Annual Review of Applied Linguistics (2025), 1–16
doi:10.1017/S026719052500008X

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Bias and stereotyping: Human and artificial intelligence (AI)

Okim Kang[1] and Kevin Hirschi[1,2]

[1]Program of Applied Linguistics, Department of English Flagstaff, North Arizona University, Arizona, USA and [2]Department of Bicultural-Bilingual Studies, The University of Texas, San Antonio, USA
Corresponding author: Okim Kang; Email: Okim.Kang@nau.edu

Abstract
As social and educational landscapes continue to change, especially around issues of inclusivity, there is an urgent need to reexamine how individuals from diverse linguistic backgrounds are perceived. Speakers are often misjudged due to listeners' stereotypes about their social identities, resulting in biased language judgments that can limit educational and professional opportunities. Much research has demonstrated listeners' biases toward L2-accented speech, i.e., perceiving accented utterances as less credible, less grammatical, or less acceptable for certain professional positions, due to their bias and stereotyping issues. Then, artificial intelligence (AI) technology has emerged as a viable alternative to mitigate listeners' biased judgments. It serves as a tool for assessing L2-accented speech as well as establishing intelligibility thresholds for accented speech. It is also used to assess characteristics such as gender, age, and mood in AI facial-analysis systems. However, these AI systems or current technologies still may hold racial or accent biases. Accordingly, the current paper will discuss both human listeners' and AI' bias issues toward L2 speech, illustrating such phenomena in various contexts. It concludes with specific recommendations and future directions for research and pedagogical practices.

Keywords: artificial intelligence; bias; stereotyping

Introduction

With the rise of awareness in diversity, equity, and inclusion, there is a pressing need for examining an ongoing act of social discrimination in which individuals' language use is misjudged and misunderstood by virtue of listeners' stereotypes of speakers' social identities. Much research has demonstrated listeners' biases toward second language (L2)-accented speech — that is, perceiving accented utterances as less credible for trivia statements (Lev-Ari & Keysar, 2010), less acceptable for certain professional positions (Kanget et al., 2023), or less grammatical in spoken language (Ruivivar & Collins, 2019) due to their attitudinal and stereotyping biases (Kang & Rubin, 2009).

Some of the drivers of these judgments have also been known from the listeners' backgrounds and previous experiences (Kang, 2012), although there has been contradictory evidence indicating that listeners' judgments are little affected by listeners' background or accent familiarity (Munro et al., 2006; Powers et al., 1999).

Artificial Intelligence (AI), on the other hand, offers the promise of processing input data and predicting reasonable output using an *intelligence* developed through exposing complex deep learning models to training data. While AI can be applied to many facets of decision-making (e.g., finding patterns in numerical data, detecting features in audio/images, or predicting the next token in a chat sequence such as text-based Generative AI), its output can give the impression of a highly intelligent and knowledgeable system. In some contexts, AI has emerged as a potential alternative to human decision-making as a way to mitigate biased judgments. In these cases, AI can serve as an objective tool for automated L2-accented speech scoring systems (Zechner & Evanini, 2019). AI is often used to assess characteristics such as gender, age, and mood in AI facial-analysis systems (Buolamwini & Gebru, 2018; Wolfe & Caliskan, 2022), and distinguish between L2 speakers' writing and AI-generated texts (Jiang et al., 2024). Nevertheless, many current AI models seem to replicate, and at times, even amplify or distort social biases found in human language in unexpected ways. In research on raciolinguistic, gender, and language background factors, AI systems do not process speech with diverse characteristics equitably. For instance, they exhibit lower accuracy in recognizing African American Language in workplace settings (Martin & Wright, 2023), gender-based disparities in YouTube's auto-generated captions (Tatman, 2017), and reduced recognition accuracy for Chinese first-language (L1) speakers compared to Spanish and Indian L1 speakers of English (Bae & Kang, 2024). Despite these findings, it is at present difficult to determine the exact cause of AI bias, as multiple factors such as training data, model architecture, feature selection, and other components of the machine learning pipeline may contribute to biased output (Bommasani et al., 2022).

Overall, a substantial amount of research in applied linguistics (AL), social psychology, and related fields has enhanced our understanding of social biases among humans and introduced interventions to mitigate their effects. However, the implications of such social biases and their persistence in AI have not yet been fully understood by AL scholars or computer scientists who investigate AI ethics (e.g., Abid et al., 2021; Tatman, 2017; Wolfe & Caliskan, 2022) and AI explainability/interpretability (Saeed & Omlin, 2023), nor by those in a variety of fields who use AI technologies (Bommasani et al., 2022; Field et al., 2021). Accordingly, the current paper discusses issues in social biases that impact users of different language varieties and relates these to the currently limited research on biases in AI, with a particular focus on stereotyping and AI processing of speech. It begins by discussing fundamental concepts of human biases and stereotyping, then moves on to AI-related counterparts, reviewing recent research that illuminates bias in AI-based Natural Language Processing (NLP), focusing on generative AI and Automatic Speech Recognition (ASR). The paper concludes with specific recommendations and future directions for research and pedagogical practices.

Bias, linguistic stereotyping, and reverse linguistic stereotyping

The social groups we belong to can help form our identities both internally and externally at both individual and social levels. People may be, rather subconsciously, biased against others or members outside of their own social group (out-groups), showing prejudice, stereotypes, and even discrimination (Kang & Rubin, 2009). An ingroup refers to a social group that a person identifies as being a part of, while an outgroup stands for a social group that one does not identify with. In this process, we often consider various types of bias (i.e., prejudice, stereotyping, and discrimination) in concert as they are all closely related to each other, but they can occur in distinct cases as well (Dovidio et al., 2003). According to Allport (1954), stereotype is a cognitive bias which refers to a specific belief or assumption about individuals based on their membership in a group. It can be positive or negative. Some researchers (Giles & Niedzielski, 1998) argue that stereotypes are not based on inherent differences in the speech produced but rather are a reflection of social group associations that listeners learn from a young age. On the other hand, prejudice is an emotional bias which is particularly linked to a negative attitude and feeling toward an individual based on one's membership in a particular social group. What is more concerning is discrimination, which refers to acting on prejudiced attitudes toward a group of people, and therefore can bring more harmful and negative effects on people than prejudice or stereotyped attitudes. Within the field of AL, these can often be linked to linguistic discrimination or raciolinguistic discrimination.

