refer to the unfair distribution of resources and opportunities as a result of bias in a system. Representational harms, on the other hand, can be harder to quantify and to spot. They have to do with widespread, often implicit, social and cultural beliefs. A computational system that reflects or reconstructs negative stereotypes about a group of people, for example, has potential for representational harms.

In review, allocational harms are **immediate, easily quantifiable, and discrete**, while the consequences of representational harms are more **long-term, difficult to measure, and cultural** (Crawford, 2017).

Q2: Consider the following two hypothetical examples of harm that could arise from the previously described scenario of the Australian game show contestant. Which type of harm (allocation or representation) is reflected in each example?

Scenario A: The game show implements a feature in which contestants need to say a certain phrase as a way of “buzzing in” to answer a question. The chatbot doesn’t recognize the phrase when it’s said by the Australian contestant, and as a result, that contestant has fewer opportunities to answer questions.

- □ Allocation
- □ Representation

Scenario B: Because this game show is televised, viewers see the Australian game show contestant be penalized for the way they speak and lose opportunities to win points. This has the potential to reinforce any negative stereotypes that viewers may hold about both this language variety and Australian people more generally.

- □ Allocation
- □ Representation

This caselet was developed by student Chloe Evered as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and supervised by Dr. Vandana P. Janeja, Dr. Christine Mallinson and Dr. Karen Chen at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

Q3: When is it important for you, as a developer of language technologies, to keep linguistic discrimination in mind as an important concern?

a. In medical contexts, when a patient’s health or well-being is at stake
b. When developing technology that will be used on live television
c. When creating a chatbot for commercial purposes that will be sold to a diverse market of consumers with many different language varieties
d. All of the above (and all other applications of language technology, even those that seem low-stakes)

Bias in deepfake detection

It’s important to keep in mind that, in the world of audio deepfake detection, these considerations of preventing linguistic discrimination and avoiding the reproduction of harmful language ideologies are of utmost concern. **Deepfakes can be created in any language variety**, and developing detection systems that perform best on one particular language variety disadvantages speakers of other varieties, which can exacerbate existing inequalities that fall along linguistic lines.

This caselet was developed by student Chloe Evered as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and supervised by Dr. Vandana P. Janeja, Dr. Christine Mallinson and Dr. Karen Chen at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

References & Resources

Ashtari, N., & Krashen, S. (2023). Confronting linguistic racism | USC Rossier School of Education. https://rossier.usc.edu/news-insights/news/confronting-linguistic-racism

Crawford, K. (2017). *The trouble with bias* [Keynote]. Neural Information Processing Systems (NIPS), Long Beach, CA. https://youtu.be/fMym_BKWQzk?si=qthmIxwxEekFNkq6

Understanding Bias I | *Machines Gone Wrong.* (n.d.). Machinesgonewrong.com. https://machinesgonewrong.com/bias_i/

What is Algorithmic Bias? (n.d.). Center for Critical Race + Digital Studies. https://www.criticalracedigitalstudies.com/peoples-guide-posts/what-is-algorithmic-bias

Martin, J.L., & Wright (2023) Bias in Automatic Speech Recognition: The Case of African American Language, *Applied Linguistics, 44*(4), pp. 613–630. https://doi.org/10.1093/applin/amac066