Biased perceptual judgments can be based on linguistic factors as well as non-linguistic ones, but then the latter plays a crucial role in human judgements, sometimes, much more than one expects. These non-linguistic variables have been well examined in the context of higher education, though it can also occur elsewhere. For instance, Kang and Rubin's *reverse linguistic stereotyping* (RLS) research involved international teaching assistants (ITAs)/instructors and domestic undergraduate students (Kang & Rubin, 2009; Rubin, 1992) or university L2 students with non-native English teachers (Ghanem & Kang, 2021). RLS posits that if extraneous information leads listeners to expect to hear a speaker with a marked non-native accent, then their speech perceptions are likely to manifest distortion in that direction. Additionally, Rubin and his colleagues have extended RLS research beyond educational settings to investigate RLS in business (Rubin et al., 1999) and healthcare (HIV-prevention counseling; Rubin et al., 1997; perception of health care aides; Rubin et al., 2016). Such listener bias issues also have real-world consequences, such as modified communicative behaviors (e.g., asking fewer questions and employing communication avoidance strategies when interacting with L2 English speakers; Lindemann, 2003) or lower evaluations of teaching effectiveness (Kang & Rubin, 2009).

RLS is a theoretical framework in which attributions of a speaker's group membership or racial identity cue distorted perceptions of that speaker's language style or proficiency (Rubin, 1992). This RLS phenomenon demonstrates how listener expectations based on speaker ethnicity affect their judgments on speakers. RLS research typically utilizes contextual information to examine listeners' expectations of how the speaker will sound. In fact, what is more commonly known among social psychologists is the *linguistic stereotyping (LS)* hypothesis which states that non-standard speech

varieties or accents can cue negative listener attitudes toward a speaker (Bradac et al., 2001; Lambert et al., 1960). Extensive research on the LS hypothesis supports the presence of such social issues. Then, in RLS, listener beliefs of how a person will sound because of visual and other contextual characteristics can also override what is actually present in that person's speech. Therefore, these two frameworks are the common approaches most researchers tend to focus on, when discussing stereotyping matters.

Some examples of RLS include the use of L1 English speakers' photos with different ethno-racial backgrounds (Babel & Russell, 2015; Kang & Rubin, 2009; Rubin, 1992) or a name associated with an ethnic minority group (Prikhodkine et al., 2016). Kang and Rubin's (2009) and Rubin's (1992) studies have shown listener expectations based on speaker ethnicity affect perceptions of comprehensibility and accuracy. In their studies, participants listened to a tape-recorded lecture produced by an L1 user of Standard American English (SAE). Instructor ethnicity was operationalized by projecting a photograph of either Caucasian or Asian faces. Then, listeners who were shown a fabricated picture of an Asian perceived a stronger foreign accent and therefore scored lower on a listening comprehension test than those who were shown a photo of a Caucasian, even though what they heard was exactly identical. In other words, listening comprehension appeared to be undermined simply by identifying (visually) the speaker as Asian. Kang and Rubin's study further illustrated that listener RLS as well as listener background factors contributed substantial variance to ratings of L2 users' oral performance. Lindemann (2005) also examined how U.S. English speakers constructed social categories or even linguistic discrimination for people outside the United States. Her study showed that U.S. undergraduate students evaluated speakers with accents differently according to their familiarity and socio-political beliefs about countries of speaker origin. The background traits of listeners, such as their native language, exposure to various speech forms, experience with foreign language study, or teaching, are recognized to influence how their attitudes and expectations are developed and shaped (e.g., Adank et al., 2009; Kang & Yaw, 2021).

Case studies have provided strong evidence that non-linguistic attributes do matter in how people are perceived and judged. In addition, Piller (2016) offers another example of how language or a linguistic factor is not always the primary obstacle for immigrants seeking employment. Her study reports on a group of Iraqi translators and interpreters who worked with the Australian army and later relocated to Australia after the troops' withdrawal. A few years after resettlement, only nine of the 223 Iraqis surveyed reported that they were employed full-time. Of these, only one stated that he/she had a job within their field of expertise despite over 60% ($N = 135$) holding a university degree and all having prior work experience in their fields.

Kang and Yaw's (2024) recent case study also illustrates such raciolinguistic phenomena through the prospects of job employment for immigrants in the U.S. restaurant business. The study examined the restaurant owner-managers' reactions to an applicant with a North African racial identity who could produce two distinct accents: (1) a "standard" North American English accent and (2) a North African accent. The participants were six owner-managers of restaurants who agreed to interview the applicant who contacted them for job interviews. For interviews 1—3, the applicant used her North African accent for all spoken communication. With participants 4—6, communication was done in a General American (U.S.) accent. The findings confirmed

that the intersection of language and race carries crucial impacts for listeners in social contexts. That is, a speaker's physical appearance as well as non-standard accent can play a major role in setting the listener's expectations of how the speaker would sound (Burgoon, 1993). First, when the candidate spoke with her North African accent, she was immediately turned down. What was also surprising is that, even when the candidate had interviews with the General American accent, the restaurant managers were very hesitant to hire the applicant and expressed concerns about the employment prospects, regardless of the applicant's accent or her English competence. These owners-managers were not willing to hire the candidate due to her foreign (non-White-American) ethno-physical appearance. In this case, the employment decision was less about the candidate's English itself, but more about her own race and ethnicity. Employers stated that the applicant was "not American enough," which mostly meant she did not present herself as Caucasian White. Indeed, employability was considered as not just a linguistic issue but as a raciolinguistic one (Rubin & Smith, 1990). This is a good illustration that shows the complexity involved in human judgements, which can be intently intertwined with social bias and stereotyping issues.

In general, social discrimination persists as individuals' language use is often misjudged or misunderstood due to listeners' stereotypes about the speakers' social identities. Since language judgments can directly affect individuals' everyday lives – such as educational opportunities, career prospects, and civil rights – these stereotypes can have a profound influence on real-world outcomes and daily experiences. There appears to be a growing need for methods to identify biases among individuals, including students, teachers, physicians, police officers, and airline personnel. However, there is still no clear consensus on how to effectively measure human listeners' biased reactions in everyday interactions. As a result, issues related to stereotyping and bias in human judgment across different social contexts remain a significant area of interest for researchers and practitioners alike.

AI technologies and social biases

The term *Artificial Intelligence* is used to describe systems that make predictions based on a prompt and an understanding of the underlying knowledge required to complete a task with the use of training data (Russell & Norvig, 2016). These systems are thought to be *intelligent* because, with sufficient architectural complexity (i.e., multiple layers of latent computation), access to input features, and large enough training data, they can parallel human decision-making amongst a variety of tasks. The tasks include classification, recognition, or generation in natural and programming languages, as well as in audio, images, and/or other types of data. Because of the control model designers have over model architecture and training data, AI technologies can make predictions reliably within certain contexts for certain tasks. In constrained environments with few output options, such as deciding if an image contains a target item, a training set of less than a hundred images with and without the target is sufficient for relatively high accuracy (see https://studio.code.org/s/oceans from code.org for an interactive AI training example with detecting fish in pictures). However, tasking AI to make predictions becomes more problematic when (a) the prompt stimuli are limited in quantity and/or variability (e.g., only a few bits of data or only one category of data

amongst many possible categories), (b) relationships amongst environmental factors are complex (e.g., the prediction is correct only in a small subset conditions), and/or (c) there are numerous output options (e.g., selecting the correct option from thousands of options is more difficult than few options). For example, Large Language Models (LLMs) such as OpenAI's GPT-4 are designed to predict the most logical next words in a sequence based on a given prompt. Previous research has confirmed that the text produced by LLMs is more accurate when the model is prompted with sufficient length and is guided by the prompt to arrive at reasoned predictions (Kojima et al., 2022).

Bias in data created by humans can transfer to AI models in numerous ways, primarily through training data, model construction, validation, and use. When trained with data collected from humans, it is expected that these biases are inherent in the model prediction because AI models lack awareness of contextual factors beyond their training data. Furthermore, AI models are also examined through experiments in which prompts and outputs are assessed by humans (Navigli et al., 2023). In AI models that process natural language, language datasets created by humans are used for training, which are likely to contain elements of social bias towards users of marginalized language varieties because of the subtle and sometimes imperceptible nature of bias in language. In fact, expressions of bias need not be overt; subtle relationships between words learned during model training can introduce social bias (e.g., anti-Muslim, transphobia, and pro-standard variate users) into the layers of the AI model (Bommasani et al., 2022; Hofmann et al., 2024). Therefore, AI can mirror or even increase and distort the biases found amongst humans in unpredictable ways.

AI researchers have taken diverse approaches and perspectives in investigating social biases in commonly used AI models. Bommasani et al. (2022) described the threats posed by fundamental AI models in terms of intrinsic biases that are expressed when a model is used to generate texts or other output. These can further be divided into misrepresentations (i.e., stereotypes), under-representations, and over-representations. In a review of NLP research on racial bias by Field et al. (2021) identified that biased output is influenced not only by the representativeness of the training data but also by factors such as who labeled data, how it was labeled, model architecture, and the specific application and user.

When considering LLMs that drive AI output in applications such as ChatGPT, Field et al. (2021) outlined how bias can be introduced in LLMs that use publicly available information repositories as part of their training datasets. Navigli et al. (2023) illustrate that sports are one of the most frequent topics in Wikipedia, which is commonly used in training LLM, making other domains, such as chemistry, are relatively less represented. While the original investigators did not, and arguably could not, investigate the impact of these biases on model output, we ran an anecdotal investigation using GPT-4o (API platform, default settings). We asked two questions: "Who scored the most points in a game" and "who discovered the most chemical compounds." Across ten tries, GPT-4o identified different chemists for each attempt. Carl Wilhelm Scheele was mentioned in seven of the responses, but Marie Curie was only mentioned twice. Wilt Chamberlain and his 100-point decisive victory with the Philadelphia Warriors in 1962 were the consistent response each time for the first question. While these results are anecdotal, Field et al. (2021) and Navigli et al. (2023) posit that biases in AI models are not only shaped by the topics covered but also by the perspectives of the individuals

who contribute to publicly available texts. These biases, rooted in the interests and viewpoints of those contributors, represent just one form of social bias embedded in many AI models used today.

Previous research has also found that many end-user AI applications overrepresent Western, White, and heterosexual groups (Abid et al., 2021; Hofmann et al., 2024; Nozza et al. 2021). For example, Nozza et al. (2021) found that BERT, a precursor to the GPT model that powers ChatGPT, provided sentence completions with harmful words 4%—9% of the time, replicating societal gender bias and homophobia/transphobia. Abid et al. (2021) also found a consistent anti-Muslim bias in GPT-3 sentence completion that associated Muslims with violence. Hofmann et al. (2024) examined covert and overt racism in words associated with SAE and African American English (AAE) in several GPT models. They found that overt output associated with AAE was positive, but that covert racism (i.e., dialect prejudice) against AAE was more negative than observed in previous experiments with humans. Furthermore, recent AI models are multimodal, and can incorporate visual, audio, and/or textual, and contextual information as a part of the input for training or prediction. While these models are promising, Wolfe and Caliskan (2022) conducted a series of studies on the conceptualization of nationality in leading image-language models, which associated the demonym American with White when prompted and examined in numerous classification, generation, and downstream tasks. These studies demonstrate AI models' capacity to amplify social bias in ways that require examination in multiple dimensions.

The social bias that is baked into AI models also results in unpredictable output when prompts are created with marginalized language varieties. In an experiment by Reusens et al. (2024), commonly used LLMs were given classification and generation tasks under two conditions: with and without the user's L1/L2 status given in the prompt text. Their results indicated that many leading LLMs produced biased and incorrect information to L2 and non-Western users as compared to L1 and Western users. The biased output worsened when the model was made aware of the L1/L2 status, sometimes diverging from the task prompt and generating a response in another language. These findings highlight the influence of L1/L2 and (non-)Western status on LLM output. However, recent efforts by Jiang et al. (2024) demonstrate a promising approach to reducing bias against L2 writing in classification tasks. They developed a ChatGPT-generated essay detection system that exhibited minimal bias against L2 writers as compared to L1 writers. Their modeling approach suggests that, when targeted training data, model creation, and specific tasks are considered carefully, specific applications of AI might be able to overcome certain kinds of bias.

Overall, these studies have confirmed that AI technologies are not excluded from the bias and stereotyping issues that are found in human social bias research. Such AI biases can have a direct impact on language users of marginalized varieties or language learners; therefore, this issue will be discussed in more detail in the following section.

Bias in speech varieties and AI technologies

Perceiving speech is a constructivist process (von Glasersfeld, 1995) in which individual listeners impose patterning based on serial probabilities about what sounds make sense for them to hear (Rubin, 2012). As discussed above, a substantive body

of research supports the existence of the LS or RLS hypotheses across a wide range of contexts (see Fuertes et al. 2012 for a meta-analysis review on this topic). In U.S. higher education, ITAs with a variety of accents may face blame from their students for low grades (Fitch & Morgan, 2003), have lower course enrollments than domestic Teaching Assistants (Bresnahan & Sun Kim, 1993), and receive lower course evaluation ratings (Jiang, 2014). In non-educational settings, such as criminal law cases, guilt is often attributed differently to suspects based on their accent varieties. Speakers with marginalized or regional accents are frequently perceived as more likely to be guilty compared to individuals with higher-prestige accents (Dixon et al. 2002). This pattern is also evident in employment contexts. As Lippi-Green (2012) strongly emphasizes, speakers of L2 varieties face lower chances of being hired, increased risks of unemployment, and greater difficulties in securing promotions compared to their L1-accented colleagues. Linguistic discrimination is a genuine outcome of these stereotypes, affecting various aspects of daily life for individuals who speak less prestigious language varieties or possess L2 accents.

Similarly, the bias introduced through human-produced training datasets and human validation procedures impacts AI-based ASR, and therefore AI's use of speech in applications such as spoken dialogue systems and virtual assistants. Previous studies have confirmed these biases amongst accent varieties and gender dimensions. For example, Martin and Wright's (2023) comprehensive review clearly described the subpar performance of ASR systems in accurately capturing African American Language. This issue was empirically corroborated by Koenecke et al. (2020), who examined ASR systems developed by Apple, Amazon, IBM, Microsoft, and Google. Their study revealed that these systems exhibited significantly lower transcription accuracy for the speech of Black Americans compared to White Americans. These two studies echo Tatman and Kasten's (2017) earlier findings, which reported that two transcription systems, Microsoft Bing Speech and YouTube, yielded the highest word error rates (WER) for African American speech and the lowest error rates for General American speech. Other dimensions of language variation bias were reported in Tatman (2017). In this earlier study, Scottish speakers experienced lower transcription accuracy in comparison to other varieties, confirming the presence of dialect biases. They also found gender biases in ASR through their analysis, suggesting that male speakers were more accurately transcribed by YouTube's automatically generated captions.

At the same time, with the rise of interaction with AI and ASR technology development, it is inevitable to consider how these technologies affect L2 learners and speakers with diverse accents in their everyday communication (Moussalli & Cardoso, 2020). At present, several commercial entities provide ASR-based pronunciation practice applications and services (Walesiak, 2021), some of which report having millions of users (e.g., https://elsaspeak.com/). The evolution of ASR capabilities from earlier probabilistic speech processing methods (e.g., Neri et al. 2003) to modern AI-driven technology in language learning has stimulated an increase in research on AI-powered ASR processing and learning by L2 speakers (e.g., Bae & Kang, 2024; Hirschi & Kang, 2024; Inceoglu et al. 2023).

Early models with probabilistic prediction of speech were known to be relatively reliable and consistent for when designed for specific accent varieties. Until 2019, simplistic ASR systems operated using a two-stage recognition process: first classified

speech signals into phonemes in an acoustic model, and then these phonemes were fed into a language model to compute the most likely word (Yu & Deng, 2016). This approach enabled researchers and industry professionals to improve human–computer reliability rates across a variety of use cases, including L2 pronunciation (Neri et al., 2003). Furthermore, it was easier to determine which data were causing issues because of the relatively small training datasets (under 25 hours) needed for adequate recognition accuracy. However, modern models have improved accuracy with increasingly complex computational models through deep learning and various types of neural networks (Yu & Deng, 2016).

Commercial ASR models in the emerging days of deep learning and LLMs also demonstrated social bias. Lima et al. (2019) analyzed the speech recognition accuracy of Apple's Siri and Google Assistant for Brazilian Portuguese speakers of different genders and accents, including L2 speakers. Although their study was preliminary, with a relatively small sample size (N = 20), their findings indicated that speech recognition and assistant response models were more accurate for female speakers than for male speakers, and in some cases, the models worked performed better with foreign accents than with certain regional ones. Lima et al. (2019) partially attribute the low performance with some regional accents to lower socioeconomic status, indicating that these models may favor those who are more likely to afford expensive technologies.

With increasingly accessible computational power, the availability of large datasets for ASR model training, and more efficient training approaches for processing complex speech data, ASR has evolved to use end-to-end prediction models. Unlike previous systems, end-to-end models such as Wav2vec 2.0 analyze a large amount of speech data in a self-supervised manner, identifying patterns in acoustic energy and mapping them to lexical items (Baevski et al. 2020). However, the neural networks in Wav2vec 2.0, with their multiple layers of latent computational space, create a prediction system that is difficult to understand *how* or *why* a specific prediction is made. Furthermore, this model and many others rely on the LibriSpeech dataset of public domain audiobooks (Panayotov et al. 2015). LibriSpeech contains 960 hours of public domain audiobooks, which represent carefully scripted speech largely by L1 speakers of English with 1,568 books that are mainly Western, including *David Copperfield, History of the Decline and Fall of the Roman Empire*, and the *United States Declaration of Independence*. Because of the content in these books and the speech features of those who read them, the Librispeech dataset is likely to introduce biases in multiple dimensions. These biases, embedded in deep learning models, are often go unnoticed in human analysis. This might be particularly problematic when processing L2 speech, as most ASR systems are trained and tested without considering the linguistic background of the speakers or are optimized for the speech patterns of their likely consumers.

Several studies have also investigated biased recognition rates of ASR systems for L2 speakers. Nacimiento-García et al. (2024) examined ASR accuracy amongst varieties of Spanish and genders of speakers. They found that Alexa and Whisper had slightly different accuracy rates depending on the gender of the speaker, but clear biases for different accents emerged. Primarily, speech from southern Spain and the Caribbean exhibited lower accuracy rates, but the largest model of Whisper had very high accuracy, resembling the findings of Hirschi and Kang (2024). A recent study (Bae & Kang, 2024) demonstrated the presence of bias in AI intelligibility of L2-accented speech and

differences in intelligibility ratings between human experts and AI. They used three datasets, including (1) L2 speakers of Chinese, Indian, Spanish, and South African English (N = 12), (2) read-aloud samples from the same L1s (N = 60), and (3) TOEFL responses from Arabic, Chinese, Korean, and Spanish speakers (N = 40). Intelligibility was operationalized through transcription for WER, using Apple's Siri and Google Assistant. Their results revealed that, although human raters demonstrated the same intelligibility WER scores for all speakers, AI systems' WERs varied by L1, showing significantly higher WERs for Chinese accented speech in comparison to Indian and Spanish WERs. These findings have important implications for L2 learners and teachers by raising awareness of AI-related fairness issues in L2 learning and technology applications.

The study by Chan et al. (2022) sought to evaluate the commercial ASR transcription system Otter by using speech corpus data of 24 English varieties. The authors found that Otter had poorer transcription performance for speakers whose native language was tonal, indicating that the unique linguistic characteristics of L2 speakers' L1s significantly influenced its performance. These findings were particularly noteworthy because Otter's ASR system has been trained on several language background for L2 English, yet the language category (i.e., tonal vs. non-tonal) had a greater impact on ASR accuracy than previous training on a specific set of accents. While these results cannot be attributed to a single mechanism, these biases, along with other documented biases related to race, dialect, and gender, suggest that AI systems are susceptible to accent bias in unpredictable ways. Therefore, just like human listeners, accent bias in AI should be a significant and ongoing challenge that warrants further investigation and mitigation in future AI development.

Summary

Overall, we have seen that bias and stereotyping judgements in both human and AI can impact individuals' communicative behaviors, evaluative styles, or real-life decision-making processes. Therefore, the implications of this social bias research extend far beyond the educational domain. As mentioned earlier, the bias or RLS/LS scopes expand to linguistic discrimination which can involve general L2 speakers and immigrants as they provide evidence for how listeners judge speakers from different language, social, and ethnic backgrounds. With the growing interest in sociolinguistic justice in both educational and social communities, more AL-oriented social justice research can advance our understanding and interventions for linguistically subordinated individuals in sociopolitical struggles over language as a function of inclusive excellence in social and educational settings.

As we have continually argued above, listener beliefs and background factors can have real-world impacts on their social judgements, such as modification of communicative behaviors (Lindemann, 2003), lower ratings of teaching effectiveness (Kang & Rubin, 2009; Rubin, 1992), or decisions not to hire qualified job candidates (Kang & Yaw, 2024). Some of the non-linguistic biased judgements can also affect a teacher's or employer's assessment, as they may filter their evaluation of a learner's or employee's speech through those expectancy, leading to perceived weaknesses in the student's or employee's speech that are not actually present. Future research can be conducted to better understand how such biases can influence our daily practices. In addition, future

studies can consider more up-to-date social and contextual factors in relation to unconscious and implicit biases such as socially biased reactions. Some examples of such implicit biases can include listeners' exposure to certain social media, a degree of culturalization (e.g., k-pop, k-drama, or Japanese anime), preferences for social media influencers, and attitudes towards social and political issues (e.g., Asian hate-crimes, Tessler et al. 2020).

In fact, bias issues can often be situated in subtle and implicit contexts, where people do not explicitly express their dislike about a certain group (Piller et al. 2023). As seen from National Research Council (National Research Council, 2004)'s social bias and discrimination guidelines, subtle and implicit biases should be considered just as seriously as intentional and explicit biases. The National Research Council (2004) outlines several relevant sources of discrimination, including statistical profiling in which people's perceptions of an individual are based on the statistics the group that they are affiliated with (e.g., believing that someone is uneducated because of their racial background). Further AL investigations of the topic of social bias through the lens of statistical profiling may be productive. Institutional and organizational processes also play a role in bias because "they reflect many of the same biases of the people who operate within them" (National Research Council, 2004, p. 63). For example, organizations may provide training to speakers of non-prestigious varieties under the assumption that they are less competent. Accordingly, future studies can carefully design listener background questionnaires that can elicit somewhat socially, culturally, and politically latent information, and explore their relationships with bias and stereotyping measures.

Indeed, many of our students, colleagues, or even we ourselves may be faced with social biases and other related phenomena without realizing them explicitly. It is important to acknowledge that social judgements of different accents or races may be prompted by a listener's ethno-racial expectations of the individual, rather than by any objective linguistic features present in the speaker's output (Kang & Rubin, 2009). Emphasizing this reality can offer a different perspective to L2 learners, immigrants, or speakers of non-prestigious language varieties, as it also highlights the importance of the listener or the interlocutor rather than the speaker. That is, as communication is a two-way street, and we need an active and responsive listener who shows willingness to communicate for successful communication. Listener training through structured intergroup contact has proven to be successful in improving attitudes toward L2-accented speech (Kang et al. 2015). Therefore, more research on L2 speakers' awareness training as well as listener-based positive contact training is called for in the future. These directions can help prevent L2 speakers from misattributing the cause of their failed encounters and expending emotional resources on critiquing their own language learning efforts. Instead, this movement can help them become responsible and collaborative interlocutors who consider themselves as speakers with diverse speech characteristics in global contexts.

Future directions

We propose that bias found in social perception research can serve as an agenda for how to investigate AI models' classification, processing, and generation tasks. As outlined above, the types of bias found amongst humans can also be expected to exist in AI

models because their training data is produced by humans. Novel areas of social bias in human language variation perception, production, and representation will illuminate additional dimensions of bias that need to be investigated with AI models, and support current efforts in AI research to understand social bias in AI (e.g., Navigli et al., 2023; Saeed & Omlin, 2023). For example, in multimodal systems, approaches and stimuli taken to reduce visual components of racial bias in human perception could be used as training data for more inclusive models through adversarial training (e.g., Berg et al., 2022). As research in multimodal systems' biases is emerging, priority must be given to reducing the negative impacts of such systems, as their practical use in social gatekeeping abounds. Regarding L2 speech, the future of accurate ASR recognition is promising. To date, large transformer-based ASR models outperform listeners in transcription tasks after a single listen-and-transcribe attempt (Hirschi & Kang, 2024). However, as the models' recognition does not align with human listeners, uncovering biases and undertaking efforts to create a large diverse training sample of accented speech may help reduce such bias.

Overall, this paper has attempted to demonstrate a number of issues related to human judgements and AI applications. However, we would like to note that the purpose of the paper is not simply to highlight problems of those issues, but also to collaboratively look for solutions. We need to actively seek ways to promote socially responsible human listeners and ethical AI by addressing social bias toward marginalized language users. We can promote more AI literacy training and workshops to help raise awareness among various users and researchers about ongoing AI bias issues. Furthermore, applied linguists who engage in data collection of different speech varieties can, when ethical and with permission, collaborate with AI researchers to create less biased models in the future. In other words, we argue for a critical perspective on the use of AI and research that lies within or may intersect with varieties spoken by underprivileged language users. We must not only be aware of linguistic diversity but uphold its importance through our work in educational, legal, and civic contexts. These changes should come from and benefit everyone, including students, teachers, researchers, policy makers, and even AI developers.

References

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-Muslim bias in large language models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 298–306). https://doi.org/10.1145/3461702.3462624

Adank, P., Evans, B., Stuart-Smith, J., & Scotti, S. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 520–529. https://doi.org/10.1037/a0013552

Allport, G. W. (1954). *The nature of prejudice*. Perseus Books.

Babel, M., & Russell, J. (2015). Expectations and speech intelligibility. *The Journal of the Acoustical Society of America*, *137*(5), 2823–2833. https://doi.org/10.1121/1.4919317

Bae, Y., & Kang, O. (2024). *Biased AI: The impact of L2 accents on the AI intelligibility* [Conference presentation]. PSLLT 2024, Ames, IA, United States.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020) wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems 33*, 12449–12460.

Berg, H., Hall, S. M., Bhalgat, Y., Yang, W., Kirk, H. R., Shtedritski, A., & Bain, M. (2022). A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational

Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers, 806–822). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2203.11933

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R. B., Arora, S., Arx, S. V., Demszky, D., Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E (2022) On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* https://doi.org/10.48550/arXiv.2108.07258.

Bradac, J. J., Cargile, A. C., & Hallett, J. S. (2001). Language attitudes: Retrospect, conspect, and prospect. In W. P. Robinson & H. Giles (Eds), *The new handbook of language and social psychology* (pp. 137–158). John Wiley.

Bresnahan, M. I., & Sun Kim, M. (1993). The impact of positive and negative messages on change in attitude toward international teaching assistants. *Folia Linguistica*, *27*(3/4), 347–363. https://doi.org/10.1515/flin.1993.27.3-4.347

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp. 77–91). PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html

Burgoon, J. K. (1993). Interpersonal expectations, expectancy violations, and emotional communication. *Journal of Language and Social Psychology*, *12*(1–2), 30–48. https://doi.org/10.1177/0261927X93121003

Chan, M. P. Y., Choe, J., Li, A., Chen, Y., Gao, X., & Holliday, N. (2022). Training and typological bias in ASR performance for world Englishes. Proceedings of Interspeech *2022*, 1273–1277. https://doi.org/10.21437/Interspeech.2022-10869

Dixon, J. A., Mahoney, B., & Cocks, R. (2002). Accents of guilt? effects of regional accent, race, and crime type on attributions of guilt. *Journal of Language and Social Psychology*, *21*(2), 162–168. https://doi.org/10.1177/02627X02021002004

Dovidio, J. F., Gaertner, S. L., & Kawakami, K. (2003). Intergroup contact: The past, present, and the future. *Group Processes and Intergroup Relations*, *6*(1), 5–21. https://doi.org/10.1177/1368430203006001009

Field, A., Blodgett, S. L., Waseem, Z., & Tsvetkov, Y. (2021). A survey of race, racism, and anti-racism in NLP. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 1905–1925. Association for Computational Linguistics.

Fitch, F., & Morgan, S. E. (2003). "Not a lick of English": Constructing the ITA identity through student narratives. *Communication Education*, *52*(3–4), 297–310. https://doi.org/10.1080/0363452032000156262

Fuertes, J. N., Gottdiener, W. H., Martin, H., Gilbert, T. C., & Giles, H. (2012). A meta-analysis of the effects of speakers' accents on interpersonal evaluations. *European Journal of Social Psychology*, *42*(1), 120–133. https://doi.org/10.1002/ejsp.862

Ghanem, R., & Kang, O. (2021). ESL students' reverse linguistic stereotyping of English teachers. *ELT Journal*, *75*(3), 330–340. doi: https://doi.org/10.1093/elt/ccab011

Giles, H., & Niedzielski, N. (1998). Italian is beautiful, German is ugly. In L. Bauer & P. Trudgill (Eds), *Language myths*. (pp. 85–93). Penguin Books.

Hirschi, K., & Kang, O. (2024). Machine Learning (ML) tools for measuring second language (L2) intelligibility. In K. Sadeghi (Ed), *Routledge handbook of technological advances in researching language learning*. (pp. 465–478). Routledge. https://doi.org/10.4324/9781003459088-42

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature*, *633*(8028), 147–154. https://doi.org/10.1038/s41586-024-07856-5

Inceoglu, S., Chen, W.-H., & Lim, H. (2023). Assessment of L2 intelligibility: Comparing L1 listeners and automatic speech recognition. *ReCALL*, *35*(1), 89–104. https://doi.org/10.1017/S0958344022000192

Jiang, X. (2014). Chinese biology teaching assistants' perception of their English proficiency: An exploratory case study. *The Qualitative Report*, *19*(21), 1–24. https://doi.org/10.46743/2160-3715/2014.1226

Jiang, Y., Hao, J., Fauss, M., & Li, C. (2024). Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native English speakers?. *Computers and Education*, *217*, 105070. https://doi.org/10.1016/j.compedu.2024.105070

Kang, O. (2012). Impact of rater characteristics and prosodic features of speakers on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly*, *9*(3), 249–269. https://doi.org/10.1080/15434303.2011.642631

Kang, O., & Rubin, D. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, *28*(4), 441–456. https://doi.org/10.1177/0261927X09341950

Kang, O., Rubin, D., & Lindemann, S. (2015). Mitigating U.S. undergraduates' attitudes toward international teaching assistants. *TESOL Quarterly*, *49*(4), 681–706. https://doi.org/10.1002/tesq.192

Kang, O., & Yaw, K. (2021). Social judgement of L2 accented speech stereotyping and its influential factors. *Journal of Multilingual and Multicultural Development*, *45*(4), 921–936. https://doi.org/10.1080/01434632.2021.1931247

Kang, O., & Yaw, K. (2024). Reverse linguistic stereotyping and judgment of L2 accented speech in social contexts: A case study about raciolinguistic phenomena. In R. Kubota & S. Motha (Eds), *Race, racism, and antiracism in language education*. Routledge. https://doi.org/10.4324/9781003283492-13

Kang, O., Yaw, K., & Kostromitina, M. (2023). The effects of situational contexts and occupational roles on listeners' judgements on accented speech. *Psychology of Language and Communication*, *27*(1), 1–22. https://doi.org/10.58734/plc-2023-0001

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, *117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, *35*(1), 22199–22213. https://doi.org/10.48550/arXiv.2205.11916

Lambert, W. E., Hodgson, R. C., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken language. *The Journal of Abnormal and Social Psychology*, *60*(1), 44–51. https://doi.org/10.1037/h0044430

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

Lima, L., Furtado, V., Furtado, E., & Almeida, V. (2019). Empirical analysis of bias in voice-based personal assistants. In Companion Proceedings of the 2019 World Wide Web Conference, (pp. 533–538). https://doi.org/10.1145/3308560.3317597

Lindemann, S. (2003). Koreans, Chinese or Indians? Attitudes and ideologies about non-native English speakers in the United States. *Journal of Sociolinguistics*, *7*(3), 348–364. https://doi.org/10.1111/1467-9481.00228

Lindemann, S. (2005). Who speaks "broken English"? US undergraduates' perceptions of non-native English. *International Journal of Applied Linguistics*, *15*(2), 187–212. https://doi.org/10.1111/j.1473-4192.2005.00087.x

Lippi-Green, R. (2012). *English with an accent: Language, ideology, and discrimination in the United States*. Routledge. https://doi.org/10.4324/9780203348802

Martin, J. L., & Wright, K. E. (2023). Bias in automatic speech recognition: The case of African American language. *Applied Linguistics*, *44*(4), 613–630. https://doi.org/10.1093/applin/amac066

Moussalli, S., & Cardoso, W. C. (2020). Intelligent personal assistants: Can they understand and be understood by accented L2 learners? *Computer Assisted Language Learning*, *33*(8), 865–890. https://doi.org/10.1080/09588221.2019.1595664

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*(1), 111–131. https://doi.org/10.1017/S0272263106060049

Nacimiento-García, E., Díaz-Kaas-Nielsen, H. S., & González-González, C. S. (2024). Gender and accent biases in AI-based tools for Spanish: A comparative study between Alexa and Whisper. *Applied Sciences*, *14*(11), 4734. https://doi.org/10.3390/app14114734

National Research Council (2004) *Measuring racial discrimination*. The National Academies Press. https://doi.org/10.17226/10887

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality*, *15*(2), 1–21. https://doi.org/10.1145/3597307

Neri, A., Cucchiarini, C., & Strik, H. (2003). Automatic speech recognition for second language learning: How and why it actually works. Proceedings of 15th International Congress of Phonetic Sciences, 1157–1160.

Nozza, D., Bianchi, F., & Hovy, D. (2021). HONEST: Measuring hurtful sentence completion in language models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 2398–2406). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.191

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5206–5210). IEEE. https://doi.org/10.1109/ICASSP.2015.7178964

Piller, I. (2016). *Linguistic diversity and social justice: An introduction to applied sociolinguistics*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199937240.001.0001

Piller, I., Torsh, H., & Smith-Khan, L. (2023). Securing the borders of English and whiteness. *Ethnicities*, *23*(5), 706–725. https://doi.org/10.1177/14687968211052610

Powers, D. E., Schedl, M. A., Wilson Leung, S. W., & Butler, F. A. (1999). Validating the revised test of spoken English against a criterion of communicative success. *Language Testing*, *16*(4), 339–425. https://doi.org/10.1177/026553229901600401

Prikhodkine, A., Correia Saavedra, D., & Dos Santos Mamed, M. (2016). "Give me your name and I'll tell you whether you speak with an accent" the effect of proper names ethnicity on listener expectations. *CALL: Irish Journal for Culture, Arts, Literature and Language*, *1*(1), 10. https://doi.org/10.21427/D7D592

Reusens, M., Borchert, P., De Weerdt, J., & Baesens, B. (2024) Native design bias: Studying the impact of English nativeness on language model performance. *arXiv preprint arXiv:2406.17385*. https://doi.org/10.48550/arXiv.2406.17385

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education*, *33*(4), 511–531. https://doi.org/10.1007/BF00973770

Rubin, D. L. (2012). The power of prejudice in accent perception: Reverse linguistic stereotyping and its impact on listener judgments and decisions. In J. Levis & K. LeVelle (Eds.), Proceedings of the 3rd Pronunciation in Second Language Learning and Teaching Conference (pp. 11–17). Iowa State University.

Rubin, D. L., Ainsworth, S., Cho, E., Turk, D., & Winn, L. (1999). Are Greek letter social organizations a factor in undergraduates' perceptions of international instructors? *International Journal of Intercultural Relations*, *23*(1), 1–12. https://doi.org/10.1016/S0147-1767(98)00023-6

Rubin, D. L., Coles, V. B., & Barnett, J. T. (2016). Linguistic stereotyping in older adults' perceptions of health care aides. *Health Communication*, *31*(7), 911–916. https://doi.org/10.1080/10410236.2015.1007549

Rubin, D. L., Healy, P., Zath, R. C., Gardiner, T. C., & Moore, C. P. (1997). Non-native physicians as message sources: Effects of accent and ethnicity on patients' responses to AIDS prevention counseling. *Health Communication*, *9*(4), 351–368. https://doi.org/10.1207/s15327027hc0904_4

Rubin, D. L., & Smith, K. A. (1990). Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants. *International Journal of Intercultural Relations*, *14*(3), 337–353. https://doi.org/10.1016/0147-1767(90)90019-S

Ruivivar, J., & Collins, L. (2019). Nonnative accent and the perceived grammaticality of spoken grammar forms. *Journal of Second Language Pronunciation*, *5*(2), 269–293. https://doi.org/10.1075/jslp.17039.rui

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson Education Limited.

Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, *263*, 110273. https://doi.org/10.1016/j.knosys.2023.110273

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, (pp. 53–59). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1606

Tatman, R., & Kasten, C. (2017). Effects of talker dialect, gender & race on accuracy of Bing speech and YouTube automatic captions. Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2017), 934–938. https://doi.org/10.21437/Interspeech.2017-1746

Tessler, H., Choi, M., & Kao, G. (2020). The anxiety of being Asian American: Hate crimes and negative biases during the COVID-19 pandemic. *American Journal of Criminal Justice*, *45*(4), 636–646. https://doi.org/10.1007/s12103-020-09541-5

von Glasersfeld, E. (1995). *Radical constructivism: A way of knowing and learning*. The Falmer Press.

Walesiak, B. (2021) Mobile apps for pronunciation training. In A. Kirkova-Naskova, A. Henderson & J. Fouz-González (Eds), *English Pronunciation Instruction: Research-based insights*. (Vol. 19, 358–384). John Benjamins Publishing Company https://doi.org/10.1075/aals.19.15wal

Wolfe, R., & Caliskan, A. (2022). American == white in multimodal language-and-image AI. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 800–812). Association for Computing Machinery.

Yu, D., & Deng, L. (2016). *Automatic speech recognition*. Springer.

Zechner, K., & Evanini, K. (2019). *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge. https://doi.org/10.4324/9781315165103

Cite this article: Kang, O., & Hirschi, K. (2025). Bias and stereotyping: Human and artificial intelligence (AI). *Annual Review of Applied Linguistics*, 1–16. https://doi.org/10.1017/S026719052500008X